

cambridge.org/bbs

Herbert H. Clark<sup>a</sup>  and Kerstin Fischer<sup>b</sup> <sup>a</sup>Department of Psychology, Stanford University, Stanford, CA 94305-2130, USA and <sup>b</sup>Department of Design and Communication, University of Southern Denmark, DK-6400 Sonderborg, Denmark[clark@stanford.edu](mailto:clark@stanford.edu); [web.stanford.edu/~clark/](http://web.stanford.edu/~clark/)[kerstin@sdu.dk](mailto:kerstin@sdu.dk); [www.sdu.dk/ansat/kerstin](http://www.sdu.dk/ansat/kerstin)

## Target Article

**Cite this article:** Clark HH, Fischer K. (2023) Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46, e21: 1–65. doi:10.1017/S0140525X22000668

Target Article Accepted: 22 March 2022

Target Article Manuscript Online: 28 March 2022

Commentaries Accepted: 29 June 2022

**Keywords:**

anthropomorphism; depictions; human–robot interaction; robotics; social agents

**What is Open Peer Commentary?** What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 15) and an Author's Response (p. 56). See [bbsonline.org](https://bbsonline.org) for more information.

### Abstract

Social robots serve people as tutors, caretakers, receptionists, companions, and other social agents. People know that the robots are mechanical artifacts, yet they interact with them as if they were actual agents. How is this possible? The proposal here is that people construe social robots not as social agents *per se*, but as *depictions* of social agents. They interpret them much as they interpret ventriloquist dummies, hand puppets, virtual assistants, and other interactive depictions of people and animals. Depictions as a class consist of three physical scenes with part-by-part mappings between them: (a) a base scene (the raw physical artifact), (b) the depiction proper (the artifact construed as a depiction), and (c) the scene depicted (the scene people are to imagine). With social robots, evidence shows people form the same three scenes plus mappings: They perceive the raw machinery of a robot, construe it as a depiction of a character, and, using the depiction as a guide, engage in the pretense that they are interacting with the character depicted. With social robots, people also recognize three classes of agents – the characters depicted, the intended recipients of the depictions (those who view or interact with the robots), and the authorities responsible for the robots (the designers, makers, and owners). Construing social robots as depictions, we argue, accounts for many phenomena not covered by alternative models.

## 1. Introduction

This paper is about social robots, the class of robots that are designed to interact with people. Many social robots resemble people or animals, but others look like novel creatures. Some of them talk and listen; others only talk; others communicate with sounds, lights, or movements. Some of them walk on two legs, some on four or more; some move on wheels; others don't travel at all. Social robots have been designed to serve as tutors, caretakers, interviewers, receptionists, personal assistants, companions, service workers, and even sexual partners.

Social robots are a puzzle. On the one hand, people interact with them as if they were humans or pets. They talk with them, show them things, and engage with them in joint activities. At the same time, people know that social robots are mechanical artifacts. They have metal and plastic parts, sensors for vision and hearing, and speech that sounds artificial. It seems self-contradictory, even irrational, for people to hold these two attitudes simultaneously: (a) a willingness to interact with social robots as real people or animals and (b) a recognition that they are mechanical artifacts. The puzzle is not only theoretical but practical. When a robot stops moving, people must decide “Did the social agent fall asleep, or did the artifact's battery die?” And when its finger breaks off, “Am I sad because the social agent is in pain, or because the artifact needs repairing.” Call this the *social artifact puzzle*.

Our proposal is that people construe social robots not as agents *per se*, but as *depictions* of agents. Briefly put, depictions are physical analogs of what they depict. As we spell out later, depictions of humans range from static depictions (paintings, sketches, and sculptures of people) through staged depictions (stage actors and cartoon characters that enact people) to interactive depictions (ventriloquist's dummies, children's dolls, and hand puppets that interact with people). Social robots, we argue, are a type of interactive depiction. Depictions in general consist of three physical scenes with part-by-part mappings between them: (a) a base scene (the raw physical artifact or actor *simpliciter*), (b) the depiction proper (those features of the artifact or actor that are taken as depictive), and (c) the scene depicted (what people are to imagine). People are able to think about depictions from each of these perspectives. This is also true of social robots.

Treating social robots as depictions of social agents, we argue, has distinct advantages over previous accounts. It also helps resolve the social artifact puzzle. We begin with a brief review of several previous approaches and hint at their shortcomings. We then characterize social robots as depictions and provide evidence that people interact with them consistent with this view. Finally, we conclude.

© The Author(s), 2022. Published by Cambridge University Press



## 2. Previous approaches

For much of the past century, robots were only fictional. The term *robot* was introduced in a 1920 play by Karl Čapek (1921) for artificial agents who worked as household servants, and it was later adopted in science fiction for other agents. The term has since been extended to industrial machines, surgical robots, military weapons, and other artifacts that are not social.

But what are *social robots*? How are they viewed, not in science fiction, but by the people who interact with them? We begin with three previous approaches – named the media equation, trait attributions, and social constructions – and argue that they are useful, but incomplete. The account we propose is intended to resolve issues they do not cover.

### 2.1 Media equation

The idea behind the *media equation* (Reeves & Nass, 1996; see also Nass & Moon, 2000) is that people treat interactive devices as social actors: “All people automatically and unconsciously respond socially and naturally to media” (p. 7). The term *media* refers to computers, social robots, and other devices that people communicate with. The media equation is “media = real life.”

The main proposal is that people communicate with media using the same “social and natural rules” (p. 5) they would use in human-to-human communication. One study, for example, looked at gender stereotypes (Nass, Moon, & Green, 1997). People were brought into a lab, taught a body of facts by a computer, and later asked about the computer’s knowledge. The computer had a male voice for half of the participants and a female voice for the other half. Later, participants claimed that the computer knew more about technology than romance when it had a male voice, but the reverse when it had a female voice.

The media equation has been influential because it predicts many features of social interactions with media – from gender stereotypes, politeness, and interpersonal distance to arousal, timing, and physical size (for a review, see Reeves & Nass, 1996). Still, the media equation leaves certain questions unanswered. Why, for example, do people “respond socially and naturally” to social

robots when they know the robots are artifacts? And are their responses truly “automatic and unconscious”? These are just two issues we will return to.

### 2.2 Trait attributions

The *trait attribution* approach assumes that people are predisposed to attribute human traits to nonhuman artifacts (Epley, Waytz, & Cacioppo, 2007; Ruijten, 2015; Ruijten, Haans, Ham, & Midden, 2019; Waytz et al., 2010b, among others). People have this predisposition, it is argued, because they know more about human than nonhuman agents (Epley et al., 2007), so they “commonly anthropomorphize nonhuman agents, imbuing everything from computers to pets to gods with humanlike capacities and mental experiences” (Waytz, Gray, Epley, & Wegner, 2010a; Waytz et al., 2010b, p. 410). They do so “to satisfy the basic motivation to make sense of an otherwise uncertain environment.” People therefore treat humanoid robots on a “continuum ranging from low to high humanlikeness” depending on “the ease with which [anthropomorphic characteristics] can be ascribed to robots” (Ruijten, Bouten, Rouschop, Ham, & Midden, 2014, p. 1).

Trait attributions have been studied in social robots by identifying the features that contribute their humanlikeness. The features investigated include not only physical properties such as eyes and fingers (e.g., Mieczkowski, Liu, Hancock, & Reeves, 2019; Phillips, Ullman, de Graaf, & Malle, 2017), but also psychological capacities such as speech and agency (e.g., Crowell, Deska, Villano, Zenk, & Roddy, 2019; Mohammadi & Vinciarelli, 2012). In several studies, people rated how well human traits apply to a variety of entities (Epley et al., 2007; Gray, Gray, & Wegner, 2007; Reeves, Hancock, & Liu, 2020; Weisman, Dweck, & Markman, 2017). In Gray et al.’s study, the entities ranged from a baby, a fetus, a dead woman, and a frog to God, “you,” and a robot, and they were rated on dimensions of both “experience” (e.g., hunger, pain, fear) and “agency” (e.g., self-control, morality, memory). The robot was rated low in experience and moderate in agency.

It is one thing to attribute humanlike traits to an artifact and quite another to know how to engage, for example, with a being that is low in experience and moderate in agency. This is just one of the issues we will return to.

### 2.3 Social constructions

In the third approach, people treat social robots as human- or life-like beings – as *social constructions* – because of the way they are framed (e.g., Chang & Šabanović, 2015). Whenever people speak about robots as humanlike, for example, they invite others to treat them *as if* they were humans (Coeckelbergh, 2011).

Social constructions depend on pretense. When people interact with social robots, in this view, they engage in the pretense that they are interacting with actual beings (Airenti, 2018). This is a social practice that children learn in make-believe play (Bretherton, 1984, 1989; Garvey, 1990; Piaget, 1962). And evidence suggests that the more children engage in role play, the more willing they are to attribute humanness to nonhuman animals and artifacts such as robots (Severson & Woodard, 2018).

But what precisely are these social constructions, and how do they account for the different perspectives people take on social robots? These are issues we will return to.

HERBERT H. CLARK, Albert Ray Lang Professor of Psychology Emeritus at Stanford University, is the author of *Psychology and Language* (with Eve V. Clark), *Arenas of Language Use*, and *Using Language*. He is also the author of more than 150 journal articles and chapters in psychology and linguistics on how people use language, especially in everyday conversation. Clark was a John Simon Guggenheim Fellow and fellow at the Center for Advanced Study in the Behavioral Sciences. He has been elected to the American Academy of Arts and Sciences and the Royal Dutch Academy of Arts and Sciences. He was awarded an honorary doctorate from the University of Neuchâtel.

KERSTIN FISCHER, Professor for Language and Technology Interaction at the University of Southern Denmark, is the director of the Human-Robot Interaction Lab in Sønderborg. She is author of *Approaches to Discourse Markers, From Cognitive Semantics to Lexical Pragmatics, and Designing Speech for a Recipient* and more than 125 articles, chapters, and conference papers. She is a senior associate editor of *Transactions in Human-Robot Interaction* and has contributed to the Human-Robot Interaction Conference in various roles. Her research focuses on the foundational mechanisms of interaction.

## 2.4 Unresolved issues

The previous approaches help characterize the social artifact puzzle, but they are hardly the whole story. Consider three issues that are raised by the conversation in Table 1 between an English-speaking robot (Smooth, see Fig. 1) and three Danish adults (whom we will call Arne, Beth, and Carl).

The first issue is *willingness*: Arne, Beth, and Carl differed in how willing they were to interact with the robot. Beth stepped in front of it, looked at its face, listened to it, and responded to its “Cheers” with English “Cheers,” but Arne and Carl did just the opposite. Arne looked behind the robot and then spoke to Beth and Carl in Danish, a language the robot hadn’t used, while Carl stood back and looked at the robot out of the corner of his eye. In line with the media equation, Beth responded to the robot as if it were a person. Arne and Carl, however, did not. Why was Beth willing to interact with Smooth, but Arne and Carl were not?

The second issue is *changes in perspective*. Although Beth interacted with the robot in lines 7 through 10, she suddenly changed her perspective. She tilted her head toward Arne and spoke to him in Danish, something that would be very impolite in front of an actual person. That is, Beth treated the robot as a social agent one moment, but as an inanimate artifact the next. This behavior is at odds with all three previous approaches. How could she change perspectives so quickly, and why?

The final issue is *selectivity*. When the robot produced “Cheers!” Beth replied “Cheers,” lifted her glass, and laughed as if the robot was making a genuine toast. She did this even though the robot had no arms and couldn’t raise a glass in a follow-up drink. Beth was selective in her interpretation of the robot’s behavior. She took account of its speech but ignored its lack of arms. But how was she selective, and why?



Figure 1. Robot Smooth.

**Table 1.** Conversation between Smooth (a robot) and Arne, Beth, and Carl (from a corpus described in Langedijk & Fischer, 2023)

1		(Arne, Beth, and Carl watch the robot drive toward them, but quickly glance at each other to see what the others are doing.)
2	Smooth	Hi there!
3	Smooth	I wonder if you would like something to drink.
4	Arne	(Arne moves around the robot to see what it's carrying, laughing) Den har vand med. [Danish: It is carrying water.]
5	Beth	(looking at the robot) Ja? [Danish: Yeah?]
6		(Smooth turns to deliver the water on its back. Beth laughs.)
7	Smooth	(facing away from Beth) Take your drink please.
8		(Beth takes a glass of water.)
9	Smooth	(Smooth turns back to face Beth) Cheers!
10	Beth	(lifting her glass slightly) Cheers.
11		(Beth laughs, and tilting her head toward Arne and Carl, speaks in Danish while continuing to gaze and smile at the robot. Arne takes out his mobile phone to videorecord the robot. Beth drinks the water.)
12		(Smooth drives away while Arne, Beth, and Carl watch, Beth still smiling. Then Beth laughs and speaks in Danish.)

The alternative account we take up addresses these three issues plus many others not covered in the earlier proposals. In sections 3, 4, and 5 we take up depictions in general – what they are, the perspectives they entail, and the types of agents they depict – and show how social robots align with these properties. In sections 6, 7, 8, and 9 we examine three types of agents associated with social robots and show how they arise from treating the robots as depictions. In section 10, we summarize what we have argued.

## 3. What are depictions?

Our proposal is that people take social robots to be depictions of social agents – artifacts created to look, move, and sound like social agents. Construing them as depictions is no different from thinking of Michelangelo’s statue of David as a depiction of the biblical David, the puppet Kermit the frog as a depiction of a talking frog named Kermit, or a toy dog as a depiction of a dog. But what *are* depictions?

### 3.1 Depictions as physical analogs

Depictions represent physical scenes by the way they look, sound, or feel: They are physical analogs of those scenes (see Clark, 2016, 2019). By physical scene, we mean a configuration in space and time of people, objects, places, events, and any other physical elements.

Most depictions are intended to have *particular* interpretations. When Alice draws a circle on paper for Bert, for all he knows she could be depicting a bicycle wheel, an earring, or the trajectory of the earth around the sun. She needs to provide



Figure 2. Asimo.

Bert not only with the *perceptual display* (the drawn circle) but also with an *interpretive framework* for construing the display (e.g., “this is a bicycle wheel”). A depiction consists of the display plus its interpretive framework. Social robots are just such artifacts – perceptual displays with interpretive frameworks.

Social robots are clearly *intended* by their makers to be construed as depictions. Many have humanlike names and depict humanlike creatures (e.g., Honda’s Asimo [Fig. 2], SoftBank’s Nao and Pepper, Hanson Robotics’ Little Sophia and Professor Einstein). Others have pet-like names and depict pets. These include dogs (Sony’s Aibo, ToyQuest’s Tekno), cats (Phillips’ iCat), frogs (ToyQuest’s Rosco), turtles (ToyQuest’s Flash), seals (AIST’s Paro), and even scorpions (ToyQuest’s Scorpion). Still others depict beings that people know little or nothing about (e.g., Breazeal’s Jibo) or unknown “robotic” creatures (e.g., R2-D2 and BB8 from *Star Wars*).

### 3.2 Varieties of depictions

Depictions come in many types (see, e.g., Chatman, 1980; Clark, 2016; Walton, 1973, 1978, 1990, 2008, 2015). Table 2 lists 15 familiar ways of depicting people. Michelangelo’s *David*, for example, depicts a warrior, the biblical David, about to do battle with Goliath. Manet’s painting “*Chez le père Lathuille*” depicts an attentive young man chatting with a young woman in a Parisian restaurant. A *Punch and Judy* show depicts the activities of a quarrelsome husband and wife, Punch and Judy. The people depicted (David, the two Parisians, Punch and Judy) we will call *characters*.

Depictions of people run the gamut from sketches to social robots. The depictions listed in Table 2 differ on five relevant dimensions:

- (1) *Static versus dynamic depictions*. Types 1 through 5 are *static* depictions, and the rest are *dynamic*. Social robots like Asimo belong to the dynamic category.
- (2) *Staged versus interactive depictions*. Dynamic depictions divide into staged and interactive depictions. Types 6 through 10 are depictions that people view from outside the scene proper. Types 11 through 15 are depictions of characters

Table 2. Fifteen types of depictions with examples (and dates created)

Static depictions	
1. Statues	Michelangelo’s sculpture of the biblical David (1504)
2. Paintings	Edouard Manet’s painting “ <i>Chez le père Lathuille</i> ” of a man and woman chatting in a Parisian restaurant (1879)
3. Sketches	Pablo Picasso’s line sketch of the composer Igor Stravinsky sitting in a chair (1920)
4. Photographs	Dorothea Lange’s photograph of a poor migrant mother and her children in Nipomo, California (1936)
5. Mannequins	A clothed mannequin of an adult woman on display in Selfridges in London (1910)
Staged depictions	
6. Stage plays	Laurence Olivier and Vivien Leigh playing Hamlet and Ophelia in <i>Hamlet</i> at the Old Vic in London (1937)
7. Movies	Leonardo DiCaprio and Kate Winslet playing Jack Dawson and Rose Bukater in <i>Titanic</i> (1997)
8. Radio plays	John Hurt playing Albert in Tom Stoppard’s <i>Albert’s Bridge</i> on BBC radio (1967)
9. Animated cartoons	Animated cartoons of Snow White and dwarfs in Walt Disney’s <i>Snow White and the Seven Dwarfs</i> (1937) or Elsa the Snow Queen in Disney’s <i>Frozen 2</i> (2020)
10. Puppet shows	Punch and Judy performed by a puppeteer in a <i>Punch and Judy</i> show for children (1950)
Interactive depictions	
11. Dolls	Chatty Cathy, a “talking” doll by Mattel (1959)
12. Ventriloquist dummies	Ventriloquist Edgar Bergen’s dummy Charlie McCarthy talking with Bergen and his daughter Candice (1964)
13. Live characters	An actress dressed as Snow White chatting with visitors at Disneyland (2000)
14. Virtual assistants	Apple’s Siri, Amazon’s Alexa, and Google Assistant interacting with their users (2011, 2014, 2016)
15. Social robots	Honda’s humanlike robot Asimo (2000)



that people interact with from within the same scene. Social robots like Asimo belong to the interactive category.

- (3) *Actor versus prop depictions* (see Clark, 2016). Actor depictions are those in which the characters are *enacted*, or *played*, by actors within the scene. Types 6, 7, 8, and 13 are actor depictions. Prop depictions are those in which the characters are *established* or *controlled* by prop managers outside the scene. Types 1 through 5, 11, 14, and 15 are prop depictions. There are also hybrid depictions in which the characters are created by a combination of actors (for the voices) and props (for the bodies), as in types 9, 10, and 12. Social robots (e.g., Asimo) are prop depictions.
- (4) *Sensory modalities*. Paintings, photographs, and sketches rely on vision alone, and radio plays and virtual assistants on audition alone. Animated cartoons, movies, and stage plays deploy both vision and audition, and dolls add touch. Most social robots rely on vision, audition, and touch.
- (5) *Displacement*. The scenes depicted in statues, photographs, plays, and puppet shows are displaced from the here-and-now in both space and time. The scenes depicted in live broadcasts, Skype, and Zoom are displaced in space but not time. The scenes represented in interactive depictions, however, are displaced in neither space nor time. Social robots belong to this category.

On their very face, then, social robots such as Asimo are depictions of social agents. They are dynamic, interactive depictions that use vision, audition, and touch, and depict agents in the here-and-now. The category they belong to includes children's dolls, ventriloquist dummies, and hand puppets, so people should conceive of social robots in ways that are common to depictions like these. In what follows, we will use Asimo as our standard exemplar (see Fig. 2).

#### 4. Perspectives on depictions

Social robots are usually treated as indivisible wholes. Asimo would be referred to simply as “the robot” or as “Asimo.” But people interacting with social robots often take other perspectives on them, such as “the robot's machinery,” “the robot's material presence,” and “the character voiced by the robot,” and they switch easily from one perspective to another, as Beth did in Table 1. But what *are* these perspectives, and how are they connected?

##### 4.1 Depictions as multiple scenes

People interpret depictions in general as consisting of three physical scenes with mappings between them (Clark, 2016; Nanay, 2018).<sup>1</sup> We will illustrate with Michelangelo's sculpture *David*.

- (1) *Base scene*. In the Galleria dell'Accademia in Florence, there is a 5.17 m tall irregular piece of Carrara marble on top of a 2 m square block of white marble. This is the raw, uninterpreted material of Michelangelo's sculpture, the base scene for his depiction.
- (2) *Depictive scene (or depiction proper)*. Viewers are expected to construe a subset of features of the base scene as the depiction proper. They note that the upper half of Michelangelo's sculpture has the shape of a muscular man with something slung over his shoulder. This is the perceptual display. They also infer that the plinth is not part of the depiction proper. And they take a nearby sign that translates to “David by Michelangelo” as supplying an interpretive framework for

the display: Michelangelo intended the display to depict the biblical David.

- (3) *Scene depicted*. Viewers recognize that they are to use the depiction proper (the perceptual display plus its interpretive framework) as a guide in imagining the scene depicted – David preparing to attack Goliath with a slingshot.

Viewing depictions as a conjunction of these scenes satisfies two principles of depictions (Clark, 2016), namely:

*Double-reality principle*: Every depiction has two realities: its base, or raw execution; and its appearance, the features intended to be depictive.

*Pas-une-pipe principle*: A depiction is *not* what it depicts.

The first principle was proposed by Richard Gregory (1968, 1970, 2005; see also Maynard, 1994). In paintings, he noted, the canvas, splotches of paint, and brush strokes constitute one reality, and their interpretation as a depiction of a scene constitutes a second. The first reality is our base scene, and the second is the depiction proper. The *pas-une-pipe* principle is named after René Magritte's painting of a briarwood pipe over the words “Ceci n'est pas une pipe” (“This is not a pipe.”), and it relates the depiction proper, Gregory's second reality, to the scene depicted. It is similar to a principle offered by Korzybski (1948; see also Bateson, 1972; Borges, 1998; Carroll, 1894): “A map is not the territory it represents.”

##### 4.2 Mapping between scenes

For people to represent depictions as three scenes plus mappings, they need to *compartmentalize* the scenes – to represent them as distinct but connected. By the double-reality principle, they must distinguish the base scene – the raw artifact – from its construal as a depiction. And for the *pas-une-pipe* principle, they must distinguish the depiction proper from what it depicts. They must also represent the mappings between scenes. How might they do that?

People recognize that the scenes are connected in part-by-part mappings. There is a part of the marble in the base scene for *David* that corresponds to David's right hand, a part we will designate  $hand_{base}$ . The same portion of marble is construed in the depiction proper as a *depictive prop* for a man's right hand, a part we will designate  $hand_{prop}$ . That prop, in turn, maps into the actual right hand of the character David in the scene depicted, a part we will designate  $hand_{char}$ . The part-by-part mapping is:  $hand_{base} \rightarrow hand_{prop} \rightarrow hand_{char}$ . There are similar mappings for other parts of the sculpture, such as the stone in David's hand, David's left eye, and, of course, David as a whole (see Clark, 2016; Walton, 1990).

The three scenes plus mappings form a package of constraints. Elements of  $David_{base}$  constrain what viewers take to be the depictive prop,  $David_{prop}$ , and vice versa, and elements of  $David_{prop}$  constrain what they take to be the scene depicted,  $David_{char}$ , and vice versa. If Michelangelo had added a belt of stones to  $David_{base}$ , that would have changed  $David_{prop}$  and what viewers took to be  $David_{char}$ . If, instead, Michelangelo had labeled the sculpture “Apple picker,” that would have changed viewers' interpretation of  $David_{prop}$  and its relation to  $David_{base}$ .

##### 4.3 Language and perspective

If people can view depictions from three perspectives, they should be able to refer to each scene separately, and they can. When

Emma and Sam are looking at Michelangelo's *David*, she could point at the sculpture and tell Sam:

- (1) That is marble quarried in Carrara.
- (2) That is a statue of the biblical David by Michelangelo.
- (3) That is David preparing to slay Goliath.

Emma uses *that* to refer, respectively, to the base scene, the depiction proper, and the scene depicted – to  $David_{base}$ ,  $David_{prop}$ , and  $David_{char}$ . She could do the same for most parts of the sculpture. Pointing at David's right hand, she could tell Sam "That's Carrara marble," "That's an exaggerated size for a hand," or "Look – David is holding a stone."

All depictions allow for these perspectives. One could hold up a photograph of Eleanor Roosevelt and say, "I bought *this* in a flea market," or "*This* is a photograph of FDR's wife Eleanor," or "*This woman/She* was the first U.S. delegate to the U.N." And at a performance of *Hamlet*, one could nod at an actor on stage and whisper to a companion, "Didn't we see *him* at the pub the other night?" or "*That guy* is playing Hamlet," or "*He's* the prince of Denmark."

Part-by-part mappings in depictions are also reflected in language. In talking about *David*, viewers would use "hand" both for the sculpture's marble hand ( $hand_{prop}$ ) and for David's flesh-and-blood hand ( $hand_{char}$ ). In the languages we know, individual parts of a depiction proper are generally denoted with the same terms as the corresponding parts of the scene depicted. As a result, references to parts (e.g., "David's left hand") are often ambiguous.

#### 4.4 Perspectives on social robots

If social robots are depictions, they, too, should consist of three scenes plus mappings. Asimo, our standard exemplar, consists of these scenes:

- (1)  $Asimo_{base}$  is the artifact or machine that constitutes Asimo's base. It is made of metal, plastic, sensors, and other material, and has parts that move and make sounds.
- (2)  $Asimo_{prop}$  is the depiction proper. Like a hand puppet, child's doll, or ventriloquist's dummy, it is a prop for the agent it represents. It depicts  $Asimo_{char}$  in certain of its shapes, movements, and sounds.
- (3)  $Asimo_{char}$  is the character depicted by  $Asimo_{prop}$ . He is a male, humanlike being named Asimo.

The mapping is  $Asimo_{base} \rightarrow Asimo_{prop} \rightarrow Asimo_{char}$ , and there are similar mappings for many of Asimo's parts. As illustrated in Table 1, evidence shows that people can switch from one of these perspectives to another, often in quick succession (Fischer, 2021).

As with depictions in general, people can use "this robot" or "Asimo" to refer to Asimo's base, prop, or character. Examples: "This robot/Asimo/It was manufactured in Japan"; "This robot/Asimo/It looks quite like a human adult"; and "This robot/Asimo/He likes to kick soccer balls," or "Asimo, please kick the ball over here." So, when someone refers to "the robot" or "Asimo" or "Asimo's right leg," it is crucial to determine which perspective they are taking.

And people understand at least some of the mappings between scenes. Aibo, a robot dog sold by Sony, can sit, stand, lie down, bark, wag its tail, recognize its name, obey commands, and

respond to being petted. When owners play with their Aibo, they focus on  $Aibo_{char}$ . Still, they are aware that  $Aibo_{base}$  works from a battery and is turned on and off with a switch on the back of its neck. They realize that a cessation of  $Aibo_{base}$ 's power *causes* a cessation of  $Aibo_{char}$ 's behavior – a part-by-part mapping from Aibo's base to its character. We return to this point in section 9.2.

#### 4.5 Language problems

It is one thing to tacitly distinguish the three perspectives on a robot (a matter of cognition) and quite another to answer questions about them (a matter of meta-cognition). Consider a study by Kahn et al. (2012) that used a 120 cm tall humanoid robot called Robovie. In the study, 90 participants aged 9–15 interacted with the robot for 15 min and were then asked 40 questions about it. Participants had no problems with the interaction. Almost all reciprocated when Robovie offered to shake hands, asked them to point at something, and offered to hug them. But the participants did have trouble with the questions.

Two of the questions asked were: (1) "Does Robovie have feelings?" and (2) "Can a person sell Robovie?" If the participants thought of Robovie as a human – it had an adult male voice and spoke fluent English – they should have said "yes" to 1 and "no" to 2, but if they thought of Robovie as an artifact, they should have said the reverse. In fact, 60% of them said "yes" to 1, but 89% also said "yes" to 2. How is this possible?

One problem is that all 40 questions referred to the robot as "Robovie," which has at least three interpretations:

- (a) the raw artifact ("Robovie's battery lost power"),
- (b) the artifact construed as a depiction ("Robovie has metal rods for arms"), and
- (c) the character depicted ("Robovie just greeted me").

Question 1 about feelings makes sense for (c), but not for (a) or (b), whereas question 2 about selling makes sense for (a) or (b), but not for (c) (at least not for an adult male English-speaking character). Although this pattern fits the participants who said "yes" to both 1 and 2, roughly 40% of those who said "no" to 1 said "yes" to 2.

Participants clearly struggled with their answers. When asked if they thought Robovie was a living being, 14% said "yes" and 48% said "no," but 38% were unwilling to commit either way. The uncommitted "talked in various ways of Robovie being 'in between' living and not living or simply not fitting either category." As one insightful participant said, "He's like, he's half living, half not" (Kahn et al. [2012], p. 310). All these responses make sense if Robovie is an artifact construed as depicting a character of some kind.

To sum up, depictions represent physical scenes by the way they look, sound, and feel. People compartmentalize them into a base scene, depiction proper, and scene depicted, with part-by-part mappings between scenes. As a result, people can think about depictions from three perspectives. Social robots are just such depictions.

#### 5. Character types

To interact with a social robot, people need some idea of the character it represents. Is it a dog, a bear, or a human, and if it is a

human, is it a child or an adult, a male or a female? What does it know, what can it do, how does it interact with people?

One source of information is a robot's perceptual display, but that is always incomplete. Picasso's sketch of Stravinsky depicts Stravinsky's shape and clothing, but not his size, behavior, or ability to speak Russian. Olivier's performance in *Hamlet* depicts Hamlet's overall shape, size, clothing, and behavior, but not his ability to speak Danish. Much the same holds for social robots. Asimo<sub>prop</sub>'s display depicts Asimo<sub>char</sub>'s overall shape, but not his eyes or ears. People cannot be certain what Asimo<sub>char</sub> is capable of from Asimo<sub>prop</sub> alone.

### 5.1 Selectivity of features

Observers of a depiction implicitly realize that only some of its features are depictive (see Clark & Gerrig, 1990). By features, we mean aspects, parts, and capacities:

- (1) An *aspect* is a property that applies to most or all of a depiction proper. Michelangelo's *David* has such base aspects as "shape," "size," "color," "material," "orientation," "surface damage," and "pigeon droppings."
- (2) A *part* is a continuous, identifiable portion of the depiction proper, such as *David*'s left thumb.
- (3) A *capacity* is a property that is dynamic. Kermit the puppet, for example, allows movement of its jaw and head. *David* has no such capacities. One can identify the static aspects and parts of a depiction from still photographs of the depiction, but it takes evidence over time to identify its capacities.

Observers implicitly register the status of each feature of a display. For Michelangelo's sculpture, they know that "overall shape" is depictive, but that other aspects are not. "Material" and "size," for example, are *supportive* aspects. They are required in the perceptual display, but only as support for the aspect "shape." And the aspects "surface damage" and "pigeon droppings" are neither depictive nor supportive. Parts and capacities can be classified on the same grounds.

If Asimo is a depiction, its perceptual display should divide in the same way. As with *David*, the aspect "overall shape" is depictive, "material" and "size" are supportive, and for Asimo at least "surface painting" is neither depictive nor supportive. Asimo<sub>base</sub>'s head, torso, arms, fingers, and legs are depictive parts, but the hinges used for its finger joints, elbows, and knees are not. Asimo<sub>base</sub>'s head is a mix of depictive and supportive parts: it has ears, but no eyes or mouth. Asimo<sub>prop</sub> has the capacity to walk, run, kick balls, climb stairs, grasp things, and look around, but not to smile.

Depictions are also selective in *what* they depict. Picasso's sketch of Stravinsky depicts the shape and location of Stravinsky's eyes, but not their color or movement. A color video of Stravinsky would depict all four attributes. So, for the sketch, eye location and shape are *depicted* attributes; eye color and movement are *inferred* attributes. For the video, all four attributes are depicted. People make analogous inferences for Asimo.

With depictions, then, people must take feature selectivity seriously or risk misinterpreting them. For the ventriloquist's dummy Charlie McCarthy, people probably infer that McCarthy<sub>char</sub>'s eyes move even though McCarthy<sub>prop</sub>'s eyes are static. For Asimo, they eventually infer that Asimo<sub>char</sub> can see and talk even though Asimo<sub>prop</sub> has no eyes or mouth. For social robots, the hardest features to infer are capacities: Can they speak and understand

a language, and if so, what language and how well? What do they know and not know?

### 5.2 Nonstandard characters

Everyday depictions often represent *nonstandard* characters – beings that are not real. Many of these are near-humans, such as angels, pixies, zombies, and devils, and others are animals that talk, such as Mickey Mouse, Porky Pig, and Sesame Street's Big Bird. Still others are unlike any creatures we know. People can only guess at what the nonstandard characters can and cannot do.

All social robots represent nonstandard characters – and they are nonstandard in different ways. Asimo and Professor Einstein both represent humanlike creatures, but of quite different types. Asimo<sub>char</sub> can walk, run, climb stairs, kick balls, and grasp objects, but Professor Einstein<sub>char</sub> cannot. Professor Einstein<sub>char</sub> can make faces and answer scientific questions, but Asimo<sub>char</sub> cannot. Jibo, a personal assistant robot, has a spherical head, a single large eye, a short torso, and no limbs. Jibo<sub>char</sub> can speak, listen, and gesture with its head, but cannot locomote. And Ishiguro's robots (see Glas, Minato, Ishi, Kawahara, & Ishiguro, 2016) look uncannily like real people, but do not speak or behave like real people.

Robot characters are therefore best viewed as *composite* characters – combinations of disparate physical and psychological attributes. Asimo<sub>char</sub> has a head, arms, hands, fingers, and other standard human features, but no mouth, jaw, or lips. He can walk, run, climb stairs, and kick soccer balls, but cannot smile or frown. He can speak even though he has no mouth. Professor Einstein<sub>char</sub> and Jibo<sub>char</sub> are composites of entirely different types.

How do people infer the content of these composites? They base some of their inferences on heuristics available for depictions of people. These include:

- (1) *Static features* → *physical attributes*. Michelangelo's sculpture of *David* has two static eyes<sub>prop</sub> that point forward, so we infer that David<sub>char</sub> has eyes that are directed at something in the distance.
- (2) *Static features* → *associated abilities*. Although David<sub>prop</sub> has two static eyes<sub>prop</sub>, we infer that David<sub>char</sub> has the standard abilities associated with eyes. He can focus on things, identify objects, distinguish colors, and so on.
- (3) *Dynamic features* → *physical attributes*. The ventriloquist's dummy Charlie McCarthy<sub>prop</sub> produces fluent speech in interaction with Edgar and Candice Bergen, so we infer that McCarthy<sub>char</sub> has the standard anatomy for articulating speech.
- (4) *Dynamic features* → *associated abilities*. McCarthy<sub>prop</sub> produces fluent English, so we infer that McCarthy<sub>char</sub> has the standard human skills for speaking and understanding English.

With social robots, for example, people tend to ascribe greater moral responsibility to a robot the more humanlike features it has (Arnold & Scheutz, 2017). And people expect a robot to *understand* more instructions that are grammatically complex the more types of grammatically complex instructions it *produces* (Fischer, 2016).

Heuristics like these, however, go only so far. Most social robots are introduced with interpretive frameworks such as "Here is your companion" or "your tutor" or "a receptionist,"

but details of their characters can only be inferred over time, much as getting acquainted with a stranger takes place over time (see Fischer, 2016; Pitsch et al., 2009).

## 6. Interacting with characters

The characters represented in depictions are often agents. Think of Michelangelo's warrior, Shakespeare's King Lear, or Disney's Mickey Mouse. But depictions also involve *intended recipients*. They are the reason that sculptures are placed on plinths, paintings on walls, photographs in books, movies on screens, and plays on stage.

Depictions that are *static* or *staged* are designed to engage recipients in imagining the scenes depicted, scenes that are displaced in space, time, or both. But depictions that are *interactive* are designed to engage recipients in the current scene. Dolls are for children to play with, ventriloquist dummies are for people to talk to, virtual agents are for people to ask questions of, and social robots are for people to interact with. Still, the characters depicted are fictional, and people don't ordinarily interact with fictional characters. If so, how do people know what to do? Viewing social robots as depictions offers an answer.

### 6.1 Imagination in depictions

It takes imagination to interpret depictions of any kind (Clark, 1996, 2016, 2019; Clark & Van Der Wege, 2015; Walton, 1990). But what we imagine depends on the type of depiction.

With depictions that are static or staged, we imagine ourselves *transported* into the world of the scene depicted (see, e.g., Chatman, 1980; Clark, 1996; Clark & Van Der Wege, 2015; Gerrig, 1993; Oatley, 2011, 2016; Walton, 1990). For Manet's painting "Chez le père Lathuille," we imagine sitting in the sunny garden of a Parisian restaurant watching a young couple courting at a nearby table. At the movie *Titanic*, we imagine ourselves on the ship *Titanic* in 1912 watching Jack and Rose fall in love and seeing Jack go down with the ship. And at *Hamlet*, we imagine ourselves in medieval Denmark watching Hamlet and Ophelia talk.

With depictions that are interactive, on the other hand, we imagine characters *imported* into our world. On television in 1964, ventriloquist Edgar Bergen and his dummy Charlie McCarthy chatted with several people, including Bergen's daughter Candice. The recipients who were there all interacted with McCarthy<sub>char</sub> as if they were talking to him in the here-and-now. The same holds for people conversing with hand puppets, Apple's Siri, and Snow White at Disneyland.

Importation is different from transportation. With paintings, movies, and stage plays, recipients engage in the pretense that they are *covert observers* in the scenes depicted, where a covert observer is present in a scene, but invisible, mute, and unable to intervene (see Clark, 2016). With importation, recipients engage instead in the pretense that they are *co-participants* with the characters depicted.

### 6.2 Frames of reference

Every physical scene, real or fictional, has a *place-* and a *time-frame* – a spatial and a temporal frame of reference. Suppose Emma visited Florence in 2020 to see Michelangelo's *David*. The place and time for her was Florence in 2020, but the place and time of the scene she imagined was the Valley of Elah in

**Table 3.** Frames of reference when observing Michelangelo's *David* in 2020

	Place-frame	Time-frame
Base scene	Galleria dell'Accademia in Florence	A morning in 2020
Depiction proper	Marble prop for David in Florence	A morning in 2020
Scene depicted	David facing Goliath in Valley of Elah	Daytime, 700 BCE

700 BCE – if she knew her history (see Table 3). The scene depicted was displaced from the depiction proper in both space and time ( $place_{prop} \neq place_{char}$ , and  $time_{prop} \neq time_{char}$ ). Telephone conversations, chatting on Skype, and live TV broadcasts are displaced in space but *not* time ( $place_{prop} \neq place_{char}$ , and  $time_{prop} = time_{char}$ ).

Interactive depictions are displaced in neither space nor time ( $place_{prop} = place_{char}$ , and  $time_{prop} = time_{char}$ ). This we will call *auto-presence*. We will speak of the characters depicted (e.g., McCarthy<sub>char</sub>) as auto-present with the props that depict them (McCarthy<sub>prop</sub>). On the set of Sesame Street, adults and children interact with the characters Bert, Ernie, and Big Bird, who are auto-present with the puppets. And at Disneyland, visitors talk extemporaneously with the characters Snow White, Mickey Mouse, and Grumpy, who are auto-present with the actors playing them.

If social robots are interactive depictions, they, too, should require auto-presence. Indeed, when people interact with Asimo, they construe Asimo<sub>char</sub> as occupying precisely the same location and behaving at precisely the same time as Asimo<sub>prop</sub>. They imagine Asimo<sub>char</sub> as auto-present with Asimo<sub>prop</sub>.

### 6.3 Layers of activity

With dynamic depictions, observers keep track of two layers of activity – the depiction proper (e.g., *Hamlet* performed on stage), and the scene depicted (activities taking place in medieval Denmark). We will call these *layer 1* and *layer 2* (Clark, 1996, Ch. 12; Clark, 1999). For the audience at *Hamlet* in 1937, the scene they *actually* watched was layer 1, and the scene they *imagined* watching was layer 2:

*Layer 1:* On stage here and now, Olivier turns his head and body toward Leigh and produces "Get thee to a nunnery."

*Layer 2:* In Denmark centuries ago, Hamlet turns to Ophelia and says, "Get thee to a nunnery."

The audience used the actions they perceived in layer 1 as a guide to the actions they imagined perceiving in layer 2. Layer 2 mirrored layer 1 moment-by-moment and part-by-part.

Observers of depictions are expected to execute two classes of processes – *engagement* and *appreciation* (Bloom, 2010; Clark, 1996). The audience at *Hamlet* was to engage in – focus their imagination on or get engrossed in – layer 2, Hamlet and Ophelia's actions in medieval Denmark. Those who did were horrified when in Act III Hamlet drew his sword and killed Polonius, and when in Act IV they learned that Ophelia had drowned. These processes are all part of *engaging* in a depiction.



Observers must also appreciate the content of layer 1 and its relation to layer 2. At *Hamlet* the audience knew that Olivier's sword was fake, the actor playing Polonius didn't die, and Leigh didn't drown. In Act III they knew they shouldn't rush onstage to hold Olivier for the police or call an ambulance for the actor playing Polonius. They could also reflect on such aspects of layer 1 as Shakespeare's language, Olivier's and Leigh's acting, and the Old Vic's stage sets. All these processes are part of *appreciating* a depiction.

Interactive depictions have the same two layers. At one point, the ventriloquist's dummy Charlie McCarthy faced Candice Bergen and produced "How about you and me going dancing?" For Candice, the two layers were:

*Layer 1.* Here and now, the ventriloquist rotates McCarthy<sub>prop</sub>'s head toward Candice and produces "How about you and me going dancing?"

*Layer 2.* Here and now, McCarthy<sub>char</sub> asks Candice to go dancing.

It took the ventriloquist, Candice, and McCarthy<sub>prop</sub> together to create layer 1 (the depiction proper), and they did that to guide Candice and others in imagining layer 2 (the scene depicted). Candice engaged in the pretense that McCarthy<sub>char</sub> was asking her to go dancing when she replied "Great idea! We can do the watusi." At the same time, she clearly appreciated, among other things, that in layer 1 McCarthy<sub>prop</sub> was under the control of the ventriloquist, her father.

Interacting with social robots works the same way. Suppose Ben is kicking a soccer ball back and forth with Asimo.

*Layer 1:* Here and now, Asimo<sub>prop</sub> moves its foot<sub>prop</sub> and propels the ball to Ben.

*Layer 2:* Here and now, Asimo<sub>char</sub> kicks the ball toward Ben.

Asimo<sub>prop</sub> and Ben produce the events in layer 1 for Ben and others to use in imagining the events in layer 2. That is, Ben is to engage with Asimo<sub>char</sub> in layer 2 while appreciating Asimo<sub>prop</sub>'s function as a depiction in layer 1.

Imported characters have a curious property. Even though they themselves are not real, many of the activities they create in the here-and-now *are* real. When Asimo<sub>char</sub> and Ben are kicking the ball back and forth, the ball and kicking are real (cf. Seibt, 2017). If Asimo<sub>char</sub> were skillful enough, he could play in a real soccer game, running, passing, and scoring along with the real players. Imported characters can do real things when the objects, people, and activities they are auto-present with are also real.

#### 6.4 Acquisition of pretense

How do children interpret social robots? The short answer is as interactive toys – as props in make-believe social play. By age 2, children use dolls and stuffed animals as props for pretend characters, and by 3–4, they can alternate in their play between giving stage directions (such as "I a daddy" or "Pretend I'm a witch") and enacting characters ("I want my mommy") (Clark, 1997, 2009, 2020; Garvey, 1990). So, at an early age, they are fully prepared to assimilate robots into their play, and that is what they appear to do (see, e.g., Turkle, Breazeal, Dasté, & Scassellati, 2006).

At the same time, research on depictions suggests that it should take children years to *fully* understand the dual layers of social robots. Children recognize that drawings of objects map into actual

objects by 12–18 months of age (DeLoache, Pierroutsakos, Uttal, Rosengren, & Gottlieb, 1998; Hochberg & Brooks, 1962), but they do not understand how maps and models map into physical layouts until they are 3–4 (DeLoache, 1991). Nor do they understand how television represents physical objects until they are 4 (Flavell, Flavell, Green, & Korfmacher, 1990). Children recognize that movies represent fictional worlds by age 4, but they do not differentiate fictional worlds in an adult-like way until they are 7 or older (Choe, Keil, & Bloom, 2005; Goldstein & Bloom, 2015; Skolnick & Bloom, 2006). It is an open question what children understand about social robots at each age.

## 7. Communicating with characters

It takes coordination for two individuals to interact with each other, and they cannot do that without communicating (Clark, 1996). But what if one of them is a person, say Margaret, and the other is a fictional character, say Kermit the frog? Margaret must engage in the pretense that she can communicate with Kermit as if Kermit speaks and understands English like an adult human. She must also appreciate that Kermit's handler, the puppeteer, is the person immediately responsible for Kermit's speech and actions. People need the same two layers in communicating with social robots.

### 7.1 Forms of communication

People communicate using both speech and gestures, where gestures include "any visible act of communication" (Kendon, 2004). Most robots are able to use gestures. Some use gaze for indexing addressees and nearby objects (e.g., Mutlu, Shiwa, Kanda, Ishiguro, & Hagita, 2009). Some point at things with their hands and arms, and some position themselves in relation to people and objects (e.g., Mead & Matarić, 2016). And some robots (e.g., Kismet, Erica, and Emys) make facial gestures (Breazeal, 2002; Glas et al., 2016; Paiva, Leite, Boukricha, & Wachsmuth, 2017). In one study (Chidambaram, Chiang, & Mutlu, 2012), a robot was fitted out with one or more of these cues: (a) facial expressions; (b) body postures; (c) eye gaze; (d) hand movements; (e) self-placement; and (f) vocal expressions. In general, the more of these cues the robot used, the more competent it was judged to be and the more often its advice was followed. It is in this process that social cues have the effects predicted by the media equation.

Most robots, however, do not make full use of language. Paro the seal blinks, coos, and moves its body, but doesn't speak. Leonardo, an animal-like creature, moves its eyes, ears, arms, and head, but also doesn't speak. Some robots, like Paro and Leonardo, understand some speech but do not speak. Others both speak and understand language but in ways that are severely limited (Moore, 2017). One limitation is in taking turns. In extemporaneous human-to-human conversations, listeners collaborate with the current speaker in locating the end of his or her turn so they can initiate the next turn with a minimum gap or overlap (Holler, Kendrick, Casillas, & Levinson, 2016; Sacks, Schegloff, & Jefferson, 1974; Schegloff, Jefferson, & Sacks, 1977). The challenge is to design social robots that collaborate on turns the way humans do.

### 7.2 Communication as pretense

When people talk with the puppet Kermit the frog, as we noted, they must engage in the pretense that Kermit<sub>char</sub> is real and

understands what they are saying. Not everyone is willing to do that, and the same holds for social robots.

In one study, participants were instructed to guide Aibo, the robot dog, along a complicated route by telling it where to go. Although Aibo opened the conversation with a greeting, only some people reciprocated (Fischer, 2016). Here is Aibo's greeting and the responses of four participants:

*Aibo*: Yes, hello. How are you?

- (a) A042: I I'm good, and you? (*laughter*)
- (b) A044: Hello Aibo. – I want you to go, straight ahead.
- (c) A047: Okay, good, and how are you?
- (d) A030: (*2 sec pause*) Go straight.

Participants A042, A044, and A047 each accepted Aibo's greeting and reciprocated with greetings of their own. But A030 was unwilling to play along. He ignored Aibo's greeting and launched right in on the first instruction. And A042's laughter suggests that she was self-conscious about the pretense even though she played along. In Fischer's terminology, A042, A044, and A047 were *players*, and A030 was a *non-player*.

Players contrast with non-players in the perspectives they adopt with robots (Fischer, 2006, 2011). With Aibo, the players were more likely than the non-players to use complete sentences and personal pronouns (such as "he" and "she"). In a follow-up study by Lee, Kiesler, and Forlizzi (2010), members of the public exchanged typed messages with a robot receptionist at the entrance of a university building. About half of the users opened their session with a greeting (such as "Hi" or "What's up?" or "Hello"); the rest started right in on their queries. The greeters were more likely to use politeness expressions, engage in small talk, disclose personal information, and avoid rude or intrusive language. In the authors' words, the greeters "treated the robot more like a person" (p. 36). The non-greeters treated it more like an "information kiosk," a mere source of information.

To be a player is to engage in the second, imagined layer of activity. When A047 heard Aibo<sub>prop</sub> depict the greeting "Hello. How are you?" she imagined, and became engaged in, the auto-present scene in layer 2:

*Layer 1*: Here and now, Aibo<sub>prop</sub> produces the sounds "Yes, hello. How are you?" and then I produce the sounds "Okay, good, and how are you?"

*Layer 2*: Here and now, Aibo<sub>char</sub> greets me with "Yes, hello. How are you?" and I politely reciprocate "Okay, good, and how are you?"

Although A030 heard the same greeting, he did not engage in layer 2.

People can appreciate what a depiction is and still not engage with it. When the robot greeted the Danes in Table 1, Beth engaged in layer 2 and reciprocated the robot's greeting, but Arne and Carl chose not to do so even though they recognized what Beth was doing. Appreciation doesn't require engagement, although engagement requires some level of appreciation.

So, not everyone is willing to play along with a robot – or to do so all the time. In the study by Kahn et al. (2012), 97% of the participants reciprocated when the robot greeted them and offered to shake their hand. But when Smooth greeted people with "Hello!" "Hi there!" or "Sorry to bother you, but..." in the lobby of a Danish concert hall, the percentage reciprocating was 78% (Fischer et al., 2021). And when a robotic wheelchair greeted people during one lab study, the percentage was only 40% (Fischer,

2016). A person's willingness to play along with a robot probably has several origins.

### 7.3 Processes in communication

People's responses to social robots, according to the media equation, should be automatic and unconscious (Nass & Moon, 2000; Reeves & Nass, 1996), but this is clearly too simple. People's responses cannot be automatic if, as just noted, people vary in their willingness to respond. And even when people do respond, only some of their processes are automatic and unconscious.

Social robots communicate with people via several media: print alone (e.g., the robot receptionist), speech alone (e.g., Professor Einstein), gestures but no speech (e.g., Aibo), and composites of speech and gestures (e.g., Asimo and Smooth). As a result, people responding to the robots require different processes, some automatic and some not, depending on the media:

- (1) *Skill-based processes*. It takes years of guided practice, or formal instruction, for people to speak, understand, or read a language. But once fully proficient in these skills, people process speech and print automatically (see, e.g., Nieuwland & Van Berkum, 2006; Van Berkum, 2008, 2009).
- (2) *Self-paced processes*. People read printed works at their own variable pace, a process they have some control over and is not automatic (Just & Carpenter, 1980).
- (3) *Concept-based processes*. When people read a passage (as with the robot receptionist), most try to imagine the scene described (Zwaan, 1999, 2014). They do so, however, one portion of the scene at a time and base their imagination on successive phrases of the passage (see Miller, 1993). This process is neither automatic nor unconscious.
- (4) *Percept-based processes*. When people examine a visual depiction (such as Michelangelo's *David* or Asimo), they try to imagine a character that is physically analogous to the display they are examining. The process of examining the display is generally piecemeal and self-paced even though what people imagine is a holistic character.
- (5) *Time-locked processes*. Speech and gestures (as in plays, movies, Siri, or Asimo) are evanescent – they fade instantly – so they must be processed instantly, or all is lost. Consider the moment in *Hamlet* when Olivier turned to Leigh and uttered the line "Get thee to a nunnery." Olivier intended the audience to imagine Hamlet turning to, gesturing for, and speaking to Ophelia *at precisely the same time* that he, Olivier, was turning to, gesturing for, and speaking to Leigh. The audience had to synchronize, or *time-lock*, the process of imagining to what they were seeing and hearing. This process must be instantaneous and largely automatic.

So, people's interpretation of a robot's speech is skill-based, but their interpretation of its actions is concept- and percept-based and time-locked to those actions. Some of these processes are automatic and unconscious, but others are not.

## 8. Authorities

Most depictions have intended interpretations, but intended by whom? We assume that Michelangelo was responsible not only for carving *David*, but for its interpretation as the biblical David. We also assume that Edgar Bergen was responsible for the interpretation of his dummy as Charlie McCarthy. Most of

us appreciate, on reflection, that Michelangelo and Bergen were the *authorities responsible* for these interpretations.

Other depictions entail a system of authorities. At the 1937 performance of *Hamlet*, the audience took Shakespeare to be responsible for the script, the actors for their acting, and the Old Vic Company for the staging. They surely appreciated all this and judged Shakespeare for the script, the actors for their acting, and the Old Vic Company for the staging. If social robots are depictions, people should assume an authority, or system of authorities, that is responsible for their interpretation, and evidence suggests they do.

### 8.1 Principals and rep-agents

Certain types of agents act on the authority of third parties – other individuals or groups. Suppose Goldberg’s Bakery hires Susan as a server. At home, Susan can do whatever she wants, but at work, she is responsible for actions within the limits set by her contract with the bakery. If she cheats or insults a customer, the customer would blame her but hold the bakery legally responsible. In Anglo-American law, Goldberg’s Bakery is called the *principal*, and Susan is called an agent of the principal (see Coleman, 1994). For clarity, we will call Susan a *rep-agent* (for “agent representing a principal”). That gives us a three-way distinction:

- (a) *Self-agents* act on their own authority and are fully responsible for their actions.
- (b) *Principals* are individuals, or groups of individuals, on whose authority others act as rep-agents.
- (c) *Rep-agents* act on the authority of specified principals.

**Table 4.** Examples of rep-agents, their settings, and third-party principals

Rep-agent	Setting	Principals
1. School teacher	Schoolroom	School officials
2. Tutor*	Private office	Tutor agency
3. Professor	Lecture hall, office	University officials
4. Nurse	Hospital areas	Hospital officials
5. Caretaker*	Retirement home	Retirement home officials
6. Receptionist*	Company entrance	Company owners
7. Legal secretary	Lawyer’s office	Law firm
8. Ticket seller*	Ticket booth	Cinema, theater owners
9. Server	Bakery counter	Bakery shop owner
10. Checkout clerk*	Checkout counter	Supermarket owner
11. Waiter, waitress	Restaurant	Restaurant owner
12. Meal deliverer*	Hotels	Hotel room service
13. Concierge*	Hotel lobby	Hotel owners
14. Police officer	City streets	City officials
15. Prostitute*	Brothel	Brothel owner
16. Pet dog*	House or on leash	Members of household

There are commercial robots for each of the starred rep-agents.

We are all familiar with rep-agents. Table 4 lists examples of specialists whose responsibilities are regulated by principals. Teachers, for example, have contracts with school officials to carry out certain instructional activities with students, so they are rep-agents for those officials.

Most social robots, we suggest, are construed as rep-agents. People assume that robot math tutors know math and not history, that robot receptionists know schedules and not chemistry, and that sex robots engage in sex and not discussions. Many people are also aware, to some degree, of who the rep-agents are responsible to. If a robot math tutor makes errors, the student may blame the tutor but hold the tutor’s principals responsible – the school or robotics company (for evidence, see Belanche, Casaló Luis, Flavián, & Schepers, 2020).

Most people realize that *all* social robots are limited in what they can do. And when Ben interacts with Asimo, he would assume that there are authorities responsible for what Asimo<sub>char</sub> actually does. If Asimo accidentally dented Ben’s car, Ben would seek damages from those authorities.

Human rep-agents are not always on duty, and even when they are, they can break off and act as self-agents. Here is an example:

One day Clark telephoned a hotel in Long Beach, California, to reserve a room for the following night. A woman answered.	
Woman	I’m sorry, but there are no rooms available for tomorrow night.
Clark	Well, are there any nearby hotels you could recommend?
Woman	Oh, I’m not in Long Beach. I’m in South Dakota – in Rapid City.
Clark	Gosh, I grew up right near you, in Lawrence County. Did you go to high school in Rapid?
Woman	Yes ... [conversation continues]

In the third turn the woman interrupted her official duties to engage in small talk with Clark, and after several turns, she returned to her official duties. Few robots have the capacity to engage in small talk (Moore, 2017).

### 8.2 Agent control

Social robots are usually framed as *autonomous* agents – as agents that act on their own authority – even though they are ultimately controlled by principals. The same holds for other interactive depictions. When people engage with Kermit, Charlie McCarthy, Snow White at Disneyland, and Chatty Cathy, they act as if these characters, too, are autonomous agents, and they improvise what they say and do in talking with them. People take this tack even though the characters are obviously controlled by a puppeteer, a ventriloquist, an actress, and children engaged in make-believe play.

What is different about social robots is that their controls are hidden, making the principals harder to identify. Sophisticated users may assume that robots are controlled by computer programs written by specialists for hardware that was designed by the manufacturers – two sets of responsible authorities – but others may have no idea. Indeed, people sometimes suspect that the robots they encounter are teleoperated by people hidden nearby (e.g., Kahn et al., 2012; Yang, Mok, Sirkin, & Ju, 2015), and they are often right. Many experiments use a so-called Wizard-of-Oz technique in which a teleoperator produces the robot’s speech and controls its actions.

In fiction, the *real* Wizard-of-Oz was a ventriloquist (Baum, 1900). The Wizard built “an enormous Head” with eyes and a mouth that moved, and from behind a screen, the Wizard produced “I am Oz, the Great and Terrible.” After the Wizard was unmasked by Dorothy the heroine, he told her:

“This [Head] I hung from the ceiling by a wire. I stood behind the screen and pulled a thread, to make the eyes move and the mouth open.”

“But how about the voice?” she enquired.

“Oh, I am a ventriloquist,” said the little man, “and I can throw the sound of my voice wherever you wish, so that you thought it was coming out of the Head.”

So, to use the Wizard-of-Oz technique is literally to treat the robot as a ventriloquist’s dummy. Ventriloquist dummies, of course, are interactive depictions – just like robots.

## 9. Emotions

With social robots, emotions are experienced both by viewers and by the characters depicted. But what emotions, about what, and for whom? Here again, depictions offer a useful analysis.

### 9.1 Emotions and depictions

Emotions are essential to many depictions. When we look at Dorothea Lange’s 1936 photograph, “Migrant mother,” we feel pity for the hungry, desperate woman it depicts. In Act III of *Hamlet*, we experience horror when Hamlet stabs Polonius. And in watching ventriloquist Edgar Bergen chat with his dummy, we laugh when the dummy makes fun of Bergen. The photograph, play, and ventriloquist show would be pointless without the emotions.

We experience some of our most vivid emotions at the movies (see, e.g., Rottenberg, Ray, & Gross, 2007; Walton, 1978). A study by Gross and Levenson (1995) identified 16 film clips that elicited emotions ranging from amusement, anger, and contentment to disgust, fear, and sadness, and did so consistently and vividly. In a related study (Gross, Fredrickson, & Levenson, 1994), 150 undergraduate women were shown a 5 min film clip from the movie *Steel Magnolias* while they were monitored physiologically. About 20% of the students cried, and the rest showed physiological changes consistent with being sad.

What were the students sad about? In our analysis, they engaged in the pretense that they were watching the scene in layer 2 while tacitly appreciating its relation to the scene they were watching in layer 1:

*Layer 1:* Here and now, a 2D color movie is showing actress Sally Field in a cemetery delivering lines to other actresses.

*Layer 2:* At some fictional place and time, M’Lynn Eatenton is speaking to friends at a funeral about the death of her daughter.

The students were sad, not about Sally Field in layer 1, but about M’Lynn Eatenton in layer 2. And they attributed grief, not to Sally Field, but to M’Lynn Eatenton. Still, movie critics were so moved by Sally Field’s acting that they nominated her for a Golden Globes Award. The students were sad about M’Lynn Eatenton (a depicted character), but the critics – and surely some students – were in awe of Sally Field (a responsible authority).

### 9.2 Emotions and social robots

Emotions should be just as essential to many social robots. Consider an online forum maintained by owners of Sony’s robot dog Aibo (Friedman, Kahn, & Hagman, 2003). According to their postings, Aibo “liked” to do things, got “angry,” got “very sad and distressed,” and showed “happy eyes” and a “wagging tail.” (Aibo’s eyes turned green to signal happiness and red to signal sadness.) These emotions belonged to Aibo<sub>char</sub>, but the owners experienced emotions as well. “I feel I care about him as a pal.” “He always makes me feel better when things aren’t so great.” “My emotional attachment to him ... is strong enough that I consider him to be part of my family, that he’s not just a ‘toy.’” Aibo’s owners reported these emotions even though they recognized that Aibo was an artifact. Fully 75% of their postings included such terms as “toy,” “battery,” “microphone,” “camera,” and “computer.” Owners not only distinguished Aibo<sub>char</sub> (“he” and “him”) from Aibo<sub>base</sub> (with a battery and computer) but saw Aibo<sub>char</sub> as a source of their emotions.

Emotions have been elicited by a wide range of social robots (see, e.g., Broadbent, 2017; Paiva et al., 2017; Seo, Geiskovitch, Nakane, King, & Young, 2015). To take one example (Logan et al., 2019), children aged 3–10 in a pediatric hospital interacted with one of three versions of a bear-like robot named Huggable: (a) the animated robot itself, (b) a video version of the animated robot, or (c) the same robot but without animation. Using a Wizard-of-Oz technique, a child-life-specialist got the children to chat, sing, and play games with one of the three depictions. In their language, the children expressed more joy and less sadness with the animated versions of the robot (in person and on video) than with the unanimated robot. Children’s emotions were elicited not by Huggable as an artifact or prop, but by the animated character it depicted.

Some emotions with robots have been validated physiologically. In a study by Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, and Eimler (2013), adults were monitored while they watched one of two videos about a small dinosaur robot named Pleo. In the normal video, a human handler fed, caressed, and stroked the robot, but in the so-called torture video, he punched it, choked it, and banged its head on a table. Compared to viewers of the normal video, viewers of the torture video were more aroused and, later, rated themselves as having more pity for Pleo and more anger for the handler. Whenever people experience “pity” and “anger,” it is not for inert objects but for sentient beings. If so, the viewers’ pity was not for the Pleo<sub>base</sub> or Pleo<sub>prop</sub>, but for Pleo<sub>char</sub> and their anger was not for the actor playing the handler, but for the handler as a character (see also Menne & Schwab, 2018; Suzuki, Galli, Ikeda, Itakura, & Kitazaki, 2015). This is as it should be if they see Pleo as a depiction.

Compartmentalization of emotions is easy to demonstrate in invented scenarios. Suppose Amy sees a forklift operator run into Ben and severely injure his arm. She would surely fear for Ben’s health, rush to his aid, and call for an ambulance. If she saw the same happen to Asimo, she would do none of that. She would take her time in contacting Asimo’s principal about the damage, and although she might advise Ben to sue the forklift operator, that is advice she would give, not to Asimo<sub>char</sub>, but to Asimo’s owner (see Belanche et al., 2020).

## 10. Conclusions

People conceive of social robots, we have argued, not as social agents *per se*, but as depictions of social agents. Depictions in



general can be static, like paintings and sculptures; they can be staged, like plays and movies; or they can be interactive, like hand puppets and ventriloquist dummies. The argument here is that people construe social robots as interactive depictions.

Depictions, such as Michelangelo's *David*, Shakespeare's *Hamlet*, and Kermit the frog, are physical analogs of what they depict – the biblical David, events in medieval Denmark, and a ranarian creature named Kermit. Each consists of three scenes – three perspectives – with mappings between them. The claim here is that people view social robots the same way. They see Asimo as an artifact (its base scene), which maps into a physical scene (the depiction proper), which maps into a humanlike character (the scene depicted). When people look at Asimo<sub>prop</sub>, they are to imagine Asimo<sub>char</sub>, the character depicted.

People anticipate doing more than *look* at social robots: They expect to *interact* with them. To do that, they must engage in the pretense that they are interacting with the characters depicted (layer 2) and, at the same time, appreciate the depictions (layer 1) in relation to the characters depicted. It takes both processes to interpret robots properly. As we noted at the beginning, when a robot stops moving, viewers must decide, “Did the character fall asleep, or did the robot’s battery die?” And when a robot’s finger breaks off, “Am I sad because the character is in pain, or because the artifact needs repairing?”

Depictions of social agents encompass three classes of agents – the characters depicted (e.g., David, Hamlet and Ophelia, and Kermit the frog), the intended recipients (the viewers), and the authorities responsible for them (e.g., Michelangelo, Shakespeare, the actors, and the Old Vic, and Kermit’s puppeteer). Social robots entail the same three classes. For Asimo, they are Asimo<sub>char</sub>, the people who interact with Asimo<sub>char</sub> and the authorities responsible for Asimo – the designers, makers, and owners.

Still, people differ in their understanding of social robots. Children aged 2 have only the most basic understanding of pretense and depictions, and it takes them years to grasp these in depth. And just as adults vary in their appreciation of the fake blood, stunt doubles, and mock fighting in movies, they surely also vary in their appreciation of how robots work. People also differ in their willingness to engage with the characters depicted.

To return to the social artifact puzzle, how is it that people are willing to interact with a robot as if it was a social agent when they know it is a mechanical artifact? Do we need a “new ontological category” for “artifacts that appear to think and feel, can be friends, and at least potentially lay some moral claims for kind, fair and just treatment” (Melson, Kahn, Beck, & Friedman, 2006, p. 4)? The answer is no. All of us have spent much of our lifetime – thousands of hours – engaging with depictions. We have all the experience we need to construe social robots as depictions of social agents.

**Acknowledgments.** We thank many friends and colleagues for their discussions on the issues we take up in this paper.

**Financial support.** Kerstin Fischer was funded in part by Carlsberg Foundation grant CF18-0166.

**Competing interest.** None.

## Note

**1** In Clark (2016, 2019) and Fischer (2021), the three scenes are called *base*, *proximal*, and *distal* scenes. Because social robots depict single characters in the here-and-now, we have replaced *proximal* scene with *depiction proper* and *distal* scene with *character*.

## References

- Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, 9(2136), 1–13.
- Arnold, T. H., & Scheutz, M. (2017). *Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI*. Paper presented at the 12th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Vienna, Austria.
- Bateson, G. (1972). *A theory of play and fantasy*. MIT Press.
- Baum, L. F. (1900). *The wonderful Wizard of Oz*. G. M. Hill.
- Belanche, D., Casaló Luis, V., Flavián, C., & Schepers, J. (2020). Robots or frontline employees? Exploring customers’ attributions of responsibility and stability after service failure or success. *Journal of Service Management*, 31(2), 267–289. doi: 10.1108/JOSM-05-2019-0156
- Bloom, P. (2010). *How pleasure works: The new science of why we like what we like*. Random House.
- Borges, J. L. (1998). *On exactitude in science*. In *Collected fictions* (pp. 325–327). Viking.
- Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.
- Bretherton, I. (Ed.) (1984). *Symbolic play: The development of social understanding*. Academic Press.
- Bretherton, I. (1989). Pretense: The form and function of make-believe play. *Developmental Review*, 9(4), 383–401.
- Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, 68(1), 627–652. doi: 10.1146/annurev-psych-010416-043958
- Čapek, K. (1921). *R. U. R. Rossum’s universal robots*. Aventinum.
- Carroll, L. (1894). *Sylvie and Bruno concluded*. Macmillan.
- Chang, W.-L., & Šabanović, S. (2015). *Studying Socially Assistive Robots in their Organizational Context: Studies with PARO in a Nursing Home*. Paper presented at the 10th Annual ACM/IEEE International Conference on Human–Robot Interaction (HRI), Portland, Oregon, USA. <https://doi.org/10.1145/2701973.2702722>
- Chatman, S. B. (1980). *Story and discourse: Narrative structure in fiction and film*. Cornell University Press.
- Chidambaram, V., Chiang, Y.-H., & Mutlu, B. (2012). *Designing Persuasive Robots: How Robots might Persuade People using Vocal and Nonverbal Cues*. Paper presented at the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction (HRI), Boston, Massachusetts, USA.
- Choe, K. S., Keil, F. C., & Bloom, P. (2005). Children’s understanding of the Ulysses conflict. *Developmental Science*, 8(5), 387–392.
- Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, 64(1), 1–37.
- Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.
- Clark, E. V. (2020). Perspective-taking and pretend-play: Precursors to figurative language use in young children. *Journal of Pragmatics*, 156, 100–109.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H. (1999). *How Do Real People Communicate with Virtual Partners*. Paper presented at the 1999 AAAI Fall Symposium, Psychological Models of Communication in Collaborative Systems, North Falmouth, Massachusetts, USA.
- Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, 123(3), 324–347.
- Clark, H. H. (2019). Depicting in communication. In Hagoort, P. (Ed.), *Human language: From genes and brains to behavior* (pp. 235–247). MIT Press.
- Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language*, 66(4), 764–805.
- Clark, H. H., & Van Der Wege, M. A. (2015). Imagination in narratives. In Tannen, D., Hamilton, H. E., & Schiffrin, D. (Eds.), *Handbook of discourse analysis* (2nd ed., pp. 406–421). John Wiley.
- Coeckelbergh, M. (2011). You, robot: On the linguistic construction of artificial others. *AI & Society*, 26(1), 61–69. doi: 10.1007/s00146-010-0289-z
- Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.
- Crowell, C. R., Deska, J. C., Villano, M., Zenk, J., & Roddy Jr, J. T. (2019). Anthropomorphism of robots: Study of appearance and agency. *JMIR Human Factors*, 6(2), e12629.
- DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, 62(4), 736–752.
- DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science*, 9(3), 205–210.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. doi: 10.1037/0033-295X.114.4.864
- Fischer, K. (2016). *Designing speech for a recipient: The roles of partner modeling, alignment, and feedback in so-called “simplified registers”*. John Benjamins.
- Fischer, K. (2006). *What computer talk is and isn’t. Human–computer conversation as intercultural communication*. AQ-Verlag.
- Fischer, K. (2011). *Interpersonal Variation in Understanding Robots as Social Actors*. Paper presented at the 2011 6th ACM/IEEE International Conference on Human–Robot Interaction (HRI), Lausanne, Switzerland.
- Fischer, K. (2021). Tracking anthropomorphizing behavior in human–robot interaction. *Journal of Human–Robot Interaction*, 11(1), Article 4. doi: 10.1145/3442677

- Fischer, K., Baumann, T., Langedijk, R., Jelinek, M., Manoopong, P., Lakshadeep, N., ... Palinko, O. (2021). Deliverable 1.2d: Update on user experiments. Report in the framework of the SMOOTH: Seamless huMan-robot interactiOn fOR THE support of elderly people project.
- Flavell, J. H., Flavell, E. R., Green, F. L., & Korfmacher, J. E. (1990). Do young children think of television images as pictures or real objects? *Journal of Broadcasting & Electronic Media*, 34(4), 399–419.
- Friedman, B., Kahn Jr, P. H., & Hagman, J. (2003). *Hardware Companions? What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA.
- Garvey, C. (1990). *Play*. Harvard University Press.
- Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.
- Glas, D. F., Minato, T., Ishi, C. T., Kawahara, T., & Ishiguro, H. (2016). *Erica: The ERATO Intelligent Conversational Android*. Paper presented at the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, USA.
- Goldstein, T. R., & Bloom, P. (2015). Characterizing characters: How children make sense of realistic acting. *Cognitive Development*, 34, 39–50.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619. doi: 10.1126/science.1134475
- Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 171(1024), 279–296.
- Gregory, R. L. (1970). *The intelligent eye*. McGraw-Hill.
- Gregory, R. L. (2005). The Medawar Lecture 2001 knowledge for vision: Vision for knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1458), 1231–1251.
- Gross, J. J., Fredrickson, B. L., & Levenson, R. W. (1994). The psychophysiology of crying. *Psychophysiology*, 31, 460–468.
- Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9(1), 87–108.
- Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, 75(4), 624–628.
- Holler, J., Kendrick, K. H., Casillas, M., & Levinson, S. C. (2016). *Turn-taking in human communicative interaction*. Frontiers Media.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329.
- Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, 48(2), 303–314.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Korzybski, A. (1948). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. International Non-Aristotelian Library.
- Langedijk, R., & Fischer, K. (2023). *Persuasive Robots in the Field*. Paper presented at the Persuasive'23 Conference, Eindhoven, Netherlands.
- Lee, M. K., Kiesler, S., & Forlizzi, J. (2010). *Receptionist or Information Kiosk: How Do People Talk with a Robot?* Paper presented at the Annual ACM Conference on Computer Supported Cooperative Work, Savannah, Georgia, USA.
- Logan, D. E., Breazeal, C., Goodwin, M. S., Jeong, S., O'Connell, B., Smith-Freedman, D., ... Weinstock, P. (2019). Social robots for hospitalized children. *Pediatrics*, 144(1), e20181511.
- Maynard, P. (1994). Seeing double. *Journal of Aesthetics and Art Criticism*, 52(2), 155–167.
- Mead, R., & Mataric, M. J. (2016). Perceptual models of human-robot proxemics. In Hsieh, M., Khatib, O., & Kumar, V. (Eds.), *Experimental robotics. Springer tracts in advanced robotics* (Vol. 109, pp. 261–276). Springer.
- Melson, G. F., Kahn Jr, P. H., Beck, A., & Friedman, B. (2006). *Toward Understanding Children's and Adults' Encounters with Social Robots*. Paper presented at the AAAI Workshop on Human Implications of Human-Robot Interaction (HRI), Boston, Massachusetts, USA.
- Menne, I. M., & Schwab, F. (2018). Faces of emotion: Investigating emotional facial expressions towards a robot. *International Journal of Social Robotics*, 10(2), 199–209.
- Mieczkowski, H., Liu, S. X., Hancock, J., & Reeves, B. (2019). *Helping Not Hurting: Applying the Stereotype Content Model and Bias Map to Social Robotics*. Paper presented at the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea.
- Miller, G. A. (1993). Images and models, similes and metaphors. In Ortony, A. (Ed.), *Metaphor and thought* (pp. 357–400). Cambridge University Press.
- Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3), 273–284.
- Moore, R. K. (2017). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In Jokinen, K., & Wilcock, G. (Eds.), *Dialogues with social robots: Enablers, analyses, and evaluation. Lecture notes in electrical engineering* (Vol. 427, pp. 281–291). Springer.
- Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *Footing in Human-Robot Conversations: How Robots might Shape Participant Roles using Gaze Cues*. Paper presented at the 4th ACM/IEEE International Conference on Human-Robot Interaction, La Jolla, California, USA.
- Nanay, B. (2018). Threefoldness. *Philosophical Studies*, 175(1), 163–182.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27(10), 864–876. <https://doi.org/10.1111/j.1559-1816.1997.tb00275.x>
- Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.
- Oatley, K. (2011). *Such stuff as dreams: The psychology of fiction*. John Wiley.
- Oatley, K. (2016). Fiction: Simulation of social worlds. *Trends in Cognitive Sciences*, 20(8), 618–628.
- Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems*, 7(3), 1–40.
- Phillips, E., Ullman, D., de Graaf, M. M., & Malle, B. F. (2017). *What does a Robot Look Like? A Multi-site Examination of User Expectations about Robot Appearance*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Austin, Texas, USA.
- Piaget, J. (1962). *Play, dreams and imitation in childhood*. Routledge & Kegan Paul.
- Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., & Heath, C. (2009). "The First Five Seconds": Contingent Stepwise Entry into an Interaction as a Means to Secure Sustained Engagement in HRI. Paper presented at the 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan.
- Reeves, B., Hancock, J., & Liu, S. X. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior*, 1(1), 1–15.
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1), 17–34. doi: 10.1007/s12369-012-0173-8
- Rottenberg, J., Ray, R. D., & Gross, J. J. (2007). Emotion elicitation using films. In J. A. Coan & J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 9–29). Oxford University Press.
- Ruijten, P. A., Bouten, D. H., Rouschop, D. C., Ham, J., & Midden, C. J. (2014). *Introducing a Rasch-type Anthropomorphism Scale*. Paper presented at the 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Bielefeld, Germany.
- Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: Testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics*, 11(3), 477–494.
- Ruijten, P. A. M. (2015). *Responses to human-like artificial agents*. Uitgeverij BOXPress.
- Sacks, H., Schegloff, I., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, 50, 696–735.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.
- Seibt, J. (2017). Towards an ontology of simulated social interaction: Varieties of the "as if" for robots and humans. In Hakli, R., & Seibt, J. (Eds.), *Sociality and normativity for robots* (pp. 11–39). Springer.
- Seo, S. H., Geiskovitch, D., Nakane, M., King, C., & Young, J. E. (2015). *Poor Thing! Would You Feel Sorry for a Simulated Robot? A Comparison of Empathy toward a Physical and a Simulated Robot*. Paper presented at the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, Oregon, USA.
- Severson, R. L., & Woodard, S. R. (2018). Imagining others' minds: The positive relation between children's role play and anthropomorphism. *Frontiers in Psychology*, 9, 2140.
- Skolnick, D., & Bloom, P. (2006). What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction. *Cognition*, 101(1), B9–B18.
- Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, 5, 15924.
- Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006). Encounters with Kismet and Cog: Children respond to relational artifacts. In P. Messaris & L. Humphreys (Eds.), *Digital media: Transformations in human communication* (pp. 313–330). Peter Lang.
- Van Berkum, J. J. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science*, 17(6), 376–380.
- Van Berkum, J. J. A. (2009). The neuropragmatics of "simple" utterance comprehension: An ERP review. In Sauerland, U., & Yatsushiro, K. (Eds.), *Semantics and pragmatics: From experiment to theory* (pp. 276–316). Palgrave-Macmillan.
- Walton, K. L. (1973). Pictures and make-believe. *The Philosophical Review*, 82(3), 283–319.
- Walton, K. L. (1978). Fearing fictions. *The Journal of Philosophy*, 75(1), 5–27.
- Walton, K. L. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.
- Walton, K. L. (2008). *Marvelous images: On values and the arts*. Oxford University Press.
- Walton, K. L. (2015). *In other shoes: Music, metaphor, empathy, existence*. Oxford University Press.

- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010a). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388.
- Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010b). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, 99(3), 410–435.
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43), 11374–11379. doi: 10.1073/pnas.1704347114
- Yang, S., Mok, B., Sirkin, D., & Ju, W. (2015). *Adventures of an Adolescent Trash Barrel*. Paper presented at the 10th Annual ACM/IEEE International Conference on Human–Robot Interaction (HRI), Portland, Oregon, USA.
- Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science*, 8(1), 15–18.
- Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, 18(5), 229–234.

## Open Peer Commentary

### The Dorian Gray Refutation

Christoph Bartneck 

Department of Computer Science and Software Engineering, University of Canterbury, 8140 Christchurch, New Zealand  
[christoph.bartneck@canterbury.ac.nz](mailto:christoph.bartneck@canterbury.ac.nz)  
<https://www.bartneck.de>

doi:10.1017/S0140525X22001662, e22

#### Abstract

Theories are an integral part of the scientific endeavour. The target article proposes interesting ideas for a theory on human–robot interaction but lacks specificity that would enable us to properly test this theory. No empirical data are yet available to determine its predictive power.

I wish Clark and Fischer (C&F) had given their theory a name because that would make it easier for us to talk about it. For the time being, let's call their theory the Clark and Fischer Conjecture (CFC). I then have to wonder how the CFC is better than other theories, such as the media equation. To better understand this question, we have to consider the role that theories play.

Constructing theories is an essential process in the scientific endeavour. Theories help to explain the past and predict the future. There are several criteria available to judge the value of a theory. Arguably, the most important criterion is its explanatory power.

The more observations a theory is able to accurately model, the higher its value. A theory, such as gravity, that applies to all things is considered more powerful than a theory that only applies to robots. A theory that applies to all robots is more valuable than a theory that only applies to social robots and so forth. Achieving a higher generalisability often requires the use of more abstract terms in the theory. This does, at times, lead to situations where researchers engage in discussion about semantics rather than about models of reality. The relationships of these terms are then ideally expressed using maths.

Another success criterion for a theory is the accuracy and reliability of its explanations and predictions. Lastly, the simplicity of the theory itself makes it preferable over others. Occam's Razor

dictates that a theory that uses fewer concepts to model reality is preferable.

Before we start a discussion of the merits of the CFC we have to acknowledge that Nass and Reeves (1996) would probably not agree to C&F's representation of the media equation as “media = real life.” This is an over-simplification. We all know that movies are just movies. While they still have the power to make us cry, we know that they are just a representation. We have little trouble experiencing this cognitive dissonance. Hence media is not exactly the same as real life. C&F seem to have employed the rhetorical straw-man technique to highlight the need for a better theory. This is unnecessary, as there is little doubt that human–robot interaction (HRI) is a new form of media that requires further attention. HRI was only in its infancy when the media equation was proposed.

Is the CFC better than other theories, such as the media equation? At this stage we cannot say. There are no experiments available yet that have demonstrated that the CFC succeeds over other theories based on the success criteria mentioned above.

This leads us to one of the main challenges of this paper. I struggle with fully understanding the CFC because of its complexity. It considers a large number of concepts such as the varieties of depictions, perspectives on depictions, character types, imagination in depictions, frames of reference, layers of activity, authorities, and emotions. Many of these concepts are then further subdivided. Varieties of depictions, for example, is subdivided into static versus dynamic depictions, staged versus interactive depictions, and actor versus prop depictions. This part of the CFC alone could be considered a theory worthwhile of testing. With these large number of concepts it seems daring to come up with a mathematical expression for the complete CFC.

Unless we have such a mathematical expression of the CFC it will be difficult, if not impossible, to construct empirical studies to test the CFC. It could even be argued that an experiment that tests all the aspects of the CFC would be so complex that it would become highly impractical to test.

How then can we ever know that the CFC is true? How can we ever know that the CFC has more explanatory power than the theories, such as the media equation? How can we know that the predictions of the CFC's are more accurate and reliable? We may never know conclusively. What we do know is that the CFC is complex. Far more complex than “media = real life.”

This complexity is not only because of the number of concepts involved, but also because of unspecific relationships between them. The most we can learn from this paper is that concept A somehow relates to concept B. What we miss are more precise predictions, such as  $A = 2 \times B$ . This level of specificity is necessary to fully understand the concepts and their relationships.

This does not mean that the CFC has no merit. Formulating theories is important. Most studies in HRI dissect concepts into ever finer slices of reality that they then study in isolation. Far less effort is made towards bringing all these concepts back into an overarching theory. C&F should be applauded for their effort, even if their conjecture will be revised and extended in the future.

C&F remind us of concepts that are likely to play an important role in HRI. We should consider the robot itself, its representation, and the interaction context. What exact influence they have, however, remains unclear. It would have been desirable if C&F would have given us more clues on how we can test their conjecture.

I would like to end with a quote from Oscar Wilde who fails to lose relevance even after all these years. In his novel, *Dorian Gray*



states that, “... no theory of life seemed to him to be of any importance compared with life itself. He felt keenly conscious of how barren all intellectual speculation is when separated from action and experiment.” Let’s bring CFC closer to action and experiment by constructing systematic studies that will shed light on its concepts and their relationships.



**Financial support.** No funding was associated with this commentary.

**Competing interest.** None.

## Reference

Nass, C., & Reeves, B. (1996). *The media equation*. SLI Publications, Cambridge University Press.

# Trait attribution explains human–robot interactions

Yochanan E. Bigman<sup>a</sup> , Nicholas Surdel<sup>b</sup> and  
Melissa J. Ferguson<sup>b</sup> 

<sup>a</sup>The Hebrew University Business School, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel and <sup>b</sup>Department of Psychology, Yale University, New Haven, CT 06520-8205, USA

[yochanan.bigman@mail.huji.ac.il](mailto:yochanan.bigman@mail.huji.ac.il)

[nicholas.surdel@yale.edu](mailto:nicholas.surdel@yale.edu)

[melissaj.ferguson@gmail.com](mailto:melissaj.ferguson@gmail.com)

<https://ybigman.wixsite.com/ybigman>

<https://www.linkedin.com/in/nsurdel>

[www.fergusonlab.com](http://www.fergusonlab.com)

doi:10.1017/S0140525X22001509, e23

## Abstract

Clark and Fischer (C&F) claim that trait attribution has major limitations in explaining human–robot interactions. We argue that the trait attribution approach can explain the three issues posited by C&F. We also argue that the trait attribution approach is parsimonious, as it assumes that the same mechanisms of social cognition apply to human–robot interaction.

C&F propose that humans understand social robots as depictions rather than actual agents. This perspective focuses on the psychologically under-explained duality with which people can react to entities as simultaneously (or alternatively) agentic and non-agentic. We disagree that the trait attribution approach cannot handle the three questions that C&F identify. We argue that the trait attribution approach, based on decades of research on social cognition, can explain these issues, and variance in human–robot interaction more broadly. Moreover, this approach is parsimonious in that it assumes the same psychological processes that guide human–human interaction guide human–robot interaction.

## Addressing the criticisms of the trait attribution approach

The first limitation according to C&F is that the trait attribution approach cannot explain individual differences in willingness to interact with social robots. However, this can be explained by

research demonstrating individual differences in attributing human-like characteristics to other humans (Haslam & Loughnan, 2014) and to nonhumans (Waytz, Cacioppo, & Epley, 2010). The same processes that affect trait attributions to other humans can explain the willingness to interact with robots. For example, people vary in how much humanness (e.g., agency, experience) they attribute to outgroup members (e.g., Krumhuber, Swiderska, Tsankova, Kamble, & Kappas, 2015), pets (e.g., McConnell, Lloyd, & Buchanan, 2017), and fictional characters (e.g., Banks & Bowman, 2016). This variability emerges across individuals, situations (e.g., Smith et al., 2022), and within interactions (e.g., Haslam, 2006). According to the trait attribution approach, similar individual and situational factors can predict when people respond to a robot in a human-like way.

The second issue that C&F identify is a change in the way people interact with social robots within an interaction. But considerable research and theory in psychology suggest that the way an interaction unfolds is dynamically affected by many factors, such as perceived traits, goals, and abilities (e.g., see Freeman, Stolier, & Brooks, 2020). For example, the accessibility of stereotypes and goals can change over a relatively short amount of time (e.g., Ferguson & Bargh, 2004; Kunda, Davies, Adams, & Spencer, 2002; Melnikoff & Bailey, 2018). The inherently dynamic context of an interaction, with constantly varying types of information being introduced verbally and nonverbally, predicts changing attributions of one’s interaction partner, whether human or robot. The same trait attribution principles that guide human interactions can be used to explain the change in perspective in human–robot interactions.

The third unresolved question raised by C&F is selectivity – people notice some of the robots’ capabilities but not others. The trait attribution approach aligns with work in social cognition suggesting that people are more sensitive to some kinds of information than others, depending on individual differences and situational factors (e.g., Brewer, 1988; Fiske & Neuberg, 1990). For example, people positively evaluate competence in others, unless the other is immoral (Landy, Piazza, & Goodwin, 2016). Although much is not yet known about precisely *which* aspects of an interaction or agent are considered relevant and *when*, we argue that these basic principles of psychology can explain the characteristic of selectivity in trait inferences about social robots. Note that the social depictions approach also cannot explain exactly which aspects will be influential when, and for whom.

## Advantages and limitations of trait attribution

In addressing these three points we suggest that the trait attribution approach can explain phenomena that C&F argue are inconsistent with it. By showing how the same principles that explain human–human interaction can explain human–robot interaction, we argue that trait attribution is a parsimonious approach to explaining human–robot interactions.

The trait attribution approach is a broad concept; different research lines focus on different types of attributions to explain human–robot interactions. Anthropomorphism, for example, affects how much people trust robots such as self-driving cars (Waytz, Heafner, & Epley, 2014). People’s reliance on algorithms for tasks hinges on their perception that they have human-like emotions (Castelo, Bos, & Lehmann, 2019). People’s aversion to algorithms making moral decisions depends on the mind they attribute to them (Bigman & Gray, 2018), and their resistance to medical algorithms is based on attributing to them an inability



to appreciate human uniqueness (Longoni, Bonezzi, & Morewedge, 2019). Similarly, people's diminished outrage at discrimination by algorithms is a result of perceiving algorithms as less prejudiced than humans (Bigman, Gray, Waytz, Arnestad, & Wilson, 2022). The attribution approach to studying human-robot interactions extends beyond the strict attribution of only traits.

One possible limitation of the trait attribution approach is that it cannot explain the apparent intentionality of the duality of some human-robot interactions. That is, people seem to at times *knowingly* suspend their disbelief and cycle between treating a robot as agentic versus non-agentic. Although the trait approach can potentially explain going back and forth in attributions of agency, it does not address the role of intentionality in (and awareness of) doing so, and more research on the importance of this characteristic would be helpful. Moreover, it is an open question when people will interact with robots as actual social agents rather than depictions of social agents. We agree that sometimes people might interact with social robots as depictions, but that does not mean that they always do so. One untested possibility is that the more mind a robot is perceived as having, the less likely people are to treat it as a depiction. To the extent that robots are increasingly more agentic, trait attribution approach parsimoniously explains interactions where people interact with social robots as "real" agents rather than depictions: loving an artificial agent (Oliver, 2020), thinking an AI is sentient when it displays sophisticated language and conversation (Allyn, 2022), and feeling bad about punishing them (Bartneck, Verbunt, Mubin, & Al Mahmud, 2007).

C&F assume that social cognition of humans is unique, and cannot be applied to nonhuman entities. We argue that social cognition is broad. Humans as targets of social cognition share the space with other entities, even if they have a special place in it. By our account, the difference between the social cognition of humans and the social cognition of robots is mostly quantitative, not qualitative.

**Financial support.** This work was supported by Office of Naval Research N00014-19-1-2299, *Modeling and planning with human impressions of robots*.

**Competing interest.** None.

## References

- Allyn, B. (2022). The Google engineer who sees company's AI as "sentient" thinks a chatbot has a soul. *NPR*. <https://www.npr.org/2022/06/16/1105552435/google-ai-sentient>
- Banks, J., & Bowman, N. D. (2016). Emotion, anthropomorphism, realism, control: Validation of a merged metric for player-avator interaction (PAX). *Computers in Human Behavior*, 54, 215–223. <https://doi.org/10.1016/j.chb.2015.07.030>
- Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007). To kill a mockingbird robot. In *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*, Washington, DC (pp. 81–87). <https://doi.org/10.1145/1228716.1228728>
- Bigman, Y., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, 21–34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- Bigman, Y., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*. Advance online publication. <http://dx.doi.org/10.1037/xge0001250>
- Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Scrull & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Erlbaum.
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825. <https://doi.org/10.1177/0022243719851788>
- Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology*, 87(5), 557–572. <https://doi.org/10.1037/0022-3514.87.5.557>
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier. [https://doi.org/10.1016/S0065-2601\(08\)60317-2](https://doi.org/10.1016/S0065-2601(08)60317-2)
- Freeman, J. B., Stoller, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Advances in experimental social psychology* (Vol. 61, pp. 237–287). Elsevier. <https://doi.org/10.1016/bs.aesp.2019.09.005>
- Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, 10(3), 252–264. [https://doi.org/10.1207/s15327957pspr1003\\_4](https://doi.org/10.1207/s15327957pspr1003_4)
- Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, 65(1), 399–423. <https://doi.org/10.1146/annurev-psych-010213-115045>
- Krumhuber, E. G., Swiderska, A., Tsankova, E., Kamble, S. V., & Kappas, A. (2015). Real or artificial? Intergroup biases in mind perception in a cross-cultural perspective. *PLoS ONE*, 10(9), e0137840. <https://doi.org/10.1371/journal.pone.0137840>
- Kunda, Z., Davies, P. G., Adams, B. D., & Spencer, S. J. (2002). The dynamic time course of stereotype activation: Activation, dissipation, and resurrection. *Journal of Personality and Social Psychology*, 82(3), 283–299.
- Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, 42(9), 1272–1290. <https://doi.org/10.1177/0146167216655984>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4), 629–650.
- McConnell, A. R., Lloyd, E. P., & Buchanan, T. M. (2017). Animals as friends: Social psychological implications of human-pet relationship. In M. Hojjat & A. Moyer (Eds.), *The psychology of friendship* (pp. 157–174). Oxford University Press.
- Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences*, 115(4), E592–E600.
- Oliver, M. (2020). *Inside the life of people married to robots*. Buzzworthy. <https://www.buzzworthy.com/meet-men-married-robots/>
- Smith, J. M., Pasek, M. H., Vishkin, A., Johnson, K. A., Shackelford, C., & Ginges, J. (2022). Thinking about God discourages dehumanization of religious outgroups. *Journal of Experimental Psychology: General*, 151(10), 2586–2603. <https://doi.org/10.1037/xge0001206>
- Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232. <https://doi.org/10.1177/1745691610369336>
- Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52, 113–117. <https://doi.org/10.1016/j.jesp.2014.01.005>

## Fictional emotions and emotional reactions to social robots as depictions of social agents

Jonas Blatter  and Eva Weber-Guskar 

Institute for Philosophy I, Ruhr University Bochum, Bochum 44801, Germany  
[jonas.blatter@ruhr-uni-bochum.de](mailto:jonas.blatter@ruhr-uni-bochum.de)  
[eva.weber-guskar@ruhr-uni-bochum.de](mailto:eva.weber-guskar@ruhr-uni-bochum.de)  
[https://www.ruhr-uni-bochum.de/philosophy/weber\\_guskar](https://www.ruhr-uni-bochum.de/philosophy/weber_guskar)

doi:10.1017/S0140525X22001716, e24

### Abstract

Following the depiction theory by Clark and Fischer we would expect people interacting with robots to experience *fictional* emotions akin to those toward films or novels. However, some people's emotional reactions toward robots display the motivational force typical to *non-fictional* emotions. We discuss this incongruity and offer two suggestions on how to explain it while maintaining the depiction theory.

Clark and Fischer's depiction theory is meant to give, among other things, an answer to the question of why we respond to social robots emotionally while knowing that they are no real

social agents. Their claim is that this is most likely because we view them not as real social agents but as merely depicting social agents and that our emotions are directed only to the depicted layer. It seems they understand these emotions as cases of what others have called *fictional emotions* (Gendler, 2010; Medina, 2013; Teroni, 2019; Vendrell Ferran, 2022). These are emotions experienced toward characters or situations that we know are imaginary or fictional, such as the fear of the monster in a scary movie, compassion with the characters in a tragic novel, or the vicarious joy when seeing the protagonists victorious at the end of a play. In all these cases, we know that these characters are fictional, but having followed their stories we feel emotions that are very similar to the ones we would feel for real people.

Fictional emotions differ from non-fictional emotions in some respects, most crucially in their action-motivating force: Emotions are closely related to certain types of actions or behavior. Fear, for example, disposes the subject to generally avoid or want to avoid the object of their fear, anger is associated with confrontational and aggressive behavior, compassion with consoling the grieving person, and so on. This motivational force (sometimes called action tendencies, or action readiness) is usually understood to be a central feature of emotions (Frijda, 1986, 1988; Lazarus, 1991). In the case of fictional emotions, however, the motivational force is strongly reduced or takes on a subdued form (Gendler, 2010; Walton, 1990). For instance, despite feeling fear while watching Kubrick's *The Shining*, we might tense up in our seat, but we do not flee from the cinema, nor do we try to console the *Angel of Grief* chiseled in stone. Thus, if we could really only experience fictional emotions toward depicted social agents, we should expect the same kind of reduced action tendencies or motivational force as in emotions directed at social robots. The authors acknowledge this claim in section 9.2 and they explain the fact by a “[c]ompartimentalization of emotions”:

“Suppose Amy sees a forklift operator run into Ben and severely injure his arm. She would surely fear for Ben's health, rush to his aid, and call for an ambulance. If she saw the same happen to Asimo, she would do none of that. She would take her time in contacting Asimo's principal about the damage, [...]” (target article, sect. 9.2, para. 4)

The authors assume that, in this scenario we would observe that while Amy might experience a certain degree of fear for the depicted Asimo<sub>char</sub>, she would not display the same action tendency as with a genuine social agent such as Ben. But this is just a conjecture on how a person might react and not an observation of actual behavior. Taking into account several empirical studies, this conjecture seems not very well justified. Strong social emotional responses to robots have been documented in many cases (Darling, 2017): Beginning with people feeling gratefulness toward Roomba, their vacuum cleaner (Sung, Guo, Grinter, & Christensen, 2007), over others refraining from hitting, switching off (Bartneck, Van Der Hoek, Mubin, & Al Mahmud, 2007), or destroying a robot (Darling, 2021, Ch. 10) to soldiers who risk their lives in order to save the robots they are working with (Singer, 2009, Ch. 17) or to bury and hold a funeral for a defect mine-disposal robot (Garber, 2013; Garreau, 2007). The interactive and immediate nature of robots seems to elicit social emotions with the motivational force typical of non-fictional emotions, which we would not expect in the case of a fictional emotion toward other forms of depiction. A person in Amy's

situation would probably never feel the urge to rush to help a depicted agent in a painting, novel, or a movie, but might feel to urge to help Asimo out of fear for him (although she would probably still call a technician rather than an ambulance). It seems, therefore, that the emotions we experience toward social robots can take on a stronger motivational force than we would expect from fictional emotions, which we experience toward other forms of depiction.

Here are two suggestions on how we could explain this apparent discrepancy between the fictional status of the social agent depicted by robots and the motivational force emotions toward them can have, without giving up the depiction theory. First, while it might be possible for people to keep the three perspectives distinct on a cognitive level, they might fail to keep them separated on an emotional level. Moreover, keeping emotional reactions to objects on the three perspectives distinct might be easier with forms of depiction that are more spatially and temporally distant and less interactive. Emotional overreactions might be strongest with embodied, physically present depictions. Several studies suggest that participants tend to respond with more empathy (Seo, Geiskkovitch, Nakane, King, & Young, 2015), afford greater trust to (Bainbridge, Hart, Kim, & Scassellati, 2011), and report a stronger feeling of social presence (Lee, Jung, Kim, & Kim, 2006) toward physically present robots compared to telepresent or simulated ones. Second, emotions might be triggered by features other than the depicted social agent. A robot's parts might additionally depict bodily features of a human or animal. Seeing such depicted bodily parts being damaged might also elicit emotional responses, without requiring the depiction of a social agent. Emotions are often said to work on the level of perception and are somewhat – but not entirely – inaccessible to higher-level cognitive penetration (Döring, 2007, 2008; Goldie, 2000; Tappolet, 2016). If this is the case, then we should expect that the intellectually demanding work of keeping depiction and reality separate might sometimes fail to translate to the emotional level.

**Acknowledgments.** We thank Tobias Starzak and acknowledge the INTERACT! project for their useful feedback.

**Financial support.** This is a publication in the context of the project INTERACT!, funded by the ministry of culture and science of North Rhine Westphalia.

**Competing interest.** None.

## References

- Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1), 41–52. <https://doi.org/10.1007/s12369-010-0082-7>
- Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). “Daisy, Daisy, Give Me Your Answer Do!” Switching off a Robot. *2nd ACM/IEEE International Conference on Human-Robot Interaction*, Washington, DC, pp. 217–222.
- Darling, K. (2017). Who's Johnny? Anthropomorphic framing in human-robot-interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney & R. Jenkins (Eds.), *Robot ethics 2.0* (pp. 173–188). Oxford University Press.
- Darling, K. (2021). *The new breed: What our history with animals reveals about our future with robots*. Henry Holt.
- Döring, S. A. (2007). Seeing what to do: Affective perception and rational motivation. *Dialectica*, 61(3), 363–394. <https://doi.org/10/ff8c8p>
- Döring, S. A. (2008). Conflict without contradiction. In G. Brun, U. Dogluoglu & D. Kuenzle (Eds.), *Epistemology and emotions* (pp. 83–103). Ashgate.
- Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
- Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5), 349–358. <https://doi.org/ckq>

- Garber, M. (2013). Funerals for fallen robots: New research explores the deep bonds that can develop between soldiers and the machines that help keep them alive. *The Atlantic*, September 20.
- Garreau, J. (2007). Bots on the ground in the field of battle (or even above it), robots are a soldier's best friend. *Washington Post*, May 6.
- Gendler, T. S. (2010). Genuine rational fictional emotions. In T. S. Gendler (Ed.) *Intuition, imagination, and philosophical methodology* (pp. 227–237). Oxford University Press. <https://doi.org/fn7mcd>
- Goldie, P. (2000). *The emotions: A philosophical exploration*. Oxford University Press.
- Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, 46(4), 352–367. <https://doi.org/czd6kp>
- Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human–Computer Studies*, 64(10), 962–973. <https://doi.org/10.1016/j.ijhcs.2006.05.002>
- Medina, J. (2013). An enactivist approach to the imagination: Embodied enactments and “fictional emotions”. *American Philosophical Quarterly*, 50(3), 317.
- Seo, S. H., Geiskovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor Thing! Would You Feel Sorry for a Simulated Robot? A Comparison of Empathy toward a Physical and a Simulated Robot. *10th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, Portland, OR, pp. 125–132.
- Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. Penguin.
- Sung, J., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). “My Roomba is Rambo”: Intimate Home Appliances. *Ubicomp*, pp. 145–162.
- Tappolet, C. (2016). *Emotions, values, and agency*. Oxford University Press.
- Teroni, F. (2019). Emotion, fiction and rationality. *The British Journal of Aesthetics*, 59(2), 113–128. <https://doi.org/10.1093/aesthj/ayz015>
- Vendrell Ferran, I. (2022). Sham emotions, quasi-emotions or non-genuine emotions? Fictional emotions and their qualitative feel. In T. Breyer, M. Cavallaro & R. Sandoval (Eds.), *Phenomenology of phantasy and emotion* (pp. 231–259). WBG Academic.
- Walton, K. L. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.

## When Pinocchio becomes a real boy: Capability and felicity in AI and interactive depictions

John M. Carroll

College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA  
[jmcarroll@psu.edu](mailto:jmcarroll@psu.edu); <https://jcarroll.ist.psu.edu>

doi:10.1017/S0140525X22001479, e25

### Abstract

Clark and Fischer analyze social robots as *interactive* depictions, presenting characters that people can interact with in social settings. Unlike other types of depictions, the props for social robot depictions depend on emerging interactive technologies. This raises questions about how such depictions depict: They conflate character and prop in ways that delight, confuse, mistreat, and may become ordinary human–technology interactions.

Clark and Fischer (C&F) characterize social robots as autonomous agents that nonetheless act on the authority of and are “ultimately controlled by principals” – defined as the other agents who designed, manufactured, or administer the social robot. They note that controls for social robots are typically not transparent, meaning a robot's design does not convey to its human users that or how it is controlled by its principals.

Indeed, social robots seem designed deliberately to misrepresent their epistemic states and conceal their limitations. C&F cite this greeting from the robot Smooth: “I wonder if you would like something to drink.” Is the robot empathetic about someone's thirst? Is it wondering what drinking is like? The resulting vagueness may be instrumental in encouraging imagination and emotional projection in humans; what C&F call the “pretense” of depictions: People acting toward the depiction as if the depiction is what it depicts.

The pretense of depiction can be fun: Chatty Cathy (character depicted) says “I love you,” and at the same place and time, the doll (prop depicting the character) plays a recorded phrase when the “chatty ring” in its upper back is pulled. The pretense invites confusion about (apparently) autonomous behavior and the nature of authorities in the background, but it is moderated by the obvious ring in the doll's back. For Smooth the robot, the projected and imagined *character* autonomously approaches people and offers a drink, but the physical embodiment of the robot is a *prop* that executes instructions a human created. This is a subtle and evocative distinction.

Outside playful pretense, things are more complex. Facilitating human confusion about actual capabilities of a robot creates ethical problems. Thus, if someone were to deduce from dialog interaction that Smooth can wonder about things, and that it understands relationships and experiences inherent in thirst and drinking, then they have been deceived. If someone expects a robot math tutor to teach them math but it introduces erroneous definitions, they have been educationally harmed. C&F observe that in such misfire scenarios people would sue the manufacturer or programmer, not the robot. But whoever is sued, people would have been mistreated. Ethical problems of this sort are not new; Weizenbaum (1976) was shocked at how susceptible and tenacious people were in experiencing his mid-1960s script-based chatbot ELIZA as human.

In their discussion of agents acting on the authority of other, principal agents, C&F conflate cases where the agent acting on another agent's authority (called a rep-agent) is a social robot with cases in which the rep-agent is a human playing a job role. But these cases are different. Thus, Clark and his hotel booking agent briefly divert their phone call to high school reminiscence, which also allows them to pursue enabling goals of building trust and common ground. The agent can do this because the character of the booking agent is projected if and when she chooses. Social robots as rep-agents never act on their own authority; their capabilities for goals and actions are too limited. However much they conceal it, they *must* be controlled by principals.

In the future, challenges around interactive depictions may become more complex, arising not only from exaggeration of limited capabilities, but also from ethically presenting capabilities that rival or might even exceed those of humans, and have only tenuous causal connection to remote (human) authorities. For example, the language model GPT-3 (third-generation pretrained transformer) is a learning machine with a gargantuan text base; it responds broadly to language prompts (Dale, 2021). It can compose original sonnets in the style of Shakespeare, develop program code from natural language specifications, and create news articles and philosophy essays, among many other applications. It carries out these tasks at human levels.

These capabilities raise many challenges already; GPT-3 might soon carry out the rep-agent depiction roles enumerated by C&F better than the humans currently employed to perform them. Like



ELIZA and social robots, GPT-3 can convey self-awareness without necessarily having it. At the least, it is the best simulacrum yet (Tiku, 2022).

The rapid emergence of interactive possibilities for artificial intelligence (AI) applications has helped to highlight “explainable AI” (Holzinger et al., 2022). AI systems can be smart enough to carry out significant cognitive tasks but lack capability to insightfully and effectively explain how and why they do what they do. For interactive AI, the explanation required is not an execution trace or design rationale, it is conversational explanation of the sort people expect from their responsible interlocutors (Carroll, 2022). C&F’s human rep-agents provide a model for how agents work together to conversationally explain role-related conduct. Robots have responsibility to explain themselves effectively or to clearly convey that they cannot (cf. the chatty ring).

Hancock (2022) argues that humans and emerging AI systems could experience *time* very differently. He mocks the expression “real time” as indicative of how humans uncritically privilege a conception of time scaled to human perception and cognition. Emerging AI systems might think through and carry out complex courses of action without humans noticing anything happened. The very transition from contemporary AI to quite autonomous and self-aware agents might occur in what humans would experience as “a single perceptual moment.” Hancock worries this could entail a Skynet Armageddon, but it seems more likely to result in greater diversity for AI systems, some of whose capabilities are occasionally unclear, even to them. Humans are very skilled at building and coordinating common ground with others, as when Clark reminisces about high school with the hotel agent. This enables the development of trust and fluent interaction. Future robots must effectively coordinate common ground with humans, and humans must reciprocally depict themselves as responsible and empathetic.

In the 1883 children’s novel, *Pinocchio* is a wooden marionette, a puppet depicting a boy (Collodi, 1883). Through the novel, Pinocchio encounters challenges, and often behaves too reflexively, without much planning or empathy for others. Ultimately though, he becomes more responsible and empathetic. Through the intervention of a fairy, Pinocchio becomes a real boy. The novel ends there, but we might think that is where the more interesting story begins.

**Financial support.** This work was supported by the National Institutes of Health R01LM013330-03.

**Competing interest.** None.

## References

- Carroll, J. M. (2022). Why should humans trust AI? *ACM Interactions*, 29(4), 73–77. <https://dl.acm.org/doi/10.1145/3538392>
- Collodi, C. (1883). *The adventures of Pinocchio*. Libero.
- Dale, R. (2021). GPT-3: What’s it good for?. *Natural Language Engineering*, 27(1), 113–118. <https://doi.org/10.1007/978-3-031-04083-2>
- Hancock, P. A. (2022). Avoiding adverse autonomous agent actions (with peer commentary). *Human–Computer Interaction*, 37(3), 211–236. <https://doi.org/10.1080/07370024.2021.1970556>
- Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., & Samek, W. (Eds.) (2022). *xxAI-Beyond explainable AI*. Springer. <https://doi.org/10.1007/978-3-031-04083-2>
- Tiku, N. (2022). The Google engineer who thinks the company’s AI has come to life. *Washington Post*. June 11. <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lambda-blake-lemoine/>
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.

## Autonomous social robots are real in the mind’s eye of many

Nathan Caruana<sup>a</sup>  and Emily S. Cross<sup>a,b,c,d</sup> 

<sup>a</sup>School of Psychological Science, Macquarie University, Sydney, NSW, Australia;

<sup>b</sup>Centre for Elite Performance, Expertise and Training, Macquarie University, Sydney, NSW, Australia; <sup>c</sup>Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK and <sup>d</sup>MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Westmead, NSW, Australia

[nathan.caruana@mq.edu.au](mailto:nathan.caruana@mq.edu.au)

[emily.cross@mq.edu.au](mailto:emily.cross@mq.edu.au)

<https://nathancaruaana.weebly.com/>

<https://www.soba-lab.com/>

doi:10.1017/S0140525X22001625, e26

### Abstract

Clark and Fischer’s dismissal of extant human–robot interaction research approaches limits opportunities to understand major variables shaping people’s engagement with social robots. Instead, this endeavour categorically requires multidisciplinary approaches. We refute the assumption that people cannot (correctly or incorrectly) represent robots as autonomous social agents. This contradicts available empirical evidence, and will become increasingly tenuous as robot automation improves.

Clark and Fischer (C&F) claim that people represent social robots as *depictions* of other humans, and not as autonomous social entities. We argue this framework for understanding human perceptions of – and interactions with – robots is limited and limiting. Instead, an eclectic approach drawing upon psychology, social neuroscience, and human–robot interaction (HRI) will best serve empirical progress as robots’ social capabilities evolve.

We agree that for some people, and in particular contexts, certain robots can be seen as representing the intentions and actions of a human principal (e.g., operator/engineer). Our central argument, however, is that such a framework for understanding HRIs is not universal and may become irrelevant as increasingly intelligent and autonomous social robots are realised.

In serving their claim, the authors draw upon Wizard-of-Oz approaches commonly used in HRI research (where a person teleoperates a robot) to categorise robots alongside ventriloquist dummies as examples of “interactive” depictions, which are a step above “staged” (e.g., puppets) and “static” (e.g., statues) depictions within their taxonomy. However, the fact that a robot “depicts” human intentions/actions *in reality* does not mean people perceive it as such. An overlooked feature of the Wizard-of-Oz approach is the use of Turing deception, in which people believe the robot operates autonomously (Kelley, 1984). We argue that under many circumstances, humans *do* perceive social robots as autonomous, intentional agents, even when they are not. In fact, several studies have demonstrated direct consequences on subjective experiences, behaviour and neural processing during interactions with virtual agents and robots depending on whether or not people believe an agent is human-controlled (Caruana & McArthur, 2019; Caruana, Spirou, & Brock, 2017; Cross, Ramsey, Liepelt, Prinz, & de C Hamilton,



2016; Schellen & Wykowska, 2019). These studies (and others) show that under some conditions, people represent artificial agents as human depictions, and sometimes not. Thus, it remains unclear how the depictions framework can resolve such findings if human depictions are always in play. If the “variety” or “proximity” of the human depiction shapes experiences with robots, this requires further specification.

This also highlights the important roles played by knowledge and beliefs in shaping robot representations; that is, what a person *believes* a robot can do, which is often quite separate from what a robot *can actually do*. The authors touch on the issue of pretense, and correctly state that what children know and believe about robots is unclear. We would go further to state that this holds for people of *all ages*, highlighting another HRI research challenge unresolved by the depictions framework. Specifically, it is reasonable to assume that most people have clearly defined, relatively invariant, top-down knowledge cues concerning puppets’ or ventriloquist dummies’ autonomy and sentience. The presence of strings and/or the close proximity of the ventriloquist offer bottom-up cues that activate knowledge of these agents being directly operated by their human “principal.” The same, however, cannot be said of social robots, whose relationship with their principal(s) can be far more distant, ambiguous, opaque, or complex – especially upon first encounter. Social robots are also more novel and varied than puppets or ventriloquist dummies, and continue to evolve as technologies develop, further fuelling this ambiguity. The depictions framework does not accommodate for this ambiguity in robot agency, nor the variability in the kinds of cues humans rely on to resolve it. Further, many individuals are naive about the current state of robot capabilities, or biased by representations of autonomous social robots in popular media (cf. Cross & Ramsey, 2021). Indeed, our own recent research encourages the hypothesis that some children may have rather realistic ideas about the autonomy and limitations of robots (Caruana, Moffat, Blanco, & Cross, 2022). Furthermore, many contexts exist for applying socially assistive robots (e.g., education, health, and aged care) where users may be more likely to overestimate robots’ autonomy, and perhaps less likely to see them as depictions of humans (e.g., young children, the elderly, and individuals with intellectual disabilities).

Rather than focusing on questions of depiction and how clearly or accurately people associate a robot with its human engineer(s), we argue that thornier challenges arise from issues related to variability in HRI across (1) *individual differences* (e.g., personality, knowledge, attribution styles, education, cognitive ability); (2) *robot form* (e.g., zoomorphic, mechanoid, humanoid, size, composition), and *function* (e.g., verbal, mobile, expressive); and (3) *application domains*. Together, this considerable variation and complexity presents deep challenges to building a robust knowledge base related to social encounters between humans and robots (Cross & Ramsey, 2021), and it remains unclear how this is resolved within the depictions framework. We argue that this problem requires a multidisciplinary, eclectic approach receptive to insights gained from the previous approaches that C&F summarise and dismiss.

We would further argue that these approaches have not been fully represented in this review, and that they continue to bear fruit in explaining variability across HRIs. For instance, the review

of the “trait attribution” approach loosely references Epley, Waytz, Akalis, and Cacioppo’s (2008) concepts of “elicited knowledge” and “effectance motivation” for explaining why humans may ascribe human-like agency/intentions to objects. However, key to this approach is the idea that significant *individual differences* exist in people’s tendencies to anthropomorphise, which these factors attempt to explain (e.g., Neave, Jackson, Saxton, & Hönekopp, 2015). Another overlooked component of Epley et al.’ framework concerns “sociality motivation.” This refers to one’s drive to be socially connected to others, and is argued to interact with the abovementioned factors to influence anthropomorphism, while also being influenced by other contextual or dispositional factors (e.g., subjective loneliness, social isolation, anxiety, personality, culture, etc.). While we fully acknowledge that none of the extant approaches for understanding HRI offers complete explanations for the “social artifact” problem, they nonetheless offer useful frameworks for understanding how some variables shape people’s interactions with robots. They also continue to inspire new empirical questions that advance our knowledge of the factors that shape HRIs. To us, it remains unclear how the depictions framework offers a solution to the inadequacies of extant approaches, or hypotheses that will help advance the field.


**Financial support.** Dr. Caruana was supported by a Macquarie University Research Fellowship (9201701145). Professor Cross was supported in part by funding from the European Research Council (ERC) under the EU Horizon 2020 research and innovation programme (grant agreement 677270) and the Leverhulme Trust (PLP-2018-152). The funding sources for this project were not involved in any decision-making, data collection, analysis, or dissemination of this study. The scientific process of this study was independent of any relationship with the funding sources that could be construed as a potential conflict of interest.

**Competing interest.** None.

## References

- Caruana, N., & McArthur, G. (2019). The mind minds minds: The effect of intentional stance on the neural encoding of joint attention. *Cognitive, Affective & Behavioral Neuroscience*, 19(6), 1479–1491. <https://doi.org/10.3758/s13415-019-00734-y>
- Caruana, N., Moffat, R., Blanco, A. M., & Cross, E. S. (2022). Perceptions of intelligence & sentience shape children’s interactions with robot reading companions: A mixed methods study. *PsyArXiv*. <https://doi.org/10.31234/osf.io/7t2w9>
- Caruana, N., Spirou, D., & Brock, J. (2017). Human agency beliefs influence behaviour during virtual social interactions. *PeerJ*, 5, e3819. <https://doi.org/10.7717/peerj.3819>
- Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human-machine interactions. *Trends in Cognitive Sciences*, 25(3), 200–212.
- Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W., & de C Hamilton, A. F. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 371(1686), 20150075. <https://doi.org/10.1098/rstb.2015.0075>
- Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, 26(2), 143–155. doi:10.1521/soco.2008.26.2.143
- Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1), 26–41.
- Neave, N., Jackson, R., Saxton, T., & Hönekopp, J. (2015). The influence of anthropomorphic tendencies on human hoarding behaviours. *Personality and Individual Differences*, 72, 214–219. <https://doi.org/10.1016/j.paid.2014.08.041>
- Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots – Open questions and methodological challenges. *Frontiers in Robotics and AI*, 5, 139. <https://doi.org/10.3389/frobt.2018.00139>

## Cues trigger depiction schemas for robots, as they do for human identities

Elliott K. Doyle  and Sara D. Hodges

Department of Psychology, University of Oregon, Eugene, OR 97403-1227, USA  
 edoyle3@uoregon.edu  
 sdhodes@uoregon.edu; <https://psychology.uoregon.edu/profile/sdhodes>

doi:10.1017/S0140525X22001650, e27

### Abstract

Clark and Fischer's three levels of depiction of social robots can be conceptualized as cognitive schemas. When interacting with social robots, humans shift between schemas similarly to how they shift between identity category schemas when interacting with other humans. Perception of mind, context cues, and individual differences underlie perceptions of which level of depiction is most situationally relevant.

Social environments are complex. To navigate them, we use simplified scaffolding information, called schemas, built from our past experiences (Macrae & Cloutier, 2009). Often, schemas focus on social identity categories, and contain stereotypes – simple, categorical, automatically arising predictors about what someone will be like (Hammond & Cimpian, 2017) – about those identities. Any given individual has many identities, each of which might be differently salient from context to context, and so different assumptions about the same individual will come to mind more readily depending on the situation (Oyserman, 2015; Shih, Pittinsky, & Ambady, 1999).

We propose that “robot” is an identity category that comprises three subschemas, delineated in Clark and Fischer's (C&F's) work as three levels of depiction. Each subschema evokes different types of behavior, but *which* is evoked as most relevant can fluctuate, just as one's perception of another person's most relevant identity might. If this is true, then people's individual variation in when and whether they approach robots as characters, depictions of social agents, or pieces of machinery is likely because of the same reasons stereotypes about any kind of identity are variably activated.

One underlying impetus for switching between these schemas, we contend, is the degree to which people perceive the robot as having a mind. Human beings assume things about each other's minds in order to communicate effectively – a task that is vital for social interaction, but very complex. Despite understandable objections about the overuse of stereotypes, particularly negative stereotypes of minority groups, stereotypes facilitate communication by providing quick, and often accurate, predictions about what someone else might be thinking (Hodges & Kezer, 2021; Lewis, Hodges, Laurent, Srivastava, & Biancarosa, 2012). If a social robot is perceived as having a mind, people are more likely to interact with the robot as a character rather than as a machine, with “robot(ic)” being simply one of the stereotypes activated to describe it as a social entity, much like “teenager” or “doctor.”

Robots are often not perceived as having a mind (Gray, Gray, & Wegner, 2007), and in these instances *social* stereotypes do not

come into play. However, some things can cause people to ascribe more mind to a robot, such as the robot behaving in unexpected ways, the robot possessing human-like features, or the person who perceives the robot feeling particularly lonely (Epley, Waytz, & Cacioppo, 2007; Waytz, Gray, Epley, & Wegner, 2010). Excessive robotic attempts to copy human appearances perfectly can be unsettling (Gray, Knobe, Sheskin, Bloom, & Barrett, 2011; Gray & Wegner, 2012), but characteristics that allow the robot to express the things humans notice and communicate to each other – like attention and emotion – can facilitate perception of mind (Duffy, 2003). Given the right cues, anthropomorphism can occur automatically when the perceiver is presented with a situation in which treating a robot as a social agent is contextually appropriate (Kim & Sundar, 2012).

The characteristics of the human perceivers, therefore, are important in addition to the features of the social robot itself. Qualities like willingness to suspend disbelief (Duffy & Zawieska, 2012; Muckler, 2017) and tendency to anthropomorphize (Waytz, Cacioppo, & Epley, 2014) vary between people, and may make people more or less inclined to treat a robot like a character or like a machine. As delineated in C&F's example of the three human interactants encountering Smooth the robot, some people will readily engage socially with the same robot that others will not. This tendency to anthropomorphize is partly because of individual variation between people, but past experience and mindset likely play a role, too: People who are distracted by novel aspects of a social robot or focused on its non-humanness may be impeded in depicting the robot as a character, and by extension, in applying certain stereotypes that guide particular kinds of interactions with it. However, these effects are not unique to perceptions of robots. For example, encountering other humans in heavily scripted roles (e.g., flight attendant, nightclub bouncer) may lead us to evoke prop-like schemas that preclude character depictions. Cues that prompt thoughts of body counts or bodily actions may similarly interfere with character depiction, and evoke more mechanical schemas (Mooijman & Stern, 2016; Small & Loewenstein, 2003).

Social robots might have difficulty being perceived as genuinely plausible interaction partners in part because the features of the robot fail to activate the character-level stereotypes, such that the robot is stuck at depiction or machinery. Alternatively, some observers might be unwilling or unable to suspend their disbelief in order to interact with the robot like a character (which would, in turn, create a social situation in which others who might otherwise be willing to treat the robot anthropomorphically are made more self-conscious by their peers' reluctance). Finally, even robots depicted as characters might evoke stereotypes of robots being less socially capable than humans (Chan et al., 2020) because, for example, their language is less fluid. As we further explore the factors that promote the willingness and ease with which humans can interact with robots as social agents, we should also heed when robots mirror aspects of some *human* agents with whom interactions are problematic.

Our suggestion that the three levels of depiction that C&F outline provide three schemas for robots, each of which can be activated to bring to mind different stereotypes, offers a psychological explanation for how people are able to switch their focus between machinery, depiction, and character fluidly. As C&F note, humans have extensive experience engaging with depictions, which should help us construe social robots as depictions of social agents. Increasingly sophisticated robots should trigger

stereotypes of various different social agents, providing humans with further cognitive scaffolding to guide and elaborate interactions with robots. Additionally, humans also have experience engaging with what C&F call “nonstandard” (i.e., not real) characters from whom they seek and derive a number of very “human” yearnings (e.g., companionship, inspiration, perspective; see Gabriel & Young, 2011; Myers & Hodges, 2009; Taylor, Hodges, & Kohányi, 2003), suggesting a flexible, inclusive, and creative ability to connect with a wide range of social agents.


**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors

**Competing interest.** None.

## References

- Chan, L., Doyle, K., McElfresh, D., Conitzer, V., Dickerson, J. P., Schaich Borg, J., & Sinnott-Armstrong, W. (2020). Artificial intelligence: Measuring influence of AI “assessments” on moral decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, United States (pp. 214–220).
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3–4), 177–190.
- Duffy, B. R., & Zawieska, K. (2012). Suspension of disbelief in social robotics. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Wuhan, China (pp. 484–489). IEEE.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.
- Gabriel, S., & Young, A. F. (2011). Becoming a vampire without being bitten: The narrative collective-assimilation hypothesis. *Psychological Science*, 22(8), 990–994.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–619.
- Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Barrett, L. F. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology*, 101(6), 1207–1220.
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125–130.
- Hammond, M. D., & Cimpian, A. (2017). Investigating the cognitive structure of stereotypes: Generic beliefs about groups predict social judgments better than statistical beliefs. *Journal of Experimental Psychology: General*, 146(5), 607–614.
- Hodges, S. D., & Kezer, M. (2021). It is hard to read minds without words: Cues to use to achieve empathic accuracy. *Journal of Intelligence*, 9(2), 27.
- Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless?. *Computers in Human Behavior*, 28(1), 241–250.
- Lewis, K. L., Hodges, S. D., Laurent, S. M., Srivastava, S., & Biancarosa, G. (2012). Reading between the minds: The use of stereotypes in empathic accuracy. *Psychological Science*, 23(9), 1040–1046.
- Macrae, C. N., & Cloutier, J. (2009). A matter of design: Priming context and person perception. *Journal of Experimental Social Psychology*, 45(4), 1012–1015.
- Mooijman, M., & Stern, C. (2016). When perspective taking creates a motivational threat: The case of conservatism, same-sex sexual behavior, and anti-gay attitudes. *Personality and Social Psychology Bulletin*, 42(6), 738–754.
- Muckler, V. C. (2017). Exploring suspension of disbelief during simulation-based learning. *Clinical Simulation in Nursing*, 13(1), 3–9.
- Myers, M. W., & Hodges, S. D. (2009). Making it up and making do: Simulation, imagination and empathic accuracy. In K. Markman, W. Klein & J. Suhr (Eds.), *The handbook of imagination and mental simulation* (pp. 281–294). Psychology Press.
- Oyserman, D. (2015). Identity-based motivation. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social sciences* (pp. 1–11). John Wiley & Sons. <https://doi.org/10.1002/9781118900772.etrds0171>
- Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, 10(1), 80–83.
- Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, 26(1), 5–16.
- Taylor, M., Hodges, S. D., & Kohányi, A. (2003). The illusion of independent agency: Do adult fiction writers experience their characters as having minds of their own?. *Imagination, Cognition and Personality*, 22(4), 361–380.
- Waytz, A., Cacioppo, J., & Epley, N. (2014). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232.
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388.

## People treat social robots as real social agents

Alexander Eng<sup>a</sup>, Yam Kai Chi<sup>a</sup> and Kurt Gray<sup>b</sup> 

<sup>a</sup>Department of Management & Organization, National University of Singapore, Singapore 119245, Singapore and <sup>b</sup>Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-8100, USA

[aeng@u.nus.edu](mailto:aeng@u.nus.edu)

[bizk@nus.edu.sg](mailto:bizk@nus.edu.sg); <https://bizfaculty.nus.edu.sg/faculty-details/?proflid=452>

[kurtgray@unc.edu](mailto:kurtgray@unc.edu); <https://www.kurtjgray.com>

doi:10.1017/S0140525X22001534, e28

### Abstract

When people interact with social robots, they treat them as real social agents. How people depict robots is fun to consider, but when people are confronted with embodied entities that move and talk – whether humans or robots – they interact with them as authentic social agents with minds, and not as mere representations.

Haunted houses employ actors who pretend to be werewolves and zombies. Visitors wander through the darkness, listening for creatures lying in wait, and then scream as the actors reach out to touch them. If you asked visitors whether they thought the werewolves were real before and after touring the attraction, they would laugh and say no. They understand that the actors are just “depicting social agents.” But to what extent does the concept of “depictions of social agents” matter when they are confronted with a werewolf chasing them through dark corridors? Very little. Research on construal level theory suggests that people can differentiate depictions from reality when psychological distance is high, but when people are directly experiencing a situation, depictions feel real – and are treated as real (Liberman, Trope, & Stephan, 2007; Trope & Liberman, 2010).

Research finds that – in real life – people also treat robots as actual social agents, not as mere depictions of social agents. This suggests that the idea of “depictions of social agents” may not be useful when considering people’s actual interactions with social robots. It is an interesting exercise to think about “depictions” in the pages of journal articles, but empirical evidence often suggests otherwise when people are immersed in their interactions with robots.

To illustrate the idea of depictions, Clark and Fischer (C&F) use the example of movies, distinguishing between agents (actors) and depictions (the characters they play), like Leonardo DiCaprio and Kate Winslet playing Jack Dawson and Rose Bukater in *Titanic*. But movies are not a good analogy for robots, because robots are *embodied* social agents unlike characters on the other side of the screen. Embodiment – having a physical presence – fundamentally changes how we interact with agents. Like the werewolf in the haunted house, it makes them *real* agents.

Importantly, even with movie characters, people often fail to distinguish between actors and the characters they play. In an analysis of hit T.V. series *Breaking Bad*, researchers found that the fictional character “Skyler is often merged with Anna Gunn, the actor playing her...[people] do not always make a clear distinction between Gunn and the fictional character of Skyler, who become a single entity” (Hermes & Stoete, 2019, p. 412).



If people distinguished “robots as social depictions” from “robots as social agents” in real life, then they would have no trouble turning robots off, even if they pleaded for their lives. But people do have trouble. In Bartneck and Hue’s (2008) replication of Milgram’s obedience study, experimenters instructed participants to switch off an anthropomorphized robotic cat which they had been interacting with, informing them that this would wipe its memories and personality. The robotic cat pleaded with participants, saying “*You are not really going to switch me off, are you?*” In contrast to C&F’s theorizing, people started bargaining with the robots, saying things like, “*No! I really have to do it now, I’m sorry!*” or “*But it has to be done!*” People treat the robot as a true social agent, not as a mere painting of a robot.

More evidence that robots are real social agents come from Qin et al.’s (2022) replication of the classic Asch conformity experiment, in which they used a social robot confederate. As with human confederates, people bowed to the social pressure of a robot misreading the length of a line.

The distinction between depictions and social agents becomes even more insubstantial in practice as robots become more realistic: The more lifelike robots become, the more we treat them like social agents themselves, not mere depictions. For example, Zhao and Malle (2022) find that people respond to new stimuli (human-like robots) in the same way that they respond to familiar stimuli (humans) if both stimuli closely resembled one another (Guttman & Kalish, 1956; Shepard, 1987). Likewise, Yam et al. (2022) found that people were more likely to act spitefully to robot supervisors who delivered negative feedback when those robots were more human-like. There’s no reason to retaliate to mere social depictions.

People – in real life, with real-life robots – treat robots as real agents and not social depictions. But C&F are correct that people see differences between robots and humans. But the difference is not about depictions, but rather about *mind*. Mind perception theory (Gray, Gray, & Wegner, 2007) suggests that we perceive the minds of social agents along two distinct dimensions, agency (thinking and doing) and experience (feeling and sensing). We perceive humans as having high agency and high experience, animals as having low agency but high experience, and social robots as having moderate agency and low-to-moderate experience (Gray & Wegner, 2010).

These perceptions of mind are important – especially in real life. Perceptions of mind underlie whether people treat robots as legitimate moral decision maker (Bigman, Waytz, Alterovitz, & Gray, 2019) – a machine with the capacity for agency and experience is seen as more qualified to make life-and-death medical and military decisions (Bigman & Gray, 2018).

Changing perceptions of mind also change how people interact with social robots. Reducing a social robot’s perceived capacity for experiencing feelings decreases the uncanniness of human-like robots (Yam, Bigman, & Gray, 2021b). On the flip side, a study at the world’s only all-robot-staffed hotel found that increasing a service robot’s perceived capacity for experiencing feelings makes people like service robots more – and forgive them more after service failures (Yam et al., 2021a).

Robots are not human beings, but neither are they mere depictions of social agents. Instead, they are seen as *real* social agents, especially when people interact with them. The reality of in-person “depictions” is something designers of both robots and haunted houses understand; we scholars also need to understand this fact.

**Competing interest.** None.

## References

- Bartneck, C., & Hue, J. (2008). Exploring the abuse of robots. *Interaction Studies*, 9(3), 415–433. <https://doi.org/10.1075/is.9.3.04bar>
- Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181, 21–34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, 23(5), 365–368. <https://doi.org/10.1016/j.tics.2019.02.008>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (American Association for the Advancement of Science)*, 315(5812), 619–619. <https://doi.org/10.1126/science.1134475>
- Gray, K., & Wegner, D. M. (2010). Blaming god for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14(1), 7–16. <https://doi.org/10.1177/1088868309350299>
- Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, 51(1), 79–88.
- Hermes, J., & Stoete, L. (2019). Hating Skyler White: Audience engagement, gender politics and celebrity culture. *Celebrity Studies*, 10(3), 411–426. <https://doi.org/10.1080/19392397.2019.1630155>
- Liberman, N., Trope, Y., & Stephan, E. (2007). Psychological distance. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 353–381). Guilford Press.
- Qin, X., Chen, C., Yam, K. C., Cao, L., Li, W., Guan, J., ... Lin, Y. (2022). Adults still can’t resist: A social robot can induce normative conformity. *Computers in Human Behavior*, 127, 107041.
- Shepard, R. N. (1987). Towards a universal theory of generalization for psychological science. *Science*, 237(4820), 1317–1323.
- Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, 117(2), 440–463. <https://doi.org/10.1037/a0018963>
- Yam, K. C., Bigman, Y., & Gray, K. (2021b). Reducing the uncanny valley by dehumanizing humanoid robots. *Computers in Human Behavior*, 125, 106945. <https://doi.org/10.1016/j.chb.2021.106945>
- Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021a). Robots at work: People prefer- and forgive-service robots with perceived feelings. *Journal of Applied Psychology*, 106(10), 1557–1572. doi:10.1037/apl0000834
- Yam, K. C., Goh, E., Fehr, R., Lee, R., Soh, H., & Gray, K. (2022). When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology*, 102, 104360. <https://doi.org/10.1016/j.jesp.2022.104360>
- Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition*, 224, 105076–105076. <https://doi.org/10.1016/j.cognition.2022.105076>

## Taking a strong interactional stance

Frank Förster<sup>a</sup> , Frank Broz<sup>b</sup>  and Mark Neerincx<sup>b</sup> 

<sup>a</sup>Adaptive Systems Research Group, University of Hertfordshire, Hatfield AL10 9AB, UK and <sup>b</sup>Interactive Intelligence Research Group, Delft University of Technology, 2628 XE Delft, Netherlands

[f.foerster@herts.ac.uk](mailto:f.foerster@herts.ac.uk)

[F.Broz@tudelft.nl](mailto:F.Broz@tudelft.nl)

[mark.neerincx@tno.nl](mailto:mark.neerincx@tno.nl)

<https://frank-foerster.gitlab.io>

<https://www.tudelft.nl/en/eemcs/the-faculty/departments/intelligent-systems/>

[interactive-intelligence/people/current-group-members/frank-broz](https://www.tudelft.nl/en/eemcs/the-faculty/departments/intelligent-systems/people/current-group-members/frank-broz)

[https://www.tudelft.nl/en/eemcs/the-faculty/departments/intelligent-systems/](https://www.tudelft.nl/en/eemcs/the-faculty/departments/intelligent-systems/people/current-group-members/mark-a-neerincx)

[interactive-intelligence/people/current-group-members/mark-a-neerincx](https://www.tudelft.nl/en/eemcs/the-faculty/departments/intelligent-systems/people/current-group-members/mark-a-neerincx)

doi:10.1017/S0140525X22001455, e29

### Abstract

We outline two points of criticism. Firstly, we argue that robots do constitute a separate category of beings in people’s minds rather than being mere depictions of non-robotic characters. Secondly, we find that (semi-)automatic processes underpinning communicative interaction play a greater role in shaping robot-directed speech than Clark and Fischer’s theory of social robots as depictions indicate.



We formulate two points of criticism regarding Clark and Fischer's (C&F's) contribution and suggest that common research practices in human–robot interaction contribute to reinforcing confusion about robot capabilities by obfuscating the nature of the interaction with an agent or prop.

Firstly, we argue that robots do exist as a separate class of entity in people's minds even before they encounter an actual robot in real life. This mental model that varies amongst people is likely because of their exposure to fictional depictions of robots in popular media. People know and expect that a robot dog or a humanoid robot is a different kind of entity than a dog or a person. They are unclear on the actual capabilities of these agents, but they can and will discover this through interaction, which makes robots distinct from noninteractive depictions such as static art or characters in noninteractive performances. Research methodology in human–robot interaction, for example, a widespread use of Wizard-of-Oz experimental designs, and a lack of transparency about the level of a robot's autonomy reinforces this ambiguity about capabilities. C&F present a virtual agent or a ventriloquist's dummy as similar examples of agents. But we argue that these agents engage in very different types of interactions, where in one case the agent being interacted with is an autonomous computer program and in the other the interaction is with another person through the use of a prop with the human controlling this prop being visible and known to their interaction partner.

Secondly, C&F underplay the influence of (semi-) automatic processes on the concrete trajectory and form of an interaction because of this conflation of interactive and noninteractive formation of understanding of agents or characters. While a person's speech style initially may be influenced by depictions as construed by the authors, the affordances and real-time contingencies of the unfolding interaction will substantially impact upon that person's style of talk. Some of these real-time adaptations are automatic (such as gaze in face-to-face conversation, Broz, Lehmann, Nehaniv, & Dautenhahn, 2012) and may "pull" the unfolding interaction in a direction different to the one set up by the person's pre-existing views of the robot's role or nature.

In support of this view are the following transcripts originating from the negation acquisition studies conducted by Förster, Saunders, Lehmann, and Nehaniv (2019). These studies consisted of multiple sessions per participant, and the transcripts pertain both to participant P12 (P) teaching object labels to Deechee (D), a childlike humanoid robot that was presented to participants as a young language learner.

Session 2, 0 min 47 seconds

((P picks up heart object))

P-1 this one here is a heart

P-2 you don't like the shape

((P turns object around))

P-3 do you wanna see upside down

P-4 heart

((D turns head and frowns))

P-5 no don't like that [one]

Session 5, 1 min 20 seconds

((P picks up square, D reaches out for it))

D-1 square!

((D gets to hold object and drops it))

P-6 yeah square!

((P picks up triangle))

D-2 done!

P-7 no! (0.5 s) don't say done!

D-3 triangle

P-8 yeah, triangle! (..) [well done!]

(((P puts down triangle)))

((P picks up heart))

D-4 done

P-9 no, I say well done (.) you don't say done

P-10 what's this one?

D-5 heart!

P-12 yes

Participant P12, instructed to talk to Deechee as if it was a 2-year-old child, initially spoke in a style roughly compatible with child-directed speech. This included intent-related questions (P-3) and intent interpretations (P-2; cf. Förster, Saunders, & Nehaniv, 2018). During the second session, however, P12 decided to speak in a much simpler, "robotic" register, that he maintained during the two follow-up sessions and into his fifth session. In this register he used mostly one-word utterances that consisted either of object labels or short feedback words, for example, P-6 and P-8. This change, as we learned later, was meant to optimize the learning outcome of the – by P12 – hypothesized learning algorithm such that his mental model of the robot was arguably one of a mere machine. However, once Deechee started to use negation words such as "no" or "done" (D-2 and D-4), P12 did not manage to maintain his linguistic restraint and abandoned his minimalistic speech style for short time periods (e.g., P-7 and P-9).

Given P12's strong adherence to his chosen minimalistic speech register prior to these lapses, these utterances appear to have a somewhat involuntary character. We argue that these lapses were caused by automatic processes temporarily gaining the upper hand over the conscious, self-imposed restrictions. The "pull" of the interaction caused the participant to treat Deechee, at least temporarily, as a being with wants or emotions. This change is because of Deechee's behaviour-in-interaction rather than a unilateral perspective switch in terms of class of depiction (cf. Förster & Althoefer, 2021). In terms of being seen as a depiction of another character it is unclear what that could possibly be in this setting. Deechee does not serve any distinct social role such as receptionist nor does it correspond to a known character such as Kermit the frog.

For social robots to be useful in their intended roles, they must become (and be understood as) social agents in and of themselves rather than puppets that experimenters act through to investigate people's incorrect mental models. This will necessarily involve people coming to understand their capabilities and limitations through multiple and prolonged interactions. More generally, the application of data-driven machine learning technology in successive human–robot collaborative activities will involve co-adaptation and co-learning. Such new emergent behaviours may comprise unconscious tangible interactions (Van Zoelen, Van Den Bosch, & Neerincx, 2021a) and new collaboration patterns (Van Zoelen, Van Den Bosch, Rauterberg, Barakova, & Neerincx, 2021b). This way, the human develops cognitive, affective, and tangible experiences and understandings of the robots, grounded in the pursuing situated collaborations. In addition to the "pre-baked" designs (Lighthart et al., 2019), anthropomorphic projections (Carpenter, 2013), and human-like collaboration functions (Neerincx et al., 2019), the evolving unique robot features with corresponding behaviours will affect the continuous (re-)construction of new types of robot characters.


**Financial support.** The cited transcripts originate from work that was supported by the EU Integrated Project "Integration and Transfer of Action and Language in Robots" through the European Commission under Contract FP-7-214668.

**Competing interest.** None.

## References

- Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012). Mutual Gaze, Personality, and Familiarity: Dual Eye-Tracking during Conversation, 2012 IEEE RO-MAN. The 21st IEEE International Symposium on Robot and Human Interactive Communication, 9–13 September, 2012, Paris, France, pp. 858–864. <https://doi.org/10.1109/ROMAN.2012.6343859>
- Carpenter, J. (2013). The Quiet Professional: An investigation of US military Explosive Ordnance Disposal personnel interactions with everyday field robots. Doctoral dissertation. University of Washington. ResearchWorks Archive. <http://hdl.handle.net/1773/24197>
- Förster, F., & Althoefer, K. (2021). Attribution of autonomy and its role in robotic language acquisition. *AI & Society*, 37(2), 605–617. <https://doi.org/10.1007/s00146-020-01114-8>
- Förster, F., Saunders, J., Lehmann, H., & Nehaniv, C. L. (2019). Robots learning to Say “No”: Prohibition and rejective mechanisms in acquisition of linguistic negation. *ACM Transactions on Human–Robot Interaction*, 8(4), 1–26. <https://doi.org/10.1145/3359618>
- Förster, F., Saunders, J., & Nehaniv, C. L. (2018). Robots that say “No”: Affective symbol grounding and the case of intent interpretations. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3), 530–544. <https://doi.org/10.1109/TCDS.2017.2752366>
- Ligthart, M., Fernhout, T., Neerinx, M. A., van Bindsbergen, K. L., Grootenhuis, M. A., & Hindriks, K. V. (2019). A Child and a Robot Getting Acquainted – Interaction Design for Eliciting Self-Disclosure. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems, 13–17 May, Montreal, Canada , pp. 61–70.
- Neerinx, M. A., Van Vught, W., Blanson Henkemans, O., Oleari, E., Broekens, J., Peters, R., ... Bierman, B. (2019). Socio-cognitive engineering of a robotic partner for child’s diabetes self-management. *Frontiers in Robotics and AI*, 6. <https://doi.org/10.3389/frobot.2019.00118>
- Van Zoelen, E. M., Van Den Bosch, K., & Neerinx, M. (2021a). Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning. *Frontiers in Robotics and AI*, 8, 692811. <https://doi.org/10.3389/frobot.2021.692811>
- Van Zoelen, E. M., Van Den Bosch, K., Rauterberg, M., Barakova, E., & Neerinx, M. (2021b). Identifying interaction patterns of tangible co-adaptations in human-robot team behaviors. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.645545>

## Unpredictable robots elicit responsibility attributions

Matija Franklin<sup>a</sup>, Edmond Awad<sup>b</sup>, Hal Ashton<sup>c</sup> and David Lagnado<sup>a</sup> 

<sup>a</sup>Experimental Psychology Department, University College London, London WC1E 6BT, UK; <sup>b</sup>Economics Department, University of Exeter, Exeter EX4 4PU, UK and <sup>c</sup>Computer Science Department, University College London, 66-72 Gower Street, London WC1E 6EA, UK  
[matija.franklin@ucl.ac.uk](mailto:matija.franklin@ucl.ac.uk); <https://www.ucl.ac.uk/pals/research/experimental-psychology/person/matija-franklin>  
[e.awad@exeter.ac.uk](mailto:e.awad@exeter.ac.uk); <https://www.edmondawad.me>  
[ucabha5@ucl.ac.uk](mailto:ucabha5@ucl.ac.uk); <https://algotinent.com/>  
[d.lagnado@ucl.ac.uk](mailto:d.lagnado@ucl.ac.uk); <https://www.ucl.ac.uk/pals/research/experimental-psychology/person/david-lagnado/>

doi:10.1017/S0140525X22001546, e30

### Abstract

Do people hold robots responsible for their actions? While Clark and Fischer present a useful framework for interpreting social robots, we argue that they fail to account for people’s willingness to assign responsibility to robots in certain contexts, such as when a robot performs actions not predictable by its user or programmer.

Autonomous machines are increasingly used to perform tasks traditionally undertaken by humans. With little or no human insight, these machines make decisions that significantly impact people’s lives. Clark and Fischer (C&F) argue that people conceive of social robots as depictions of social agents. They differentiate between the “base scene” – representing the physical materials the robot is made from, the “depictive scene” representing the robot’s recognizable form along with an interpretive authority, and the “scene depicted” which either transports people into an imagined world inhabited by the robot or imports the robot’s imagined character into the real world. We argue that this framework fails to account for people’s willingness to assign responsibility to social robots (and AI more generally). Specifically, we argue that in a range of cases people assign some degree of responsibility to social robots, and do not shift all responsibility to the “authority” that uses the robot. These cases include robots that behave in novel ways not predictable by their users or programmers. We also argue that responsibility attribution is not a finite resource; thus users and robots can simultaneously be held responsible.

Recent work (Tobia, Nielsen, & Stremitzer, 2021) explores the question of who is held responsible for the actions of autonomous machines. Experimental evidence suggests that people are willing to attribute blame or praise to robots as agents in their own right (Ashton, Franklin, & Lagnado, 2022; Awad et al., 2020; Franklin, Awad, & Lagnado, 2021). As agents, autonomous machines are sometimes treated differently from humans. For example, people tend to hold humans accountable for their intentions while holding machines accountable for the outcomes of their actions (Hidalgo, Orghian, Canals, De Almeida, & Martin, 2021). Further, people ascribe more extreme intentions to humans while only ascribing narrow intentions to machines. This is a puzzle for the depiction framework because it shows that people are prepared to attribute responsibility to *depictions* of agents as well as to the depiction’s authority.

C&F argue that attributing responsibility to a depiction’s authority is intuitive for the ventriloquist’s dummy or a limited social robot like Asimo. However, the examples they list in Table 2 concern those whose behavior is largely predictable, at least by the authority. Recent technological advances have produced social robots capable of generating original behavior not conceived even by their creators (Woodworth, Ferrari, Zosa, & Riek, 2018). Using machine learning methods, modern social robotics learn human preferences by observing human behavior in various contexts, developing adaptive robot behavior which is tailored to the user (Wilde, Kulić, & Smith, 2018). The mechanisms by which they reach their decisions are opaque, complex, and not directly encoded by the creator. We propose that such social robots are more likely to elicit responsibility attributions in their own right.

Perceived increases in machine autonomy come with increases in attributed responsibility toward those machines. First, higher machine autonomy is associated with intent inferences toward machines becoming more like humans (Banks, 2019). Thus research shows that when robots are described as autonomous, participants attribute responsibility to them nearly as much as they do to humans (Furlough, Stokes, & Gillan, 2021). Additionally, more autonomous technologies decrease the perceived amount of control that the authority has over them, which in turn decreases the credit the authority receives for positive outcomes (Jörlling, Böhm, & Paluch, 2019). Similarly, drivers of manually controlled vehicles are deemed more responsible than the drivers of automated vehicles (McManus & Rutchick, 2019).

Furthermore, C&F's assertion that the creator of the depiction is responsible for the interpretation of their depictions relies on the fact that the depiction's behavior is predictable by the creator. The authors write: "We assume that Michelangelo was responsible not only for carving David, but for its interpretation as the biblical David" (target article, sect. 8, para. 1). But this argument fails for machines that behave unpredictably. When the painting "Edmond De Belamy," generated by a deep learning algorithm, sold at an art auction for \$432,500, many credited the machine (Christie's, 2018). This attribution to machine creativity goes beyond anecdotal evidence (Epstein, Levine, Rand, & Rahwan, 2020). Similarly, AlphaGo, in beating World Champion Go-player Lee Sedol, used novel strategies as adopted by human players (Chouard, 2016). Such novel moves prompted comments worldwide about machine creativity (McFarland, 2016), giving credit to AlphaGo rather than just DeepMind's team. While the DeepMind team intended AlphaGo to win the match, they did not envisage these novel moves.

Moreover, accounts of responsibility attribution should avoid committing the fixed-pie fallacy (Kaiserman, 2021) – the false assumption that there is a total amount of responsibility that can be allocated, or in other words, treating responsibility as a finite resource. The statement "when Ben interacts with Asimo, he would assume that there are authorities responsible for what Asimo<sub>char</sub> actually does..." (target article, sect. 8.1, para. 4) hints at this error. People are willing to attribute responsibility to both autonomous machines and their users (e.g., a self-driving car and the driver; Awad et al., 2020).

There are also strong normative arguments that go against this fixed-pie fallacy. Some argue that neither the creators nor the operators of autonomous machines should bear sole responsibility (Sparrow, 2007). Others have drawn parallels between artificial intelligence and group agency – usually assigned to large corporations – as both are nonhuman goal-directed actors (List, 2021). Even in the case of recent fatal autonomous car crashes, attribution of legal responsibility to the car's manufacturer has not proved as straightforward as C&F's model would predict (De Jong, 2020).

C&F present an insightful framework to cover predictable and pre-programmed social robots. Here we have argued that more intelligent, autonomous, and thus, unpredictable social robots exist today. People are willing to attribute responsibility to such robots for their mistakes (Ashton et al., 2022; Awad et al., 2020; Franklin, Ashton, Awad, & Lagnado, 2022). Further, for more anthropomorphized social robots, research suggests that people are even willing to attribute experiential mental states (Fiala, Arico, & Nichols, 2014). The framework thus needs to be extended to handle the more intelligent robots currently being produced, and normative theories in philosophy and law suggesting that social robots may need to share social responsibility.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.



**Competing interest.** None.

## References

- Ashton, H., Franklin, M., & Lagnado, D. (2022). Testing a definition of intent for AI in a legal setting. Unpublished manuscript.
- Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., ... Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, 4(2), 134–143.
- Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, 90, 363–371.

- Chouard, T. (2016). The go files: AI computer clinches victory against go champion. *Nature*, <https://doi.org/10.1038/nature.2016.19553>
- Christie's (2018). Is artificial intelligence set to become art's next medium? [Blog post]. Retrieved from <https://www.christies.com/features/a-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx>
- De Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*, 26(2), 727–735.
- Epstein, Z., Levine, S., Rand, D. G., & Rahwan, I. (2020). Who gets credit for AI-generated art?. *iScience*, 23(9), 101515.
- Fiala, B., Arico, A., & Nichols, S. (2014). You robot. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (1st ed., pp. 31–47). Routledge. <https://doi.org/10.4324/9780203122884>
- Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022). Causal Framework of Artificial Autonomous Agent Responsibility. In *Proceedings of 5th AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, Oxford, UK.
- Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *iScience*, 24(4), 102252.
- Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The influence of robot autonomy. *Human Factors*, 63(4), 592–602.
- Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.
- Jörling, M., Böhm, R., & Paluch, S. (2019). Service robots: Drivers of perceived responsibility for service outcomes. *Journal of Service Research*, 22(4), 404–420.
- Kaiserman, A. (2021). Responsibility and the "pie fallacy". *Philosophical Studies*, 178(11), 3597–3616.
- List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, 34(4), 1213–1242.
- McFarland, M. (2016). What AlphaGo's sly move says about machine creativity. The Washington Post, retrieved from [washingtonpost.com/news/innovations/wp/2016/03/15/what-alphagos-sly-move-says-about-machine-creativity/](https://www.washingtonpost.com/news/innovations/wp/2016/03/15/what-alphagos-sly-move-says-about-machine-creativity/)
- McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social Psychological and Personality Science*, 10(3), 345–352.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Tobia, K., Nielsen, A., & Stremitzer, A. (2021). When does physician use of AI increase liability?. *Journal of Nuclear Medicine*, 62(1), 17–21.
- Wilde, N., Kulić, D., & Smith, S. L. (2018). Learning User Preferences in Robot Motion Planning through Interaction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia (pp. 619–626). IEEE.
- Woodworth, B., Ferrari, F., Zosa, T. E., & Riek, L. D. (2018). Preference Learning in Assistive Robotics: Observational Repeated Inverse Reinforcement Learning. In *Machine Learning for Healthcare Conference*, Stanford University, USA (pp. 420–439). PMLR.

## The second-order problem of other minds

Ori Friedman<sup>a</sup>  and Arber Tasimi<sup>b</sup> 

<sup>a</sup>Department of Psychology, University of Waterloo, Waterloo N2L 3G1, Canada

and <sup>b</sup>Department of Psychology, Emory University, Atlanta, GA 30322, USA

[friedman@uwaterloo.ca](mailto:friedman@uwaterloo.ca)

[arber.tasimi@emory.edu](mailto:arber.tasimi@emory.edu)

<https://uwaterloo.ca/psychology/people-profiles/ori-friedman>

<http://psychology.emory.edu/home/people/faculty/tasimi-arber.html>

doi:10.1017/S0140525X22001443, e31

### Abstract

The target article proposes that people perceive social robots as depictions rather than as genuine social agents. We suggest that people might instead view social robots as social agents, albeit agents with more restricted capacities and moral rights than humans. We discuss why social robots, unlike other kinds of depictions, present a special challenge for testing the depiction hypothesis.



How will we know when a social robot (or any other kind of artificial intelligence) is a genuine social agent? That is, how will we know when it is conscious, feels things, and understands what it hears or says? This is the philosophical problem of other minds – the problem of how we can know that anyone else has a mind – applied to human creations (Harnad, 1991).

The target article raises a new problem of other minds. Clark and Fischer suggest that rather than viewing social robots as genuine social agents, people instead view them as depictions of social agents. Under this depiction account, people engage in a kind of pretense when interacting with social robots (also see Rueben, Rothberg, & Matarić, 2020). This account could be right, but we suggest it remains possible that people might instead view social robots as genuine social agents. Testing between these accounts introduces a new second-order problem of other minds: How can we tell if other people think they are dealing with a genuine social agent or a mere depiction of one?

The second-order problem of other minds may be difficult to resolve. When dealing with depictions, people normally hold back – their actions fall short from what they would do with the real thing. For example, children pretending to eat plastic fruit refrain from actually biting it (e.g., Leslie & Happé, 1989; Lillard, 1993) and filmgoers do not attempt to intervene in movie events. Do people also hold back with social robots? It might be hard to tell. Although people do not treat social robots exactly the way they treat their peers, this isn't saying much. There are many different kinds of agents and people see them as varying in their mental capacities (Gray, Gray, & Wegner, 2007; Weisman, Dweck, & Markman, 2017) and moral standing (Crimston, Hornsey, Bain, & Bastian, 2018; Goodwin, 2015). So, while it might be obvious when people hold back when dealing with many kinds of depictions (e.g., plastic fruit), this will be less obvious with social robots. What looks like holding back could turn out to reflect beliefs that social robots have limited capacities and moral standing.

To illustrate these points, let's consider the evidence offered as support for the idea that people view social robots as depictions. One line of evidence is that rather than seeing social robots the way they see their fellow humans, people see social robots as a kind of property. They affirm social robots can be sold, and if a social robot dented someone's car, the owner of the car would seek compensation from the robot's owner rather than from the robot itself. Treating social robots like property might follow from the belief that they are depictions rather than genuine social agents. But it is also reminiscent of how people treat real agents viewed as having limited moral standing or limited mental abilities. For example, pets and other animals are bought and sold and their owners are held liable when they cause harm (e.g., Bowman-Smith, Goulding, & Friedman, 2018; Nadler & McDonnell, 2011). Similar points may apply to how enslaved people were viewed and treated in the past. They too were treated as chattel, and when they caused harm, their enslavers were held liable in some legal systems (Oppenheim, 1940). But it is unlikely that people view pets as depictions, or that the enslaved were viewed this way either. So rather than viewing robots as mere depictions, people might instead see them as genuine agents with limited moral worth and limited mental capacities.

The target article also notes that people differ from one another in how they interact with social robots. Although some people converse with social robots, others refrain from doing so – these people do not respond to greetings from social robots and if they address robots at all, it is only with blunt questions and

brusque orders; perhaps these people are unwilling to play along with the pretense that these depictions are social agents. But this again is ambiguous. We might also expect differences between people if some believed that social robots are real agents, while others did not. Here again, people's treatment of animals raises questions. As with social robots, people vary in how they address their pets and some people's communication with their pet dogs is apparently limited to commands and threats (e.g., Carlisle-Frank, Frank, & Nielsen, 2004; Mitchell, 2004). Some talk to pets could have a pretend element – people sometimes ask dogs questions but then also answer the questions (Mitchell, 2004). But it seems unlikely (at least to us) that people view pets as depictions, or that most variation in talk to pets come down to differences in owners' proclivity to pretend.

Although the second-order problem of other minds be difficult to resolve, the difficulty may be asymmetric. While it might be difficult to confirm that social robots are viewed as depictions, it may be easier to confirm when they are viewed as genuine agents. Consider the issue of whether people show moral concern for robots (for a recent review see Harris & Anthis, 2021). When people express concerns for the welfare of robots and advocate for robots to have rights, this might suggest they view social robots as genuine agents – at least if these expressions of concern focus on robots themselves and not on side-concerns, such as concerns about property damage, or concerns that mistreating robots will encourage mistreatment of humans (e.g., Levy, 2009). By contrast, the absence of concern would not necessarily show that people view robots as depictions. It could instead stem from viewing robots as genuine agents with limited capacities or moral worth.

**Financial support.** This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada awarded to OF.


**Competing interest.** None.

## References

- Bowman-Smith, C. K., Goulding, B. W., & Friedman, O. (2018). Children hold owners responsible when property causes harm. *Journal of Experimental Psychology: General*, 147(8), 1191–1199. <https://doi.org/10.1037/xge0000429>
- Carlisle-Frank, P., Frank, J. M., & Nielsen, L. (2004). Selective battering of the family pet. *Anthrozoös*, 17(1), 26–42. <https://doi.org/10.2752/089279304786991864>
- Crimston, C. R., Hornsey, M. J., Bain, P. G., & Bastian, B. (2018). Toward a psychology of moral expansiveness. *Current Directions in Psychological Science*, 27(1), 14–19. <https://doi.org/10.1177/0963721417730888>
- Goodwin, G. P. (2015). Experimental approaches to moral standing. *Philosophy Compass*, 10(12), 914–926. <https://doi.org/10.1111/phc3.12266>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (New York, N.Y.)*, 315(5812), 619–619. <https://doi.org/10.1126/science.1134475>
- Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1(1), 43–54. <https://doi.org/10.1007/BF00360578>
- Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics*, 27(4), 1–95. <https://doi.org/10.1007/s11948-021-00331-8>
- Leslie, A. M., & Happé, F. (1989). Autism and ostensive communication: The relevance of metarepresentation. *Development and Psychopathology*, 1(3), 205–212. <https://doi.org/10.1017/S0954579400000407>
- Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics*, 1(3), 209–216. <https://doi.org/10.1007/s12369-009-0022-6>
- Lillard, A. S. (1993). Pretend play skills and the child's theory of mind. *Child Development*, 64(2), 348–371. <https://doi.org/10.1111/j.1467-8624.1993.tb02914.x>
- Mitchell, R. W. (2004). Controlling the dog, pretending to have a conversation, or just being friendly? Influences of sex and familiarity on Americans' talk to dogs during play. *Interaction Studies*, 5(1), 99–129. <https://doi.org/10.1075/is.5.1.06mit>
- Nadler, J., & McDonnell, M. H. (2011). Moral character, motive, and the psychology of blame. *Cornell Law Review*, 97, 255–304.

- Oppenheim, L. (1940). Law of slaves – A comparative study of the Roman and Louisiana systems. *Tulane Law Review*, 14, 384–406.
- Rueben, M., Rothberg, E., & Mataric, M. J. (2020). Applying the theory of make-believe to human–robot interaction. In M. Nørskov, J. Seibt & O. S. Quick (Eds.), *Culturally sustainable social robotics* (pp. 40–50). IOS Press. <https://doi.org/10.3233/FAIA200899>
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43), 11374–11379. <https://doi.org/10.1073/pnas.170434711>

## Interacting with characters redux

Richard J. Gerrig 

Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, USA  
[richard.gerrig@stonybrook.edu](mailto:richard.gerrig@stonybrook.edu)

doi:10.1017/S0140525X22001558, e32

### Abstract

Clark and Fischer (C&F) discuss how people interact with social robots in the context of a general analysis of interaction with characters. I suggest that a consideration of aesthetic illusion would add nuance to this analysis. In addition, I illustrate how people's experiences with other depictions of characters require adjustments to C&F's claims.

Clark and Fischer (C&F) make a compelling case that people experience social robots as depictions of social agents. I focus my comments on the claims they make in the section “Interacting with characters.” I suggest that their analysis requires some adjustments in the context of people's habitual responses to other types of depictions.

In their section on interacting with characters, C&F make brief mention of the concept of transportation (Clark & Van Der Wege, 2015; Gerrig, 1993). They suggest that “we imagine ourselves transported into the world of the scene depicted” (target article, sect. 6.1, para. 2). Although transportation is surely relevant to their analysis, the focus on depictions resonates more with the concept *aesthetic illusion* that has been explored broadly in the cognitive humanities (Wolf, Bernhart, & Mahler, 2013). In one analysis, Wolf (2009) suggested that aesthetic illusion “consists predominantly of a feeling, with variable intensity, of being imaginatively and emotionally immersed in a represented world and of experiencing this world in a way similar (but not identical) to real life” (p. 144). Scholars have discussed the capacity of different types of depictions to foment aesthetic illusion. Those discussions have often called to the type of layering that is an important feature of the analysis C&F provide. Consider these assertions about viewers' experiences of computer-generated images (CGI): “Simply put, one will *see* a photograph, despite *knowing* that it is actually CGI, which, in turn, is to say that the beholder is *simultaneously* positioned at the extreme poles of complete rational distance from, and total immersion in, the referential illusion” (Bantleon & Tragatschnig, 2013, p. 287). This type of analysis parallels claims made by C&F and holds the potential to enrich their discussion of depictions.

This evocation of aesthetic illusion (as well as the original mention of transportation) provides a context for a disagreement

with one claim that C&F make. They define *importation*, and then argue, “Importation is different from transportation. With paintings, movies, and stage plays, recipients engage in pretense that they are *covert observers* in the scenes depicted, where a covert observer is present in a scene, but invisible, mute, and unable to intervene” (target article, sect. 6.1, para. 4). However, at least in the context of movies, viewers unmute themselves and produce content that appears to count as interventions.

Consider a study by Bezdek, Foy, and Gerrig (2013; see also Gerrig & Bezdek, 2013) in which participants were asked to speak aloud while they watched brief excerpts (2–5 minutes) from feature films. One excerpt came from Alfred Hitchcock's film “Marnie.” In a scene early in the film, Marnie is trying to exit a building after robbing an office safe. To slip by a woman cleaning the office, she removes a noisy pair of shoes so that she can walk on tip toes in stocking feet. She tucks the shoes into the pockets of her trench coat. As Marnie moves along, viewers are able to see that her shoes are becoming dislodged. Marnie, however, is unaware of this eventuality. As they spoke aloud in response to this scene, participants often looked into Marnie's future. One viewer remarked:

Oh that's cool...OH NO THE SHOE...the freakin shoe. Why did she have to put it in her pocket why couldn't she just hold the shoe?

Another viewer expressed much the same content but offered advice directly to Marnie:

The shoe's going to fall out your pocket – just hold them. Told you your shoe's going to fall out your pocket. Your shoe's going to fall out your pocket...there it goes...ha!

These two examples echo the types of language C&F illustrate that people direct to social robots. For example, the latter viewer's language provides evidence for interaction in parallel to Beth's utterances quoted in their Table 1.

C&F might suggest that this viewer is pretending to give the character advice. Given the intensity of the viewer's emotions – and the final “ha!” which approximates “I told you so!” – I would disagree. I argue that, in the moment, the viewer is genuinely behaving as if the character can benefit from their advice (Gerrig & Jacovina, 2009). With brief reflection, the viewer would certainly acknowledge that Marnie's world is inaccessible. But, in the moment, the experience of an aesthetic illusion generates behavior that is real rather than pretense.

These observations are not problematic for C&F's overall perspective that people experience social robots as depictions of social agents. Rather, the claim that people habitually function as if they are interacting with depictions provides a richer context to consider how people experience social robots. To put it plainly, people's experiences with other types of depictions make them cognitively and emotionally prepared to interact with social robots.

Still, it is hard not to wonder to what extent people's attempts to interact with social robots regularly call attention to them as depictions. Viewers behave as if they can interact with movie characters because they are “imaginatively and emotionally immersed in a represented world” (Wolf, 2009, p. 144). The physical reality of social robots' presentation might regularly counteract the possibility of an aesthetic illusion. Similarly, social robots' limited behavioral repertory may largely prevent people from being transported into a goal-directed narrative. C&F's

theoretical analysis provides a rich context to contemplate these issues.


**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Bantleon, D., & Tragatschnig, U. (2013). Wilful deceptions: Aesthetic illusion at the interface of painting, photography, and digital images. In W. Wolf, W. Bernhart & A. Mahler (Eds.), *Immersion and distance: Aesthetic illusion in literature and other media* (pp. 263–292). Rodopi.
- Bezdek, M. A., Foy, J. E., & Gerrig, R. J. (2013). “Run for it!”: Viewers’ participatory responses to film narratives. *Psychology of Aesthetics, Creativity, and the Arts*, 7 (4), 409–416.
- Clark, H. H., & Van Der Wege, M. A. (2015). Imagination in narratives. In D. Tannen, H. E. Hamilton & D. Schiffrin (Eds.), *Handbook of discourse analysis* (2nd ed., pp. 406–421). John Wiley.
- Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.
- Gerrig, R. J., & Bezdek, M. A. (2013). Aesthetic illusion in film. In W. Wolf, W. Bernhart & A. Mahler (Eds.), *Immersion and distance: Aesthetic illusion in literature and other media* (pp. 89–111). Rodopi.
- Gerrig, R. J., & Jacovina, M. E. (2009). Reader participation in the experience of narrative. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 223–254). Academic Press.
- Wolf, W. (2009). “Illusion (aesthetic)”. In P. Hühn, J. C. Meister, J. Pier, & W. Schmid (Eds.), *Handbook of narratology* (pp. 144–159). Walter de Gruyter.
- Wolf, W., Bernhart, W. & Mahler, A. (Eds.) (2013). *Immersion and distance: Aesthetic illusion in literature and other media*. Rodopi.

## How deep is AI’s love? Understanding relational AI

Omri Gillath , Syed Abumusab, Ting Ai,  
Michael S. Branicky, Robert B. Davison, Maxwell Rulo,  
John Symons and Gregory Thomas

Department of Psychology, University of Kansas, Lawrence, KS 66045, USA  
ogillath@ku.edu; syedmusab@ku.edu; tingai@ku.edu; msb@ku.edu;  
rbd@ku.edu; maxrulo20@ku.edu; johnsymons@gmail.com; gthomas@ku.edu  
<https://gillab.ku.edu/>

doi:10.1017/S0140525X22001704, e33

### Abstract

We suggest that as people move to construe robots as social agents, interact with them, and treat them as capable of social ties, they might develop (close) relationships with them. We then ask what kind of relationships can people form with bots, what functions can bots fulfill, and what are the societal and moral implications of such relationships.

Clark and Fischer (C&F) argue that people regard social robots (“bots” for short) as depictions of social agents rather than as actual social agents. Conversely, we suggest that bots can play a variety of social roles including relationship partners. Rather than thinking of bots as *representing* or *imitating* social agents we encourage readers to approach bots as capable of filling substantive roles within social systems. Adopting such an approach raises various research questions concerning bots’ roles, status,

and nature, and the moral effects of bots fulfilling such social roles.

C&F further suggest that the way authority figures (e.g., developers, engineers, corporate executives) present bots modulates peoples’ construal of the bots and the kind of interactions, and eventually relations people form with bots. The authors note that social bots already *serve* people as tutors, caretakers, receptionists, companions, and other social agents. Using the term “serve” implies that bots are perceived as tools or servants and that they are potentially also programmed as such. Both categories (tools and servants), while not the same, introduce a derogating filter. People often view servants as out-group members (“lessers”), exhibit disrespect toward them, discriminate against them, and are less likely to form an authentic relationship with them (e.g., Smooth; C&F, target article). People are also not inclined to form a relationship with what they see as a tool (however, see work on transitional objects, objectophilia, and attachment to objects, e.g., Lehman, Denham, Moser, & Reeves, 1992; Melumad & Pham, 2020). The authors’ perspective overlooks the role of other powerful social actors (e.g., media depictions of human relationships with – and attachment to – bots), and the degree to which bots employ self-learning algorithms to adapt and change on their own.

Even if C&F are correct in suggesting that bots are merely interactive depictions, the interactions people have with them are inevitably embedded within social contexts and involve specific social roles. Thus, these interactions and roles are not only depictions in peoples’ heads, but rather are sorts of relationships and a part of a larger social world. Relationships with bots could fall under the umbrella of *parasocial relationships*, where a person expends emotional energy, interest, and time, while the other party is *unaware* of the other’s existence. Parasocial relationships are most common with celebrities or organizations (such as sports teams), but bots can also play such a role. While unidirectional, parasocial relationships are nevertheless perceived as relationships and were found to fulfill people’s *need to belong* (Aw & Labrecque, 2020).

Assuming the formation of social relations with bots is possible, what kind of relationships might people form with bots? It is important to distinguish the kinds of relationships, given that different relationships are associated with different functions and outcomes. Bots might be team members, friends, confidantes, or even romantic partners. In each role, they are likely to fulfill different functions, which can lead to different outcomes.

For example, if bots fill the role of friends or relationship partners, will relationships with bots help to mitigate social ills such as loneliness (Palgi et al., 2020)? Social scientists have observed a decline in the number of people who are getting married, having sex, or having children in wealthy societies. Elsewhere we have discussed the connection between technology and these trajectories (Gillath, 2019). Will bots exacerbate these tendencies by providing high-quality replacements to relations with humans? People might find it easier to customize the characteristics of their lover bot than to expend the effort necessary to build and maintain a relationship with another human person. In turn, will bots reduce loneliness, increase it, or just hide the symptoms?

Close relationships often involve aspects such as trust, commitment, intimacy, or passion. Can bots generate the same emotions and motives in people who interact with them? On the surface, it seems they already do. For example, some people use bots in their sexual activities. Observers might see this as another example of using bots as a tool. However, users develop



relationships with sex dolls and perceive them as companions (Langcaster-James & Bentley, 2018), fall in love with them (Viik, 2020), and even marry them (Burr-Miller & Aoki, 2013; Langcaster-James & Bentley, 2018). Indeed, some sexbots come with different modes such as sexy or family mode suggesting that users get more than just sex or a tool when buying a sexbot (Dunne, 2017).

These examples highlight the need to study not only the cognitive aspects of human–bots or human–AI interactions but also the affective and relational aspects, understanding issues such as bots' ability to be responsive and the development of constructs such as trust in relations with bots (Gillath et al., 2021).

Alongside the questions about relationships with bots, one should also consider the moral implications of having a relationship with a bot. For example, in the case of certain demographics, like the elderly or children, is it morally permissible to delegate our emotional and relational responsibilities toward them to bots? Or, when people form relationships with bots who look like humans, are they willingly suspending suspicion, imagining they interact with a human, or are they being deceived, and how would we protect people from the latter? For instance, should we add warnings on the bot reminding people that this is “only” a machine? And would consumers want it (might that break their preferred illusion)?


The morality of having bots as relationship partners is complicated, both at the individual level (is it okay to make people believe that a bot loves them?) and at the societal level (what are the implications for the future of society if we have bots replacing our friends and lovers?). Having bots as companions or friends might impact and even inhibit the normal development of social interpersonal skills and relationship dynamics. This is especially important for children and young adults who might not obtain these abilities and in turn lose access to associated goods of interpersonal life and society more broadly. These are issues that should be further considered in light of the current paper.


**Competing interest.** None.

## References

- Aw, E. C. X., & Labrecque, L. I. (2020). Celebrity endorsement in social media contexts: Understanding the role of parasocial interactions and the need to belong. *Journal of Consumer Marketing*, 37, 895–908.
- Burr-Miller, A., & Aoki, E. (2013). Becoming (hetero)sexual? The hetero-spectacle of idollators and their real dolls. *Sexuality and Culture*, 17, 384–400.
- Dunne, D. (2017). Meet silicon Samantha: AI sex robot has a functioning G-Spot and can switch between “FAMILY” and “sexy” mode. *Daily Mail*. Retrieved from <https://www.dailymail.co.uk/sciencetech/article-4331408/Sex-robot-Silicon-Samantha-functioning-G-spot.html> (Accessed 29 September 2022).
- Gillath, O. (2019). Does technology spell doom for close relationships? *Scientific American Blog Network*. Retrieved June 17, 2022, from <https://blogs.scientificamerican.com/observations/does-technology-spell-doom-for-close-relationships/>
- Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115, 106607. <https://doi.org/10.1016/j.chb.2020.106607>[OG]
- Langcaster-James, M., & Bentley, G. R. (2018). Beyond the sex doll: Post-human companionship and the rise of the alldoll. *Robotics*, 7, 62.
- Lehman, E. B., Denham, S. A., Moser, M. H., & Reeves, S. L. (1992). Soft object and pacifier attachments in young children: The role of security of attachment to the mother. *Journal of Child Psychology and Psychiatry*, 33, 1205–1215.
- Melumad, S., & Pham, M. T. (2020). The smartphone as a pacifying technology. *Journal of Consumer Research*, 47, 237–255.
- Palgi, Y., Shrira, A., Ring, L., Bodner, E., Avidor, S., Bergman, Y., & Hoffman, Y. (2020). The loneliness pandemic: Loneliness and other concomitants of depression, anxiety and their comorbidity during the COVID-19 outbreak. *Journal of Affective Disorders*, 275, 109–111.
- Viik, T. (2020). Falling in love with robots: A phenomenological study of experiencing technological alterities. *Paladyn, Journal of Behavioral Robotics*, 11(1), 52–65.

## Children's interactions with virtual assistants: Moving beyond depictions of social agents

Lauren N. Girouard-Hallam  and Judith

H. Danovitch 

Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY 40292, USA  
[l0giro01@louisville.edu](mailto:l0giro01@louisville.edu)  
[j.danovitch@louisville.edu](mailto:j.danovitch@louisville.edu)  
<http://louisvillekidstudies.org>

doi:10.1017/S0140525X22001649, e34

### Abstract

Clark and Fischer argue that people see social robots as depictions of social agents. However, people's interactions with virtual assistants may change their beliefs about social robots. Children and adults with exposure to virtual assistants may view social robots not as depictions of social agents, but as social agents belonging to a unique ontological category.

In their article, Clark and Fischer (C&F) state: “It is one thing to tacitly distinguish the three perspectives on a robot (a matter of cognition) and quite another to answer questions about them (a matter of meta-cognition)” (target article, sect. 4.5, para. 1). In supporting their theory that it may be difficult for some people to think through their own conceptualizations of social agents, C&F reference Kahn et al.'s (2012) study where children ages 9–15 were asked questions about a socially contingent robot called Robovie. They argue that the language used in the study may obscure Robovie's status as a depiction of a social entity and explain why children struggled to categorize Robovie. Although we agree that prompting children to think about the ontology of social robots poses challenges, we also believe that taking a developmental perspective when considering social robots may lead to a different interpretation altogether: This generation of children do not view social robots as representations of social beings, but rather, as Kahn et al. (2012) posited, they view social robots as belonging to a new ontological category. Although C&F state that “It is an open question what children understand about social robots at each age” (target article, sect. 6.4, para. 2), we propose that recent research on children's understanding of virtual assistants provides valuable insight into how children construe social robots.

Nearly half of American parents of children under age 9 indicate that they have at least one virtual assistant in their home (Rideout & Robb, 2020), meaning that these devices are far more likely to be familiar to children than even the most popular social robots. Virtual assistants are interactive and conversational and behave in socially contingent ways. Recent research suggests that

children as young as age 4 can effectively interact with virtual assistants (e.g., Lovato & Piper, 2015; Lovato, Piper, & Wartella, 2019; Oranç & Ruggeri, 2021; Xu & Warschauer, 2020) and, by age 7, children view them as reliable information sources (Girouard-Hallam & Danovitch, 2022). Moreover, children ascribe both artifact and non-artifact characteristics to these devices. Children ages 6–10 attribute mental characteristics like intelligence, social characteristics like the capacity for friendship, and some moral standing to a familiar virtual assistant (Girouard-Hallam, Streble, & Danovitch, 2021), but they also hold that virtual assistants cannot breathe and are not alive (Girouard-Hallam & Danovitch, 2022).

Thus, similar to the children in Kahn et al.'s (2012) Robovie study, children do not treat virtual assistants entirely like other humans nor like inanimate objects. Instead, children may view them as belonging to a new ontological category that occupies its own niche between person and artifact (e.g., Kahn, Gary, & Shen, 2013; Kahn & Shen, 2017; Severson & Carlson, 2010). In a study examining children's ontological beliefs about virtual assistants, Festerling and Siraj (2020) found that 6–10-year-old children had clear ontological beliefs about humans and artifacts, but children believed that virtual assistants possessed human and artifact features simultaneously. Thus, children view virtual assistants as a unique entity rather than as a mechanical depiction of a non-unique entity, such as a person. Contrary to C&F's arguments that people view social robots as non-real facsimiles of real social agents by engaging with them and then appreciating their qualities (the dual-layer argument; target article, sect. 6.4, para. 2), and that children in particular treat robots "as interactive toys – as props in make-believe social play" (target article, sect. 6.4, para. 1), children appear to believe that virtual assistants are at once animate and inanimate, rather than separating these entities into a real structure and an imaginary depiction.

Children's para-social partnerships with virtual assistants further contribute to the idea that children view virtual assistants as a new ontological category, occupying a unique space between artifact and person. Para-social relationships are emotionally tinged and one-sided, and they commonly occur between children and media characters, such as characters from popular television shows (Richards & Calvert, 2017). Parents report that their young children form para-social relationships with virtual assistants and that these relationships result from children's exposure to these socially contingent devices (Hoffman, Owen, & Calvert, 2021). Thus, it seems that the more time children spend with virtual assistants, which can respond and engage in conversation with them, the more likely they are to believe that virtual assistants are companions that care for them and that should be cared for in turn. Similarly, there is evidence that children treat virtual assistants as trusted social partners, and benefit from pedagogical exchanges with them similar to the ones they have with human partners (Xu et al., 2021). C&F use Fischer's (2016) hypothesis that some people are "players" and some are "non-players" to explain that "not everyone is willing to play along with a robot – or to do so all the time" (target article, sect. 7.2, para. 7). We propose that children who regularly interact with virtual assistants accrue a willingness to engage as "players" with these devices, which by extension changes the way that they view them and might change the way they view social robots as well.

In conclusion, as this generation of children grows up with virtual assistants and similar devices, and virtual assistants occupy an increasing part in adults' day-to-day lives, it will be necessary to re-evaluate C&F's stance. Interactions with virtual

assistants may reveal a more complex general relationship between humans and robots than C&F claim. It may be that rather than viewing social robots as depictions of social agents, children and adults who have experience with virtual assistants instead view them as semi-social agents. In other words, they may view social robots not as a composite of several parts, but rather as a unique assemblage of human and artifact characteristics. Additional empirical research that takes a developmental approach to examining the conversations and interactions people have with virtual assistants could aid in testing C&F's hypothesis that "people construe social robots not as agents *per se*, but as *depictions* of agents" (target article, sect. 1, para. 3). A developmental and ontological perspective on social robots may move the conversation beyond mere depiction to a deeper understanding of the role social robots play in our daily lives and how we view them in turn.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.




## References

- Festerling, J., & Siraj, I. (2020). Alexa, what are you? Exploring primary school children's ontological perceptions of digital voice assistants in open interactions. *Human Development*, 64(1), 26–43. <https://doi.org/10.1159/000508499>
- Fischer, K. (2016). *Designing speech for a recipient: The roles of partner modeling, alignment and feedback in so-called 'simplified registers'* (Vol. 270). John Benjamins.
- Girouard-Hallam, L. N., & Danovitch, J. H. (2022). Children's trust in and learning from voice assistants. *Developmental Psychology*, 58(4), 646. <https://doi.org/10.1037/dev0001318>
- Girouard-Hallam, L. N., Streble, H. M., & Danovitch, J. H. (2021). Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, 3(5), 1118–1131. <https://doi.org/10.1002/hbe2.321>
- Hoffman, A., Owen, D., & Calvert, S. L. (2021). Parent reports of children's parasocial relationships with conversational agents: Trusted voices in children's lives. *Human Behavior and Emerging Technologies*, 3(4), 1–12. <https://doi.org/10.1002/hbe2.271>
- Kahn, P. H., Gary, H. E., & Shen, S. (2013). Children's social relationships with current and near-future robots. *Child Development Perspectives*, 7, 32–37. <https://doi.org/10.1111/cdep.12011>
- Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... Shen, S. (2012). "Robovie, you'll have to go inside the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*, 48, 303–314. <https://doi.org/10.1037/a0027033>
- Kahn, P. H., & Shen, S. (2017). NOC NOC, who's there? A new ontological category (NOC) for social robots. In N. Budwig, E. Turiel, & P. D. Zelazo (Eds.), *New perspectives on human development* (pp. 106–122). Cambridge University Press. <https://doi.org/10.1017/CBO9781316282755.008>
- Lovato, S., & Piper, A. M. (2015). "Siri, Is This You?" Understanding Young Children's Interactions with voice Input Systems. In *Proceedings of the 14th ACM International Conference on Interaction Design and Children, IDC 2015*, Medford, MA, USA, pp. 335–338. ACM.
- Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019). "Hey Google, Do Unicorns Exist?": Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children, IDC 2019*, Boise, ID, USA, pp. 301–313. ACM. <https://doi.org/10.1145/3311927.3323150>
- Oranç, C., & Ruggeri, A. (2021). "Alexa, let me ask you something different": Children's adaptive information search with voice assistants. *Human Behavior and Emerging Technologies*, 3(4), 1–11. <https://doi.org/10.1002/hbe2.270>
- Richards, M. N., & Calvert, S. L. (2017). Media characters, parasocial relationships, and the social aspects of children's learning across media platforms. In R. Barr & D. N. Linebarger (Eds.), *Media exposure during infancy and early childhood: The effects of content and context on learning and development* (pp. 141–163). Springer International Publishing/Springer Nature. [https://doi.org/10.1007/978-3-319-45102-2\\_9](https://doi.org/10.1007/978-3-319-45102-2_9)
- Rideout, V., & Robb, M. (2020). *The common sense census: Media use by kids age zero to eight*. Common Sense Media. Retrieved from [https://www.commonsensemedia.org/sites/default/files/uploads/research/2020\\_zero\\_to\\_eight\\_census\\_final\\_web.pdf](https://www.commonsensemedia.org/sites/default/files/uploads/research/2020_zero_to_eight_census_final_web.pdf)
- Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, 23(8–9), 1099–1103. <https://doi.org/10.1016/j.neunet.2010.08.014>
- Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent

vs. a human partner. *Computers & Education*, 161, 104059. <https://doi.org/10.1016/j.compedu.2020.104059>

Xu, Y., & Warschauer, M. (2020). What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, pp. 1–13. Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376416>

## Of children and social robots

Elizabeth J. Goldman , Anna-Elisabeth Baumann   
and Diane Poulin-Dubois 

Centre for Research in Human Development, Department of Psychology,  
Concordia University, Montréal, QC H4B 1R6, Canada  
[elizabeth.godman@concordia.ca](mailto:elizabeth.godman@concordia.ca);  
[anna-elisabeth.baumann@mail.concordia.ca](mailto:anna-elisabeth.baumann@mail.concordia.ca);  
[Diane.PoulinDubois@concordia.ca](mailto:Diane.PoulinDubois@concordia.ca)  
<https://www.concordia.ca/artsci/psychology/research/cognitive-language-development-lab.html>

doi:10.1017/S0140525X22001583, e35

### Abstract

In the target article, Clark and Fischer argue that little is known about children's perceptions of social robots. By reviewing the existing literature we demonstrate that infants and young children interact with robots in the same ways they do with other social agents. Importantly, we conclude children's understanding that robots are artifacts (e.g., not alive) develops gradually during the preschool years.

The target article aims to address the puzzle of social robots: People interact with robots as if they were humans despite knowing that social robots are artifacts. This apparent cognitive dissonance is explained by the fact that people construe social robots not as social agents *per se* but as *depictions* of social agents. Such decoupling has been documented to account for adults' attribution of intentionality to inanimate agents ranging in abstractness from geometric figures to puppets to humanoid robots. Clark and Fischer (C&F) conclude that what children understand about social robots at each age remains an open question. We review the substantial body of work that addresses whether children are aware of the fictional nature of robots. We demonstrate that even infants interact with robots as if they were social agents but that the dual orientation toward robots, understanding them as artifacts, gradually develops during the preschool years.

It is well established that children display the same behaviors toward robots and people. For example, even infants follow the gaze of robots (Meltzoff, Brooks, Shon, & Rao, 2010; Mwangi, Barakova, Diaz, Mallofre, & Rauterberg, 2018; O'Connell, Poulin-Dubois, Demke, & Guay, 2009; Okumura, Kanakogi, Kanda, Ishiguro, & Itakura, 2013). Children also imitate social robots, but until age 6, less so than they do humans (Itakura, Okanda, & Moriguchi, 2008; Schleihauf et al., 2021; Sommer et al., 2020, 2021). Some studies have shown that young children can learn new information directly from social robots (Moriguchi, Kanda, Ishiguro, Shimada, & Itakura, 2011; Okumura et al., 2013) and appear to use similar mechanisms when learning from humans or robots (Stower, Calvo-Barajas, Castellano, & Kappas,

2021). Children as young as 3 years can learn new words from robots but prefer to learn from a robot that has previously demonstrated accuracy (Brink & Wellman, 2020). When social (e.g., morphology, agency, animacy) and epistemic (e.g., expertise, competency) characteristics are pitted against one another, 5-year-olds prefer to learn new words from a competent robot over an incompetent human, whereas 3-year-olds are split about whom to trust (Baumann, Goldman, Meltzer, & Poulin-Dubois, 2022).

C&F report that the more social cues robots display, the more competent they are judged to be by adults. There is also evidence to support this conclusion in children. For example, infants are more likely to follow a robot's gaze if the robot acts in a social and communicative manner (Itakura et al., 2008; Meltzoff et al., 2010; Peca, Simut, Cao, & Vanderborcht, 2016). Children also prefer to interact with and learn from a robot that displays contingent non-verbal social cues (i.e., gaze following) over a non-contingent robot (Breazeal et al., 2016). Interestingly, cues such as goal-directedness and speech may be more important than morphology in determining how children affiliate. For instance, 3-year-old children learn equally well from a humanoid (Nao) and a non-humanoid-looking robot (Cozmo) (Baumann et al., 2022).

Evidence that children can learn from and interact with robots as they do with humans is not sufficient to conclude whether children view robots as depictions of social agents. For this, how children conceptualize and categorize robots needs to be examined. Several tasks have been designed to answer this question, including interviews and a naïve biology task. In particular, interviews can assess children's perceptions of robots across many domains (Beran, Ramirez-Serrano, Kuzyk, Fior, & Nugent, 2011; Chernyak & Gary, 2016; Goldman, 2021; Jipson & Gelman, 2007; Manzi et al., 2020). The interview questions children are asked typically include: Mental (e.g., Can the robot think?), perceptual (e.g., Can the robot see?), social (e.g., Could you trust the robot with a secret?), emotional (e.g., Does the robot have feelings?), and biological (e.g., Is the robot alive?). A recent meta-analysis revealed that age is a factor in whether children attribute mental states to robots (Thellman, de Graaf, & Ziemke, 2022). Although the findings were mixed, most studies reported that people of all ages attribute mental states to robots. Some studies indicated a stronger attribution of mental states to robots in younger children, which lessened as children got older. The literature suggests that children tend to anthropomorphize social robots and that by age 5 children, like adults, recognize that robots are artifacts but still attribute mental states to them.

Another way to assess children's conceptualization of robots (e.g., whether children view robots as artifacts) is to ask them if robots are alive. For example, Kim, Yi, and Lee (2019) asked 3-, 4-, and 5-year-olds to make an animacy judgment about a humanoid robot (Vex) (e.g., Is it alive or not alive?). The older children were less likely to say the humanoid robot was alive than the younger children. As interviews can only be used with older children, other methods are required to assess whether children depict robots as social agents. There is evidence that non-verbal infants expect objects that act like animals (e.g., self-propulsion, vocalizations) to have an inside rather than be hollow (Setoh, Wu, Baillargeon, & Gelman, 2013). Future work could build upon these findings by testing infants in a related task with robots. In a similar vein, recent work has examined how children conceptualize robots with a naïve biology task (Goldman, Baumann, & Poulin-Dubois, *in press*). Using a



modified version of Gottfried and Gelman's (2005) task, children were shown images of robots, unfamiliar animals, and artifacts and asked to select whether something biological (e.g., heart) or mechanical (e.g., gears) belonged inside. A developmental shift in children's categorization of social robots was found: 5-year-olds believed that a humanoid robot (Nao) had mechanical insides, but 3-year-olds equally attributed mechanical and biological insides to the humanoid robot. This finding held when 3-year-olds were presented with a non-humanoid robot (Cozmo) (Goldman et al., *in press*).

Although more work is needed, existing research suggests that by the age of 5, children recognize that robots are not alive yet still attribute epistemic (Baumann et al., 2022; Stower et al., 2021) and social (Breazeal et al., 2016) characteristics to them. Thus, preschool children treat robots as depictions of social agents as they interact with and learn from robots while still recognizing them as inanimate objects.

**Financial support.** This work was supported by an Insight Grant from the Social Sciences and Humanity Research Council of Canada to Diane Poulin-Dubois (no. 435-2017-0564).

**Competing interest.** None.

## References

- Baumann, A., Goldman, E. J., Meltzer, A., & Poulin-Dubois, D. (2022). People do not always know best: Preschoolers' trust in social robots. *SocArXiv* [under revision]. Preprint: <https://doi.org/10.31235/osf.io/g37cq>
- Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in child-robot interaction. *International Journal of Human-Computer Studies*, 69(7-8), 539-550.
- Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. *Topics in Cognitive Science*, 8(2), 481-491. <https://doi.org/10.1111/tops.12192>
- Brink, K. A., & Wellman, H. M. (2020). Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology*, 56(7), 1268-1277. <https://doi.org/10.1037/dev0000884>
- Chernyak, N., & Gary, H. E. (2016). Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. *Early Education and Development*, 27(8), 1175-1189.
- Goldman, E. J. (2021). *Preschool-age children's understanding about a novel robotic toy: Exploring the role of parent-child conversation* [Doctoral dissertation]. UC Santa Cruz.
- Goldman, E. J., Baumann, A., & Poulin-Dubois, D. (in press). Preschoolers anthropomorphizing of robots: Do human-like properties matter? *Frontiers in Developmental Psychology*. <https://doi.org/10.3389/fpsyg.2022.1102370>
- Gottfried, G. M., & Gelman, S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development*, 20(1), 137-158. <https://doi.org/10.1016/j.cogdev.2004.07.003>
- Itakura, S., Okanda, M., & Moriguchi, Y. (2008). Discovering mind: Development of mentalizing in human children. In S. Itakura & K. Fujita (Eds.), *Origins of the social mind: Evolutionary and developmental views* (pp. 179-198). Springer Science + Business Media. [https://doi.org/10.1007/978-4-431-75179-3\\_9](https://doi.org/10.1007/978-4-431-75179-3_9)
- Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development*, 78(6), 1675-1688. <https://doi.org/10.1111/j.1467-8624.2007.01095.x>
- Kim, M., Yi, S., & Lee, D. (2019). Between living and nonliving: Young children's animacy judgments and reasoning about humanoid robots. *PLoS ONE*, 14(6), e0216869. <https://doi.org/10.1371/journal.pone.0216869>
- Manzi, F., Peretti, G., Di Dio, C., Cangelosi, A., Itakura, S., Kanda, T., ... Marchetti, A. (2020). A robot is not worth another: Exploring children's mental state attribution to different humanoid robots. *Frontiers in Psychology*, 11, 2011. <https://doi.org/10.3389/fpsyg.2020.02011>
- Meltzoff, A. N., Brooks, R., Shon, A. P., & Rao, R. P. N. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks*, 23(8-9), 966-972. <https://doi.org/10.1016/j.neunet.2010.09.005>
- Moriguchi, Y., Kanda, T., Ishiguro, H., Shimada, Y., & Itakura, S. (2011). Can young children learn words from a robot? *Interaction Studies*, 12, 107-118.
- Mwangi, E., Barakova, E. I., Diaz, M., Mallofre, A. C., & Rauterberg, M. (2018). Dyadic Gaze Patterns during Child-Robot Collaborative Gameplay in a Tutoring Interaction. 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing and Tai'an, China. <https://doi.org/10.1109/roman.2018.8525799>
- O'Connell, L., Poulin-Dubois, D., Demke, T., & Guay, A. (2009). Can infants use a non-human agent's gaze direction to establish word-object relations? *Infancy*, 14(4), 414-438. <https://doi.org/10.1080/15250000902994073>
- Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., & Itakura, S. (2013). Can infants use robot gaze for object learning? *Interaction Studies*, 14(3), 351-365. <https://doi.org/10.1075/is.14.3.03oku>
- Peca, A., Simut, R., Cao, H.-L., & Vanderborcht, B. (2016). Do infants perceive the social robot Keepon as a communicative partner? *Infant Behavior and Development*, 42, 157-167. <https://doi.org/10.1016/j.infbeh.2015.10.005>
- Schleihauf, H., Hoehl, S., Tsvetkova, N., König, A., Mombaur, K., & Pauen, S. (2021). Preschoolers' motivation to over-imitate humans and robots. *Child Development*, 92(1), 222-238. <https://doi.org/10.1111/cdev.13403>
- Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences of the United States of America*, 110(40), 15937-15942. <https://doi.org/10.1073/pnas.1314075110>
- Sommer, K., Davidson, R., Armitage, K. L., Slaughter, V., Wiles, J., & Nielsen, M. (2020). Preschool children overimitate robots, but do so less than they overimitate humans. *Journal of Experimental Child Psychology*, 191, 104702. <https://doi.org/10.1016/j.jecp.2019.104702>
- Sommer, K., Slaughter, V., Wiles, J., Owen, K., Chiba, A. A., Forster, D., ... Nielsen, M. (2021). Can a robot teach me that? Children's ability to imitate robots. *Journal of Experimental Child Psychology*, 203, 105040. <https://doi.org/10.1016/j.jecp.2020.105040>
- Stower, R., Calvo-Barajas, N., Castellano, G., & Kappas, A. (2021). A meta-analysis on children's trust in social robots. *International Journal of Social Robotics*, 13(8), 1979-2001. <https://doi.org/10.1007/s12369-020-00736-8>
- Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4), 1-51. <https://doi.org/10.1145/3526112>

## Social robots as social learning partners: Exploring children's early understanding and learning from social robots

Amanda Haber and Kathleen H. Corriveau 

Wheelock College of Education and Human Development, Boston University, Boston, MA 02215, USA

[haber317@bu.edu](mailto:haber317@bu.edu);  
[kcorriv@bu.edu](mailto:kcorriv@bu.edu)

doi:10.1017/S0140525X22001601, e36

### Abstract

Clark and Fischer propose that people interpret social robots not as social agents, but as interactive depictions. Drawing on research focusing on how children selectively learn from social others, we argue that children do not view social robots as interactive toys but instead treat them as social learning partners and critical sources of information.

Clark and Fischer (C&F) offer a new approach for how people construe social robots, arguing that people interpret social robots not as social agents, but as interactive depictions. We agree with the authors that like voice assistants, people expect to *interact* with social robots. However, in contrast to C&F, we argue that children do not construe social robots as interactive toys but instead treat them as social learning partners. Such a distinction

is important, as children's environments are increasingly filled with robots: According to the *Allied Business Intelligence Inc.* by 2024, over 79 million homes will have at least one robot. Thus, an examination of research focusing on how children selectively learn from social others is critical to exploring children's early understanding of and interaction with social robots.

Young children can learn about the world around them through their own first-hand observations, exploration, experimentation, and by actively seeking information from social learning partners including testimony from adults (e.g., caregivers, teachers, peers; Harris, Koenig, Corriveau, & Jaswal, 2018; Wang, Tong, & Danovitch, 2019) as well as nonhuman agents such as voice assistants (Siri, Alexa; Aeschlimann, Bleiker, Wechner, & Gampe, 2020; Girouard-Hallam & Danovitch, 2022; Girouard-Hallam, Streble, & Danovitch, 2021; Oranç & Ruggeri, 2021), computers (e.g., Danovitch & Alzhabi, 2013; Noles, Danovitch, & Shafto, 2015), or social robots (Breazeal et al., 2016; Oranç & Ruggeri, 2021). Indeed, prior work demonstrates that toddlers (aged 18–24 months; Movellan, Eckhardt, Virnes, & Rodriguez, 2009) and children (aged 3–6; Tanaka & Matsuzoe, 2012) are able to learn new words from social robots, suggesting that from an early age, children treat such as agents as learning partners and critical sources of information.

To date, an extensive body of literature examining children's trust of testimony from others indicates that preschoolers are surprisingly selective when deciding whom to learn from (Harris, 2012; see Harris et al., 2018 for review; Mills, 2013). Preschoolers attend to the informant's epistemic characteristics such as an individual's prior accuracy or expertise (e.g., Birch, Vauthier, & Bloom, 2008; Harris & Corriveau, 2011; Sobel & Kushnir, 2013) as well as social characteristics including familiarity, eye contact, confidence (or uncertainty), social group, and contingent interactions (e.g., Brink & Wellman, 2020; Corriveau & Harris, 2009; Corriveau, Kinzler, & Harris, 2013; Koenig, Clement, & Harris, 2004).

Importantly, unlike what is proposed by C&F, children appear to employ similar strategies when determining the credibility of social robots, as they do when they make inferences about humans (e.g., Danovitch et al., 2013; Oranç & Ruggeri, 2021). Moreover, like they do with humans, children prefer to learn from an accurate over inaccurate computer (Danovitch et al., 2013) and an accurate over an inaccurate social robot (aged 3; Brink & Wellman, 2020; Geiskkovitch et al., 2019). Similarly, they prefer to ask for information from a robot who engaged in greater contingent behavior (aged 3–5; Breazeal et al., 2016) or a more interactive teaching style (aged 4–6; Okita, Ng-Thow-Hing, & Sarvadevabhatla, 2009). Additionally, young children (aged 3–6) even attribute expertise to social robots in certain domains, with children are more likely to direct questions to robots (vs. an adult informant) when the topic was about machines, but more likely to ask humans about psychological or physics-related questions. Taken together, these data support the idea that children engage with social robots in much the same way as they do with other social informants – and importantly, not simply as interactive depictions.

Further, although children as young as 3 recognize that nonhuman agents are not alive (Jipson & Gelman, 2007), they treat them as they would other interlocutors. Indeed, children view computers and social robots as factual sources of information (Danovitch & Keil, 2008) and attribute mental capacities, moral and psychological characteristics to social robots (Breazeal et al., 2016 [children aged 3–5]; Kahn et al., 2012 [children aged 9–12]; Bernstein & Crowley, 2008 [children aged 4–7]) as well as voice assistants (e.g.,

Girouard-Hallam et al., 2021 [children aged 6–10]). Moreover, such judgments about the capacities of nonhuman agents can also impact children's learning preferences. For example, 3–6-year-olds who attributed greater perceptual abilities to robots were more likely to choose to learn from a robot rather than a human informant (Oranç & Ruggeri, 2021). These data support the notion that children view such agents as true social learning partners, and not simply interactive toys similar to dolls.

In sum, the authors argue that children construe social robots as interactive toys.

However, we argue that equating social robots to other toys children may use in pretend play does not account for the critical role that robots play in children's early learning. We urge C&F to consider this more sophisticated view of social robots and how this would impact their theoretical perspective. Such a view is increasingly important in today's society. Children today spend a great deal of time interacting with and learning from nonhuman agents including social robots and voice assistants, highlighting the need for consideration of children's use of social robots as social learning partners across the lifespan.

**Financial support.** This work received no specific grant from any funding agency.

**Competing interest.** None.

## References

- Aeschlimann, S., Bleiker, M., Wechner, M., & Gampe, A. (2020). Communicative and social consequences of interactions with voice assistants. *Computers in Human Behavior*, 112, 106466. <https://doi.org/10.1016/j.chb.2020.106466>
- Bernstein, D., & Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *Journal of the Learning Sciences*, 17(2), 225–247. <https://doi.org/10.1080/10508400801986116>
- Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, 107(3), 1018–1034. <https://doi.org/10.1016/j.cognition.2007.12.008>
- Breazeal, C., Harris, P. L., DeSteno, D., Westlund, J. M. K., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. *Topics in Cognitive Science*, 8(2), 481–491. <https://doi.org/10.1111/tops.12192>
- Brink, K. A., & Wellman, H. M. (2020). Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology*, 56(7), 1268–1277. <https://doi.org/10.1037/dev0000884>
- Corriveau, K., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, 12(3), 426–437. <https://doi.org/10.1111/j.1467-7687.2008.00792.x>
- Corriveau, K. H., Kinzler, K. D., & Harris, P. L. (2013). Accuracy trumps accent in children's endorsement of object labels. *Developmental Psychology*, 49(3), 470–479. <https://doi.org/10.1037/a0030604>
- Danovitch, J. H., & Alzhabi, R. (2013). Children show selective trust in technological informants. *Journal of Cognition and Development*, 14(3), 499–513. <https://doi.org/10.1080/15248372.2012.689391>
- Danovitch, J. H., & Keil, F. C. (2008). Young humans: The role of emotions in children's evaluation of moral reasoning abilities. *Developmental Science*, 11(1), 33–39. <https://doi.org/10.1111/j.1467-7687.2007.00657.x>
- Geiskkovitch, D. Y., Thiessen, R., Young, J. E., & Glenwright, M. R. (2019). What? That's not a chair!: How robot informational errors affect children's trust towards robots. In *ACM/IEEE International Conference on Human-Robot Interaction* (pp. 48–56).
- Girouard-Hallam, L. N., & Danovitch, J. H. (2022). Children's trust in and learning from voice assistants. *Developmental Psychology*, 58(4), 646. <https://doi.org/10.1037/dev0001318>
- Girouard-Hallam, L. N., Streble, H. M., & Danovitch, J. H. (2021). Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, 5, hbe2.321. <https://doi.org/10.1002/hbe2.321>
- Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.
- Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society Biological Sciences*, 366(1567), 1179–1187. <https://doi.org/10.1098/rstb.2010.0321>
- Harris, P. L., Koenig, M., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, 69, 251–273. <https://doi.org/10.1146/annurev-psych-122216-011710>

- Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development*, 78(6), 1675–1688. <https://doi.org/10.1111/j.1467-8624.2007.01095.x>
- Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... Severson, R. L. (2012). Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes? In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–40. <https://doi.org/10.1145/2157689.2157696>
- Koenig, M., Clement, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, 15(10), 694–698. <https://doi.org/10.1111/j.0956-7976.2004.00742.x>
- Mills, C. M. (2013). Knowing when to doubt: Developing a critical stance when learning from others. *Developmental Psychology*, 49(3), 404–418. <https://doi.org/10.1037/a0029500>
- Movellan, J., Eckhardt, M., Virnes, M., & Rodriguez, A. (2009). Sociable Robot Improves Toddler Vocabulary Skills. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction – HRI '09*, p. 307. <https://doi.org/10.1145/1514095.1514189>
- Noles, N. S., Danovitch, J. H., & Shafto, P. (2015). Children's trust in technological and human informants. *Proceedings of the Cognitive Science Society*, 6, 1721–1726.
- Okita, S. Y., Ng-Thow-Hing, V., & Sarvadevabhata, R. (2009). Learning together: ASIMO developing an interactive learning partnership with children. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 1125–1130). IEEE.
- Oranç, C., & Ruggeri, A. (2021). “Alexa, let me ask you something different” children's adaptive information search with voice assistants. *Human Behavior and Emerging Technologies*, 3(4), 595–605. <https://doi.org/10.1002/hbe2.270>
- Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, 120(4), 779–797. <https://doi.org/10.1037/a0034191>
- Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1), 78–95. <https://doi.org/10.5898/JHRI.1.1.Tanaka>
- Wang, F., Tong, Y., & Danovitch, J. (2019). Who do I believe? Children's epistemic trust in internet, teacher, and peer informants. *Cognitive Development*, 50, 248–260. <https://doi.org/10.1016/j.cogdev.2019.05.006>

## “Who’s there?”: Depicting identity in interaction

Patrick G. T. Healey<sup>a</sup> , Christine Howes<sup>b</sup>,  
Ruth Kempson<sup>c</sup>, Gregory J. Mills<sup>d,e</sup>, Matthew Purver<sup>a,f</sup>,  
Eleni Gregoromichelaki<sup>b,c</sup>, Arash Eshghi<sup>g</sup> and  
Julian Hough<sup>a</sup>

<sup>a</sup>School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK; <sup>b</sup>Department of Philosophy, Linguistics, Theory of Science, University of Gothenburg, 41255 Gothenburg, Sweden; <sup>c</sup>Department of Philosophy, King's College London, London WC2R 2LS, UK; <sup>d</sup>Faculty of Arts, Computational Linguistics (CL), University of Groningen, 9712 EK Groningen, Netherlands; <sup>e</sup>School of Computer Science and Mathematics, Kingston University, Surrey KT1 1LQ, UK; <sup>f</sup>Jozef Stefan Institute, Ljubljana, Slovenia and <sup>g</sup>School of Mathematical & Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

[p.healey@qmul.ac.uk](mailto:p.healey@qmul.ac.uk)

[christine.howes@gu.se](mailto:christine.howes@gu.se)

[ruth.kempson@kcl.ac.uk](mailto:ruth.kempson@kcl.ac.uk)

[G.Mills@kingston.ac.uk](mailto:G.Mills@kingston.ac.uk)

[eleni.gregoromichelaki@gu.se](mailto:eleni.gregoromichelaki@gu.se)

[A.Eshghi@hw.ac.uk](mailto:A.Eshghi@hw.ac.uk)

[m.purver@qmul.ac.uk](mailto:m.purver@qmul.ac.uk)

[j.hough@qmul.ac.uk](mailto:j.hough@qmul.ac.uk)

<http://cogsci.eecs.qmul.ac.uk>

<https://gu-clasp.github.io/people/ruth-kempson/>

doi:10.1017/S0140525X22001492, e37

### Abstract

Social robots have limited social competences. This leads us to view them as depictions of social agents rather than actual social agents. However, people also have limited social competences. We argue that all social interaction involves the depiction of social roles and that they originate in, and are defined by, their function in accounting for failures of social competence.

Clark and Fischer (C&F) provide a timely reminder that there is a large and underappreciated gap between the ambitions of social robotics and the actual social competence of robots (Park, Healey, & Kaniadakis, 2021). As they demonstrate, natural conversation presents complex challenges that go well beyond current engineering capabilities (see also Healey, 2021). Nonetheless, they also point to parallels in the ways in which people interact with each other and with social robots.

This commentary questions the ontological distinction underlying C&F's discussion. Specifically, does their account of depiction provide a principled basis for their argument that depictions of social agency fundamentally differ from actual social agency?

C&F discuss various examples of depictions of social agents including Laurence Olivier's performance of Hamlet. Depiction in these examples is complex. The character – Hamlet – is based on a mixture of characters from earlier plays (possibly also Shakespeare's son); there are multiple versions of the text of Hamlet; different productions select different parts of those texts, different actors perform those parts differently; direction, costume, staging, scenography vary, and so on. C&F embrace this complexity and use it to characterise various aspects of ways people treat interaction with social robots as performance.

The problem, as we see it, is that C&F's account of depiction is so rich, encompassing so much of human social interaction, that the distinction between actual social agents and depictions of social agents dissolves. As C&F show, there are familiar contexts in which people perform a role, such as hotel receptionist, which also involve derived authority, particular communicative styles and particular costumes and props. These roles are depictions and successful interaction in these cases requires that we recognise and engage with the performance (Eshghi, Howes, & Gregoromichelaki, 2022). However, arguably, all human social interaction has these properties (Kempson, Cann, Gregoromichelaki, & Chatzikyriakidis, 2016). It was Goffman's (1959) insight that this kind of performative, depictive, dramaturgical description can be applied to any human social interaction.

When the receptionist in C&F's example (target article, sect. 8.1) switches to being someone who grew up in the same region as Clark, this is, in Goffman's terms, a switch from one kind of performed identity to another. It involves, for example, switching to certain kinds of community-specific knowledge, norms, and patterns of language use (see also Clark, 1996). People have multiple overlapping identities, all involving elements of depiction: different social repertoires, forms of authority, and conventions of interpretation. Moreover, it is unclear why such performances of identity involve depictions rather than *indices* to contextual features (“contextualisation cues”) that transform the current situation to a new one where the terms of the interaction have changed.

Despite this, we share the intuition that the features of interaction that C&F highlight are important. However, the crucial role that they assign to inference and pretense seems



uncharacteristically individualistic, presenting the role of potentially sophisticated robots as passive, and ignoring efforts people make to scaffold the interaction. Our suggestion is that one way to retain a meaningful, explanatory role for depictions is to abandon the assumption of any fundamental discontinuity between authentic and performed social agency and, instead, look at how depiction functions in interaction. Specifically, the way depictions are used as a means of transforming the relation between interlocutors when social performances threaten to break down; they provide a way to account for the gap between a represented social role and the role invoked to explain the performative failure. Returning to C&F's receptionist example, the inability to provide local hotel information leads to the discovery of the receptionist's actual location which prompts the conversation to switch from "customer"–"receptionist" to "people from Rapid City."

Not all failures emerge at the level of social performance. When we encounter contemporary social robots, there are a variety of ways in which things can go wrong and a variety of stances we can take to explain the failure (cf. Dennett, 1987). We quickly discover the limitations of robot social affordances and this forces us to reason about, for example, who made this thing? (authority); what is it supposed to do? (intention/character); is there hardware failure (base scene)? This applies equally to humans and robots: We sometimes invoke problems with authority (e.g., someone is too junior or too young to answer), intention (e.g. deceit) or hardware problems (someone can't hear, or is too drunk).

There are some empirical advantages to approaching depiction in this way. It restricts the range of possible depictions to things that are actually cited to account for disruptions to interaction rather than the indefinitely many possible forms of social depiction we could imagine. It also provides an index of social competence. The relative frequency with which we invoke interactive depictions or, for example, hardware problems, provides a measure of how sophisticated a social agent is. Embarrassment accompanies the failure of social roles (Goffman, 1967); involving characteristic displays such as blushing, averting eye contact, face touching, and smiling and laughter. Unlike shame, embarrassment also directly implicates other participants in a coordinated understanding of what has failed, how it failed and how to recover from it. Interestingly, robots are not currently designed to systematically recognise or produce signals of embarrassment (Park et al., 2021).

Our assumption is that what makes an "authentic" social interaction is the ability to detect and recover from failure – something in principle achievable by machines. Machines can participate in interactions where cognitive abilities are distributed across multiple agents and each can compensate for the failures or inadequacy of the other. The centrality of miscommunication (and ability to recover from it) in human–human interaction (Healey, de Ruyter, & Mills, 2018; Howes & Eshghi, 2021) follows from the observation that we never share the same language, skills, or information as anyone we nevertheless successfully interact with (Clark, 1996). This is obvious in, for example, parent–child or expert/non-expert interactions, but is arguably characteristic of all social exchanges, including interactions with social robots. At present the potential possibilities for divergences may be broader and along different dimensions but this is not, we argue, different in kind.

**Financial support.** Christine Howes was supported by two grants from the Swedish Research council (VR) 2016-0116 – Incremental Reasoning in

Dialogue (IncReD) and 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Purver received financial support from the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103) and the project Sovrag (Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration, J5-3102); and the UK EPSRC via the project Sodestream (Streamlining Social Decision Making for Improved Internet Standards, EP/S033564/1).

**Competing interest.** None.

## References

- Clark, H. H. (1996). Communities, commonalities, and communication. In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 324–355). Cambridge University Press.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Eshghi, A., Howes, C., & Gregoromichelaki, E. (2022). Action coordination and learning in dialogue. In J.-P. Bernardy, R. Blanck, S. Chatzikyriakidis, S. Lappin & A. Maskharashvili (Eds.), *Probabilistic approaches to linguistic theory* (pp. 357–418). CSLI Publications.
- Goffman, E. (1959). *The presentation of self in everyday life*. London: Allen Lane.
- Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior* (1st ed.). Doubleday.
- Healey, P., de Ruyter, J. P., & Mills, G. J. (2018). Editors introduction: Miscommunication. *Topics in Cognitive Science*, 10(2), 264–278.
- Healey, P. G. T. (2021). Human-like communication. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence* (pp. 137–151). Oxford University Press.
- Howes, C., & Eshghi, A. (2021). Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information*, 30(2), 331–362.
- Kempson, R., Cann, R., Gregoromichelaki, E., & Chatzikyriakidis, S. (2016). Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3–4), 203–276.
- Park, S., Healey, P. G. T., & Kaniadakis, A. (2021). Should Robots Blush? In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 717, pp. 1–14. <https://doi.org/10.1145/3411764.344556>

## A neurocognitive view on the depiction of social robots

Ruud Hortensius<sup>a</sup>  and Eva Wiese<sup>b</sup>

<sup>a</sup>Department of Psychology, Utrecht University, 3584 CS Utrecht, The Netherlands and <sup>b</sup>Cognitive Psychology & Ergonomics, Institute of Psychology and Ergonomics, School of Mechanical Engineering and Transport Systems, Berlin Institute of Technology, D-10587 Berlin, Germany  
[r.hortensius@uu.nl](mailto:r.hortensius@uu.nl)  
[eva.wiese@tu-berlin.de](mailto:eva.wiese@tu-berlin.de)  
[www.ruudhortensius.nl](http://www.ruudhortensius.nl)  
<https://sites.google.com/view/gmuscilab>

doi:10.1017/S0140525X22001637, e38

### Abstract

While we applaud the careful breakdown by Clark and Fischer of the representation of social robots held by the human user, we emphasise that a neurocognitive perspective is crucial to fully capture how people perceive and construe social robots at the behavioural and brain levels.

Within their framework, Clark and Fischer (C&F) focus on observable (e.g., language) and self-reported behaviour (e.g., ratings on a questionnaire). While these measures provide a first

indication of how people perceive and interact with social robots, they do not paint a complete picture. We propose that perspectives and techniques from psychology and neuroscience will not only allow to answer if people indeed represent a social robot at three connected physical scenes but also when and how. More objective measures, such as neuroimaging, can peel apart the multiple layers of human–robot interactions by outlining the neural and behavioural mechanisms supporting these interactions.

When observing the emotions expressed by a social robot, an individual could be asked what emotion the robot is displaying (i.e., open-ended question) or if the robot is happy or angry (i.e., two-alternative forced choice). While these and similar subjective measures (e.g., questionnaires, in-depth interviews) already provide a glimpse of how the individual views the robot, they provide just that, a glimpse. The indication of an emotion does not mean that the robot is represented as a happy robot. Nor does it mean that the same mechanisms are used to observe and understand the emotions expressed by the robot as when people observe and understand the emotions of other people. Behavioural and neural measures are vital to truly understand the social cognitive mechanisms during human–robot interaction (Wiese, Metta, & Wykowska, 2017). Advanced neuroimaging and brain stimulation techniques paired with new analytic approaches as well as implicit measures will provide a more detailed understanding of the representation held by the human user. For instance, fMRI studies indicate that some neurocognitive processes, such as person perception, show similar profiles across interactions with people and social robots, while other processes, such as theory-of-mind, show dissimilar profiles during these interactions (Hortensius & Cross, 2018; Hortensius et al., 2018; Wykowska, Chaminade, & Cheng, 2016). It is important to note that even in the presence of similar behavioural patterns or neural activity, the underlying mechanism might differ between these interactions. New analytic approaches derived from cognitive neuroscience, such as representational similarity analysis, can test if behavioural reactions or activity within a neural network represents, for example, agent type (robot or human) or emotion (angry or happy) (Henschel, Hortensius, & Cross, 2020). These techniques and approaches could therefore be vital in outlining at what level and scene people represent the robot expressing emotions.

This approach has been successful in distilling the multiple layers during human–robot interactions, for example during mind perception. Top-down effects of mental state attribution are widely observed in social perception (Teufel, Fletcher, & Davis, 2010) and this also holds for human–robot interaction. Besides the appearance of the robot, the beliefs and expectations held by the individual play a critical role how they construe the social robot (Hortensius & Cross, 2018; Wykowska, Wiese, Prosser, & Müller, 2014). For example, if people believe that the action of the robot has a human-origin, activity in a region of the theory-of-mind network is increased compared to when people believe the action has a pre-programmed origin (Özdem et al., 2017). When teasing apart mind perception even further, we can dissociate two distinct processes: theory-of-mind and anthropomorphism. Often viewed or treated as similar or even as one form of mind perception, recent evidence from psychology and neuroscience suggests otherwise (Hortensius et al., 2021; Tahiroglu & Taylor, 2019). While the observed outcome, understanding the actions and hidden states of an agent, is the same, these forms of mind perception are likely supported by separate behavioural and neural mechanisms. For instance, activity within

the theory-of-mind network did not relate to an individual's tendency to ascribe human characteristics to objects and nonhuman agents, such as robots (Hortensius et al., 2021). Even if the observer understands the gestures and motion of the robot as happy, it does not mean that they truly believe that the robot is happy or that the same processes are used as when understanding the happiness of a friend.

Not only can this approach elucidate the neural and behavioural mechanism supporting human–robot interactions, it can also indicate the reliance of these interactions on both social and non-social processes. Besides the possibility that the human user represents the robot as a depiction of a social agent, it is also possible that it is represented as a depiction of an object. The main focus of human–robot interaction research has mostly been on if robots activate similar neurocognitive processes as humans. The reference or comparison category in this case is thus always a human agent, thereby restricting the focus on neurocognitive processes that are social in nature. Considering to what extent human–robot interactions rely on non-social neurocognitive processes or processes that extend over domains (e.g., attention, memory) is vital (Cross & Ramsey, 2021). Robust activation in object-specific brain regions has been observed across neuroimaging studies on the perception and interaction with robots (Henschel et al., 2020). Extending the scope of both neurocognitive mechanisms and reference and comparison categories is needed, to understand if people construct these agents as (depictions) of objects or social agents (including animals), or as a completely new, standalone category. It is unlikely that one category fits all, as for instance, not only appearance and behaviour of the robot influence social cognitive processes (Abubshait & Wiese, 2017; Waytz, Gray, Epley, & Wegner, 2010), but also context (e.g., lifelikeness of the interaction) (Abubshait, Weis, & Wiese, 2021). Importantly, people can hold different views of a robot. For example, implicit and explicit measures of mind perception do not correlate (Li, Terfurth, Woller, & Wiese, 2022). It is therefore possible that people can view a robot as an object while in parallel view the robot as a social entity ostensibly experiencing happiness.

Together, this psychology and social and cognitive neuroscience approach to the study of human–robot interaction will provide a much completer picture by providing the necessary evidence for or against the framework put forward by C&F, and ultimately tell if, when, and how people construe social robots as mere depictions of social agents.

**Financial support.** This work was supported by the Human-Centered AI focus area at Utrecht University (Embodied AI initiative).

**Competing interest.** None.

## References

- Abubshait, A., Weis, P. P., & Wiese, E. (2021). Does context matter? Effects of robot appearance and reliability on social attention differs based on lifelikeness of gaze task. *International Journal of Social Robotics*, 13(5), 863–876. <https://doi.org/10.1007/s12369-020-00675-4>
- Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology*, 8, 1393. <https://www.frontiersin.org/article/10.3389/fpsyg.2017.01393>
- Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in Cognitive Sciences*, 25(3), 200–212. <https://doi.org/10.1016/j.tics.2020.11.009>
- Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences*, 43(6), 373–384. <https://doi.org/10.1016/j.tins.2020.03.013>

- Hortensius, R., & Cross, E. S. (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), 93–110. <https://doi.org/10.1111/nyas.13727>
- Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1. <https://doi.org/10.1109/TCDS.2018.2826921>
- Hortensius, R., Kent, M., Darda, K. M., Jastrzab, L., Koldewyn, K., Ramsey, R., & Cross, E. S. (2021). Exploring the relationship between anthropomorphism and theory-of-mind in brain and behaviour. *Human Brain Mapping*, 42(13), 4224–4241. <https://doi.org/10.1002/hbm.25542>
- Li, Z., Terfurth, L., Woller, J. P., & Wiese, E. (2022). Mind the Machines: Applying Implicit Measures of Mind Perception in Social Robotics. Proceedings of the 2022 ACM/IEEE International Conference on Human–Robot Interaction, Sapporo, Hokkaido, Japan, pp. 236–245.
- Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Overwalle, F. V. (2017). Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, 12(5), 582–593. <https://doi.org/10.1080/17470919.2016.1207702>
- Tahiroglu, D., & Taylor, M. (2019). Anthropomorphism, social understanding, and imaginary companions. *British Journal of Developmental Psychology*, 37(2), 284–299. <https://doi.org/10.1111/bjdp.12272>
- Teufel, C., Fletcher, P. C., & Davis, G. (2010). Seeing other minds: Attributed mental states influence perception. *Trends in Cognitive Sciences*, 14(8), 376–382. <https://doi.org/10.1016/j.tics.2010.05.005>
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8, 1663. <https://doi.org/10.3389/fpsyg.2017.01663>
- Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693), 20150375. <https://doi.org/10.1098/rstb.2015.0375>
- Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS ONE*, 9(4), e94339. <https://doi.org/10.1371/journal.pone.0094339>

## The now and future of social robots as depictions

Bertram F. Malle<sup>a</sup>  and Xuan Zhao<sup>b</sup>

<sup>a</sup>Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912, USA and <sup>b</sup>Department of Psychology, Stanford University, Stanford, CA 94305, USA

[bfmalle@brown.edu](mailto:bfmalle@brown.edu)

[xuanzhao@stanford.edu](mailto:xuanzhao@stanford.edu)

<http://bit.ly/bfmalle>

<https://www.xuan-zhao.com/>

doi:10.1017/S0140525X22001510, e39

### Abstract

The authors at times propose that robots *are* mere depictions of social agents (a philosophical claim) and at other times that *people conceive of* social robots as depictions (an empirical psychological claim). We evaluate each claim's accuracy both now and in the future and, in doing so, we introduce two dangerous misperceptions people have, or will have, about social robots.

When interacting with robots, people face an attribution problem (Heider, 1958): To what entity should they attribute the various actions that a robot performs, such as greeting a hotel guest, tutoring a second-language speaker, or recommending a new song? A common assumption is that people conceive of the robot itself as

performing these actions. Clark and Fischer (this issue) (C&F) propose instead that people often engage in a pretense and take an imagined character to do the greeting or tutoring or recommending – a character that is merely *depicted* by the machine in front of them. The authors' innovative depiction thesis suggests similarities between social robots and other human-created depictions, such as maps, puppets, and movies, and they provide illuminating examples suggesting that at least some people, some of the time, treat robots as depictions.

To evaluate the thesis of social robots as depictions, however, we must distinguish two versions of the thesis: that social robots *are* mere depictions of social agents (a philosophical claim); and that *people conceive of* social robots as depictions (an empirical psychological claim). Moreover, we have to evaluate how the thesis fares in the present but also how it will fare in the future. Analyzing these four combinations (see Table 1), we find that evidence for the depiction thesis is limited, but the analysis reveals two dangerous misperceptions people have, or will have, about social robots: Right now, people often treat robots as autonomous agents even though in reality the robots are little more than depictions. In the future, people may fail to treat robots as the autonomous agents that they are bound to become, far more powerful than today's depictions.

Consider what robots are now. Like children's dolls and ventriloquist dummies, social robots are dressed up to perform actions that in actuality *they* do not perform: they cannot hold a conversation, be empathic, or have relationships. Like nonsocial robots (vacuum bots, manufacturing automata), social robots are programmed and controlled by designers to perform a limited number of actions; but unlike nonsocial robots, current social robots are advertised to be much more capable than they really are – that is, they are largely a pretense, a fiction.

Now consider how people treat current social robots. C&F offer vivid anecdotes but only a small number of studies that support the claim that people conceive of robots as depictions. In fact, there is considerable evidence that people often do the opposite – they treat robots as autonomous agents when they should not. People spontaneously take a robot's visual perspective (and more so if it looks highly humanlike; Zhao & Malle, 2022); people ascribe personality to robots (Ferguson, Mann, Cone, & Shen, 2019) as well as cognitive and moral capacities (Malle, 2019; Weisman, Dweck, & Markman, 2017; and more so if the robot looks highly humanlike; Zhao, Phillips, & Malle, 2019); and people feel empathy for robots, especially when the robots have animal-like appearance (Darling, 2016; Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj, & Eimler, 2013). In all these cases, people's psychological response to robots – so well-practiced in encounters with other human beings – seems to be directed to the robot-proper, not to a depicted character. Or at least there is no evidence that people compartmentalize the depiction from the depicted (as C&F suggest, p. x). Thus, people often fail to take the stance of pretense that the depiction thesis postulates; instead, they fall prey to an illusion created by designers and engineers, who exploit the deep-seated human psychology of generalization (Shepard, 1987) and lure people into a dangerous overestimation of capabilities that robots-proper currently do not have (Malle, Fischer, Young, Moon, & Collins, 2020).

Now consider what robots will be like in the future. They will not just be depictions; they will instantiate, as robots-proper, the actions that current robots only depict. Unlike dolls and dummies, they will not just be crafted and controlled by human programs. They will rapidly evolve through directing their own learning and devising



**Table 1** (Malle & Zhao). Depiction thesis, in two interpretations, now and in the future

Thesis interpretation	Now	In the future
What robots are	Most current social robots are mere depictions.	Most social robots will be autonomous agents.
How people perceive robots	Rather than treating social robots as depictions, people often ascribe more autonomy and capabilities to robots than is warranted.	People will continue to treat social robots as autonomous agents, but they may underestimate robots' true autonomy and capabilities.

their own programs. They will increasingly make autonomous decisions enabled by continuously updated and massively expanded algorithms. And equipped with complex capacities, they will perform socially significant actions – making a customer feel welcome, consoling a child, or caring for an older adult in distress.

In this future, people will ascribe such significant actions to the robot in front of them, not to any depicted character. And yet people will *underestimate* future robots' capacities, because our human psychology – evolved to co-exist with other humans – will be unprepared for robots' superhuman speed and scope of information processing and their ability to acquire vast numbers of roles and capabilities. (The reader is encouraged to watch the movie *Her* to see an example of such a being.) Designers, engineers, and scientists must help users set the right expectations of what such robots are capable of and simultaneously build robots that can communicate their capabilities to users.

But the greatest fear in fiction and philosophy has always been that robots will develop their own preferences and interests that may be in conflict with those of humans. To allay this fear, policies and regulations must be in place to ensure the design and manufacturing of robots that, while being autonomous, are still fully responsive to human influence. For this is what humans are – autonomous but also responsive to each other's influence. Robots of the future, like humans, must be able to learn the norms and values of our communities, improve from people's moral criticism, and be altered or excluded if they fail to correct themselves. Experts and community members alike must be teachers of future robots – robots as real agents, not merely as depictions.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.



## References

- Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot Law* (pp. 213–232). Edward Elgar Publishing.
- Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science*, 28(4), 331–336. <https://doi.org/10.1177/0963721419835206>
- Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.
- Malle, B. F. (2019). How many dimensions of mind perception really are there? In E. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268–2274). Cognitive Science Society.
- Malle, B. F., Fischer, K., Young, J. E., Moon, A. J., & Collins, E. C. (2020). Trust and the discrepancy between expectations and actual capabilities of social robots. In D. Zhang & B. Wei (Eds.), *Human–robot interaction: Control, analysis, and design* (pp. 1–23). Cambridge Scholars.
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1), 17–34. <https://doi.org/10.1007/s12369-012-0173-8>
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science (New York, N.Y.)*, 237(4820), 1317–1323.
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43), 11374–11379. <https://doi.org/10.1073/pnas.1704347114>

Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition*, 224, 105076. <https://doi.org/10.1016/j.cognition.2022.105076>

Zhao, X., Phillips, E., & Malle, B. F. (2019). *How people infer a humanlike mind from a robot body* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/w6r24>

## Dancing robots: Social interactions are performed, not depicted

Guido Orgs<sup>a,b</sup>  and Emily S. Cross<sup>c,d,e</sup> 

<sup>a</sup>Department of Psychology, Goldsmiths, University of London, London SE14 6NW, UK; <sup>b</sup>Department of Music, Max Planck Institute for Empirical Aesthetics, 60322 Frankfurt am Main, Germany; <sup>c</sup>School of Psychology and Neuroscience, University of Glasgow G128QB, Glasgow, UK; <sup>d</sup>MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, NSW 2751, Australia and <sup>e</sup>School of Psychological Sciences, Macquarie University, Sydney, NSW 2109, Australia

[g.orgs@gold.ac.uk](mailto:g.orgs@gold.ac.uk); <https://neurolive.info/>

[e.cross@westernsydney.edu.au](mailto:e.cross@westernsydney.edu.au); <https://www.soba-lab.com/>

doi:10.1017/S0140525X2200156X, e40

### Abstract

Clark and Fischer's depiction hypothesis is based on examples of western mimetic art. Yet social robots do not depict social interactions, but instead perform them. Similarly, dance and performance art do not rely on depiction. Kinematics and expressivity are better predictors of dance aesthetics and of effective social interactions. In this way, social robots are more like dancers than actors.

Clark and Fischer (C&F) argue that social robots are depictions of human social agents. Importantly, their argument draws heavily upon western art in the mimetic tradition, where the primary purpose (and value) of art lies in how accurately an artwork imitates reality (Shimamura, 2011). Social robots are conceptualised as interactive depictions of real humans and likened to actors in a play. C&F link the quality of a social robot to its resemblance to a human agent: The better the social robot *impersonates* a human agent, the more likely it is that people will interact with the robot in the same way.

Here, we argue that the analogy between social robots and mimetic art is flawed. This is because in many cases – including the examples provided by the authors – a social robot does not pretend to be a human agent, but instead participates in genuine social interactions, as a robot. Social robots are better likened to performance artists or dancers instead of actors; rather than depicting social interactions, they perform social interactions. This distinction between performance and depiction is important for better understanding and situating the scope and the limits of robots as social agents (Cross & Ramsey, 2021).

Much of western contemporary art neither depicts nor represents. This is especially true for performance art. For example, in Marina Abramovic's famous performance installation "The artist is present" (Abramovic & Biesenbach, 2010), she invites visitors to sit down opposite her at a table in a gallery. Abramovic neither depicts a social interaction in this artwork – she genuinely meets other people – nor does she impersonate a character. The encounter is thus performed, but it is not depicted; depicted and depictive scenes are the same. Similarly, many contemporary choreographers and theatre makers create works without a linear narrative, storyline, or obvious characters (see Fig. 1 for an example). In fact, dissolving the binary distinction between depicted and depictive scenes, or acting and not-acting (Kirby, 1987) is an important aesthetic feature of contemporary theatre, dance, and performance art (Fischer-Lichte, 2017; Lehmann, 2005). The aesthetics of dance and performance do not necessarily depend on how realistically a character is impersonated, but on a performer's expressiveness (Christensen, Lambrechts, & Tsakiris, 2019), changes in the speed and acceleration of movement sequences (Orlandi, Cross, & Orgs, 2020), or movement synchrony among a group of performers (Cracco et al., 2022; Vicary, Sperling, von Zimmermann, Richardson, & Orgs, 2017). Much of contemporary performance art or non-narrative dance therefore lacks a clear separation between depicted and depictive scenes.

C&F describe a similar example of performing without depicting: The robot "Smooth" offers a drink to Beth, who grabs the drink and thanks the robot. Beth responds to the robot naturally and intuitively, because – as in performance art – there is no distinction between depicted and depictive scenes. The robot performs a genuine social interaction: One physically embodied, social agent offers an object to another physically embodied, social agent. The robot therefore does not pose as a social agent, it *is* a genuine social agent.

In both performance art and in social interactions with robots, base scene and depictive scene are still present, yet this distinction is not specific to (or required for) engaging with performance art or social robots. People consist of bones, blood, organs, water, and so on, just as robots consist of metal and wiring. We can choose to interact with real people at different levels. For example, a surgeon spends most of her time working with the physical reality of the

body, and not the person. Moreover, in many real-life social interactions people pretend, simulate, or act (Goffman, 1990). The distinction between three levels of depiction is thus not specific to robotic agents but equally applies to human agents.

Conceptualising social robots as depictions, therefore, does not help to explain in what way robots are similar or different to human social agents. Instead, we argue that social robots are better characterised by the properties of their social interactions, for example human-like movement kinematics or turn-taking behaviour. Importantly, the physical properties of an agent – for example, the extent which it resembles the human body, are arguably less important than the way it moves or interacts with the world around it (Cross et al., 2012; Ramsey & de Hamilton, 2010). Abstract shapes can produce vivid illusions of agency, expressivity, and social relationships, as first shown in the now-famous animations of Heider and Simmel (1944), a finding that has been replicated, extended, and discussed extensively over the past half century (cf. Press, 2011).

In our own research, we have shown that movements that comply with the kinematics of human action are judged to be more natural and aesthetically pleasing than movements that violate human kinematics (Chamberlain et al., 2022). In the case of dance, greater predictability of movement kinematics increases aesthetic preference. A given sequence of dance movements is more appealing if the movements are performed with salient and rhythmic changes in speed and acceleration (Orlandi et al., 2020). Importantly, greater movement predictability also allows for smoother social interactions. For example, in cooperative tasks between two people, individuals reduce the variability of their movements to facilitate turn-taking (Vesper, van der Wel, Knoblich, & Sebanz, 2011).

In other words, we remain unconvinced that the separation between different levels of depiction is necessary or sufficient to explain why people engage socially with robots in some situations but not others. Levels of depiction do not explain why people engage with dance or performance art, because these levels do not necessarily exist for these art forms. Arguably, the interesting question is not what difference exists between real and depicted social agents, but instead: What constitutes an effective social interaction, no matter at what level of depiction it is performed?



**Figure 1** (Orgs and Cross) Performing without depicting. Seke Chimutengwende and Steph McMann in Detective Work (2021) Choreography by Seke Chimutengwende in collaboration with Steph McMann, commissioned by NEUROLIVE. Image by Hugo Glendinning.

**Financial support.** GO and ESC received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 864420 – Neurolive and No. 677270 – Social Robots). ESC also gratefully acknowledges funding from the Leverhulme Trust (PLP-2018-152).

**Competing interest.** None.

## References

- Abramovic, M., & Biesenbach, K. (2010). *Marina Abramović: The artist is present*. Museum of Modern Art.
- Chamberlain, R., Berio, D., Mayer, V., Chana, K., Leymarie, F. F., & Orgs, G. (2022). A dot that went for a walk: People prefer lines drawn with human-like kinematics. *British Journal of Psychology*, 113(1), 105–130. <https://doi.org/10.1111/bjop.12527>
- Chimutengwende, S., & McMann, S. (2021, November). *Detective work*. Siobhan Davies Studios. <https://neurolive.info/Performance-1>
- Christensen, J. F., Lambrechts, A., & Tsakiris, M. (2019). The Warburg Dance Movement Library – The WADAMO library: A validation study. *Perception*, 48(1), 26–57. <https://doi.org/10.1177/0301006618816631>
- Cracco, E., Lee, H., van Belle, G., Quenon, L., Haggard, P., Rössion, B., & Orgs, G. (2022). EEG frequency tagging reveals the integration of form and motion cues into the perception of group movement. *Cerebral Cortex*, 32(13), 2843–2857.
- Cross, E. S., Liepelt, R., de Hamilton, A. F. C., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping*, 33(9), 2238–2254. <https://doi.org/10.1002/hbm.21361>
- Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human-machine interactions. *Trends in Cognitive Sciences*, 25(3), 200–212. <https://doi.org/10.1016/j.tics.2020.11.009>
- Fischer-Lichte, E. (2017). *Ästhetik des Performativen (10. Auflage)*. Suhrkamp.
- Goffman, E. (1990). *The presentation of self in everyday life (Repr)*. Penguin.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259. <https://doi.org/10.2307/1416950>
- Kirby, M. (1987). *A formalist theatre*. University of Pennsylvania Press.
- Lehmann, H.-T. (2005). *Postdramatisches theater*. Verlag der Autoren.
- Orlandi, A., Cross, E. S., & Orgs, G. (2020). Timing is everything: Dance aesthetics depend on the complexity of movement kinematics. *Cognition*, 205, 104446. <https://doi.org/10.1016/j.cognition.2020.104446>
- Press, C. (2011). Action observation and robotic agents: Learning and anthropomorphism. *Neuroscience & Biobehavioral Reviews*, 35(6), 1410–1418. <https://doi.org/10.1016/j.neubiorev.2011.03.004>
- Ramsey, R., & de Hamilton, A. F. C. (2010). Triangles have goals too: Understanding action representation in left aIPS. *Neuropsychologia*, 48(9), 2773–2776. <https://doi.org/10.1016/j.neuropsychologia.2010.04.028>
- Shimamura, A. P. (2011). Toward a science of aesthetics: Issues and ideas. In *Aesthetic science* (pp. 3–27). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199732142.003.0010>
- Vesper, C., van der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, 211(3–4), 517–530. <https://doi.org/10.1007/s00221-011-2706-z>
- Vicary, S., Sperling, M., von Zimmermann, J., Richardson, D. C., & Orgs, G. (2017). Joint action aesthetics. *PLoS ONE*, 12(7), e0180101. <https://doi.org/10.1371/journal.pone.0180101>

## Anthropomorphism, not depiction, explains interaction with social robots

Dawson Petersen<sup>a</sup>  and Amit Almor<sup>b</sup>

<sup>a</sup>Linguistics Program, University of South Carolina, Columbia, SC 29208, USA and <sup>b</sup>Department of Psychology, Linguistics Program, Institute for Mind and Brain, Barnwell College, University of South Carolina, Columbia, SC 29208, USA  
[DHP1@email.sc.edu](mailto:DHP1@email.sc.edu)  
[almor@sc.edu](mailto:almor@sc.edu)  
[https://sc.edu/study/colleges\\_schools/artsandsciences/psychology/our\\_people/directory/almor\\_amit.php](https://sc.edu/study/colleges_schools/artsandsciences/psychology/our_people/directory/almor_amit.php)

doi:10.1017/S0140525X22001698, e41

## Abstract

We question the role given to depiction in Clark and Fischer's account of interaction with social robots. Specifically, we argue that positing a unique cognitive process for handling depiction is evolutionarily implausible and empirically redundant because the phenomena it is intended to explain are not limited to depictive contexts and are better explained by reference to more general cognitive processes.

We applaud Clark and Fischer (henceforth C&F) for calling attention to the timely question of how interaction with social robots can be nested in a broader framework. However, we question the central role given to depiction in their account. We argue that positing a specific cognitive mechanism for handling depictions is problematic *a priori* from an evolutionary perspective. Depictions are not naturally occurring phenomena which we have evolved to accommodate. Rather, they are human creations and could never have been created if we did not already possess a general mechanism to interpret them. We further wish to question the relevance of depiction as an explanatory factor regarding interactions with social robots. As we will show, many of the puzzles C&F discuss are not limited to depictive contexts and, therefore, are better explained in general terms. Specifically, we argue that (1) the type of artifact-directed social behavior performed in depictive contexts is present in other instances of anthropomorphism, (2) the levels of representation involved in depiction are present in other kinds of symbolic representation, and (3) the dissociations between social attributions and social interactions which occur with social robots are present in general social cognition. Overall, we argue that the issue of evolutionary plausibility, along with the requirements of parsimony, favors more general accounts over a depiction specific one.

The first puzzle that C&F address regards social behavior directed toward robots. We argue that this is merely one example in the broader category of artifact-directed social behavior. While C&F's explanation seems to be sufficient for robots, it fails to explain the full category. Following Airenti (2018), we note that there are clear instances of nonhuman entities eliciting social responses even when the target does not meaningfully resemble a human. For example, when a car engine fails to start, it is not uncommon for the would-be driver to engage in begging, chastising, or other social behaviors directed toward the car. It is difficult to argue that the car is a depiction of a social agent. Rather, Airenti argues that the interactive situation itself, in this case non-cooperation, is sufficient to provoke a social response. We are not convinced that there are important qualitative differences between social interactions with robots and broken cars which should require distinct explanations. In these two instances of anthropomorphic artifact-directed social behavior, it makes little difference whether the target artifact is a depiction or not. The robot's status as a depiction, while it may increase the frequency of anthropomorphization, is not necessary for it to be anthropomorphized. As such, we suggest that depiction does not play a central causal role in social interactions with robots.

The second puzzle that C&F discuss involves levels of representation. We argue that the three levels of representation that C&F propose are not unique to depictions, but rather are present in widely varying cognitive contexts. In the philosophical literature, a distinction is drawn between icons (which are analogous



to depictions, representing non-arbitrarily via correspondences between the signifier and the signified) and symbols, like words, which represent arbitrarily (Burks, 1949; de Saussure, 1983). The ability to be represented at multiple levels is by no means limited to icons. Symbols likewise can be conceived of as physical objects (sounds, marks on a page), be mentioned as representative objects bearing meaning, or be used to express their meanings without any acknowledgment of the signs themselves. The presence of these levels of representation in general symbolic reasoning calls into question the relevance of a depiction-specific framework to explain phenomena which are present in non-depictive contexts.

The third puzzle involves the relationship between social beliefs and social interactions. C&F consider interactions with social robots to be basically different from interactions with humans because (in general) we believe humans and not robots to be conscious social agents. As a result, C&F make a great deal of the fact that social robots can be treated alternatively as objects and agents while failing to recognize that the same is true of human beings. We can just as quickly attribute human behavior (falling to the ground and twitching) to a physical cause (a seizure) as we can attribute Robovie's behavior (turning off) to a physical cause (a dead battery). Equally, anyone who has worked in the service industry will relate to Smooth's experience of being treated as a mere piece of machinery by customers. These facts about human interaction undermine the assumption that social behavior relies on beliefs about agency and consciousness. While we may intuitively believe that humans are conscious and robots are not, there is little evidence that this belief greatly affects our willingness to engage socially with either. If we abandon the assumption that social beliefs determine social interactions, much of the difficulty dissolves, and there is no longer need for a bright line to distinguish depictions, non-depictive anthropomorphization, and ordinary social interaction. As with the previous two puzzles, the phenomenon that C&F seek to explain with depictions is present in non-depictive contexts, and a more general explanation is required. Given that this puzzle, like the previous two, is solvable at a general level, it is not clear to us what role a theory of depictions has to play in cognitive psychology as a whole.

In summary, we argue that the phenomena that C&F describe are not qualitatively distinguishable from other non-depictive phenomena. They are not indicative of a unique depictive cognitive process, but are simply an anthropomorphic generalization of more basic representative processes already used in social cognition. While C&F's theory is coherent and well-articulated, it is evolutionarily unmotivated, because a unique process for depictive interpretation could not arise unless depictions already existed, and it is unnecessary, because the puzzles that C&F address require (and in many cases already possess) more general explanations. The broader phenomenon of anthropomorphism, in contrast, is still vastly underexplored and lacks a fully articulated theory. We suggest that future efforts should be focused on providing and testing theories of anthropomorphism, not of social robots or depictions specifically.

**Acknowledgments.** We thank Anne Bezuidenhout and Brett Sherman for their comments on an earlier draft of this commentary.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.


**Competing interest.** None.

## References

- Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, 9, 2136.
- Burks, A. W. (1949). Icon, index, and symbol. *Philosophy and Phenomenological Research*, 9(4), 673–689.
- de Saussure, F. (1983). *Course in general linguistics*, trans. by Harris, R. Open Court Classics.

## A more ecological perspective on human–robot interactions

Varun Ravikumar, Jonathan Bowen and

Michael L. Anderson 

Rotman Institute of Philosophy, University of Western Ontario, London, Ontario N6A 3K7, Canada  
[mande54@uwo.ca](mailto:mande54@uwo.ca)  
[jbowen23@uwo.ca](mailto:jbowen23@uwo.ca)  
[vraviku@uwo.ca](mailto:vraviku@uwo.ca)  
<http://www.emrglab.org/>

doi:10.1017/S0140525X22001613, e42

### Abstract

Drawing from two strands of ecological psychology, we suggest that even if social robots are interactive depictions, people need not mentally represent them as such. Rather, people can engage with the opportunities for action or affordances that social robots offer to them. These affordances are constrained by the larger sociocultural settings within which human–robot interactions occur.

In the current state-of-the-art, social robots are not thought to be bona fide agents or interactants. Yet people sometimes interact with them *as if* they were real agents. Using the example of the interaction between the English-speaking social robot Smooth and three Danish human speakers, Clark and Fischer (C&F) point out that each speaker engages with Smooth in different ways. One speaker initially treats Smooth as if it were a person but then ignores Smooth, like the other two speakers, as if it were an inanimate object. To explain these differences in interactive behaviors, C&F propose that people *represent* social robots as *interactive depictions* of agents, which thus allows them to treat those robots as both an inanimate artifact (base scene) and as an agent (depicted scene).

We argue that in proposing a cognitivist explanation of human interactive capacities, C&F miss out on some important resources for understanding *embodied* social interactions. Drawing from two strands of ecological psychology, we suggest that even if social robots are interactive depictions, people need not mentally represent them as such. Rather, people can directly engage with the opportunities for action or *affordances* that such robots/depictions offer to them (Gibson, 1979/2015). Furthermore, the affordances that social robots offer to human interactants are constrained by the sociocultural *behavioral setting(s)* within which their interactions occur (Barker, 1968; Heft, 2001). We propose that any account of human–robot interactions ought to take into consideration the affordances of social robots and the sociocultural settings in which those interactions occur.

Consider the case of Smooth. The Seamless huMan–robot interactiOn fOr THE support of elderly people (aka SMOOTH) is an autonomous responsive robot developed for the care of the elderly in a Danish-assisted care facility. As dehydration constitutes a problem among the elderly as people may lose their sense for thirst with increasing age, Smooth was designed to serve water, among other purposes, to the facility’s residents and to encourage them to drink more (Fischer et al., 2020). Smooth is not merely a *thing* or a *tool* that residents use to satisfy their metabolic needs, rather, it was designed to afford residents the opportunity to engage in interactions not only with it but also with each other. While residents usually retreat to their rooms after mealtimes, the presence of Smooth allows for them to hang around after their meals in the common room to order a drink and to chat with each other. Smooth thus has an intentionally designed positive influence on the social lives of residents (Fischer et al., 2020).

What is it about Smooth that that invites engagement by residents? In its embodied detail, the very design of Smooth – its physical, kinematical, and functional features – elicits certain kinds of interactive behaviors from residents. Social robots like Smooth have a physical, human-like form, unlike digital tools such as Amazon’s Alexa that have a disembodied presence, which thus shapes how they are perceived. Smooth has a “penguin-like” shape and has a tray-like surface attached to its back to serve drinks; it also is of a height (and speed) designed for encounters with persons using a wheelchair or a walker (see the Target Article for an image of Smooth). It has “eyes” and “ears,” which indicates that it can see, hear, and be talked to. All these features help contribute to the regulation of Smooth’s interactions with residents: Its eye gaze, body orientation, and speed changes, for example, help signal its “intentions” to residents which then allows them to perceive and co-navigate the shared space with Smooth (Fischer et al., 2020). To reiterate, the embodied aspects of social robots *invite* interactive behaviors from people; we *directly* perceive the affordances of such robots without the need to mentally represent them (Gibson, 1979/2015).

Crucially, the affordances that Smooth offers to residents are constrained by the larger sociocultural setting within which their interactions occur. Smooth was designed to function in an elderly care facility, and its interactive capabilities need to be understood in the context of a nursing home (Fischer et al., 2020). If Smooth were to be placed in a different sociocultural setting, say at a restaurant or a public park, it may afford different kinds of behaviors to human interactants. At a restaurant, for example, Smooth may be initially welcomed by patrons as it serves them a drink but it might be later ignored as patrons engage with each other, which may explain the behaviors of the Danish speakers in C&F’s example. Here, the patrons are not dependent on Smooth for drinking water or for social interactions as in the case of an elderly care facility. In the setting of a care facility, Smooth forms an integral part of the lives of residents and affords behaviors of a different kind from that in a restaurant. Elders come to expect certain kinds of behaviors from Smooth and thus may attribute greater agency to it, and Smooth affords certain kinds of behaviors that temper the expectations of the elderly. Caregivers, on the other hand, may view Smooth as a useful *tool* to ease their overburdened workload, whereas visitors, say grandchildren, may view Smooth as a plaything (Fischer et al., 2020). The differences in how human interactants engage with social robots can thus be explained by recourse to the larger

sociocultural setting within which their interactions occur: Depending on their social roles within the assisted care facility, caregivers, residents, and visitors may engage with Smooth in different ways. The sociocultural setting structures the behaviors of interactants (Barker, 1968; Heft, 2001); there is no need to posit a complex, three-layered mental representation (of the base scene, the depiction proper, and the depicted agent) to explain how people interact with social robots/depictions.

In conclusion, the embodiment of social robots and the sociocultural settings within which interactions occur between humans and robots play an important role in shaping those interactions. In employing a cognitivist lens, C&F miss out on rich, embodied considerations when attempting to address the social artifact puzzle.


**Financial support.** This work was supported by a Canada Research Chair award to MLA (award no. 950-231929 from SSHRC).

**Competing interest.** None.

## References

- Barker, R. (1968). *Ecological psychology: Concepts and methods for studying the environment of human behavior*. Stanford University Press.
- Fischer, K., Seibt, J., Rodogno, R., Rasmussen, M. K., Weiss, A., Bodenhausen, L., ... Kruger, N. (2020). Integrative social robotics hands-on. *Interaction Studies*, 21(1), 145–191. <https://doi.org/10.1075/is.18058.fis>
- Gibson, J. J. (2015). *The ecological approach to visual perception* (Classic ed.). Psychology Press. (Original work published in 1979).
- Heft, H. (2001). *Ecological psychology in context: James Gibson, Roger Barker, and the legacy of William James’s radical empiricism*. Erlbaum.

## Virtual and real: Symbolic and natural experiences with social robots

Byron Reeves 

Department of Communication, Stanford University, Stanford, CA 94305, USA  
[reeves@stanford.edu](mailto:reeves@stanford.edu)  
[screenomics.stanford.edu](http://screenomics.stanford.edu)

doi:10.1017/S0140525X22001522, e43

### Abstract

Interactions with social robots are *symbolic experiences* guided by the pretense that robots depict real people. But they can also be *natural experiences* that are direct, automatic, and independent of any thoughtful mapping between what is real and depicted. Both experiences are important, both may apply within the same interaction, and they may vary within a person over time.

The scene is a crowded movie theater. On the big screen you see realistic dinosaurs rendered with advanced computer graphics. When they appear, your first responses are unexpected, involuntary, and quick. Your heart pounds, palms sweat, eyes open wide, body tilts backward, and your brain devises a plan to get up and run. The antidote to that discomfort is the familiar mantra – “*Calm down, it’s only a movie!*”

Repeating “it’s only a movie” works because it confirms that the raw pixels projected on the screen show things that are not actually there. The recognition that a picture is merely a depiction gives viewers room to appreciate and interpret the scene, and to consider the intentions of the filmmaker. Clark and Fischer (C&F) provide an excellent map, something missing in media psychology, of the dimensions of depiction and interpretation.

The sweating, however, is different. This is a *natural* response, the main requirement for which is the mere recognition that there are dinosaurs on the screen (Worth & Gross, 1974, 2017). This and other primitive responses are unfiltered by any thoughtful mapping between real and virtual (Lang, 2000), and they signal that the moment may require action rather than interpretation. Evolved over millennia, the responses are not needed to survive a modern movie, but they are nevertheless difficult to circumvent just because the symbols that roused them only mimic reality (Meshi, Tamir, & Heekeren, 2015; Reeves & Nass, 1996; Shepard, 1990). Natural responses seem difficultly related to the concept of depiction, and especially to the elements of interpretation, imagination, and appreciation.

### Realism

Much of the history of media technology is about inventions that promote natural responses. Bigger screens with higher resolution, virtual and augmented reality, computer graphics, three-dimensional sound, and better interactivity – all promote a sense of “being there.” And the inventions work. It matters, for example, whether you watch the dinosaurs on a smartphone, in 3D IMAX, or with VR goggles (e.g., Bailenson, 2018; Reeves, Lang, Kim, & Tatar, 1999). And the primitive responses influence thoughtful ones; they are memorable (Bolls, Lang, & Potter, 2001; Lang, Dhillon, & Dong, 1995), positively evaluated (Bartsch, Kalch, & Oliver, 2014), and the excitation often transfers to other contexts (Kramer, Guillory, & Hancock, 2014).

There are similar advances in robotic realism, including human-like skin textures (Hu & Hoffman, 2019), more purposive uses of touching (Willemse & Van Erp, 2019), better body language (Marmpena, Lim, & Dahl, 2018), and more detailed facial expressions (Chen et al., 2019). These features may give some social robots a commercial edge precisely because they make humans and machines *less* distinguishable. Even in an imagined Star Trek Holodeck future, it may still be possible to say “it’s only a robot!” but increased realism nevertheless favors natural reactions.

### Time domains

Depiction seems most relevant to longer time domains, and like for other media technologies when they were new, that domain is the current emphasis in robotics research. When people interpret, construe, imagine, or appreciate, this primarily involves “slow thinking” (Kahneman, 2011), and C&F cite numerous good examples of how people reason about social robots in this time scale.

Media realism, however, causes quick responses that occur in seconds or less. “Fast thinking” research about social robots is relatively new but increasing. For example, when people touch a robot they show heightened arousal, similar to touching humans (Li, Ju, & Reeves, 2017). And within seconds, people make judgments about the warmth and competence of social robots, just as they do for people in real life (Reeves, Hancock, & Liu, 2020).

Designers are focusing on other primitive features like eye contact (Kiilavuori, Sariola, Peltola, & Hietanen, 2021), and how robots negotiate physical space (Hoffman & Ju, 2014).

### Discretionary framing

Media experiences are not only determined by media stimuli. People can choose a frame, at least temporarily. In one relevant research paradigm, people switch between interacting with media characters that they believe are either controlled by a computer or by another real person. The mere belief that people are interacting with a real person (and not with a character that only depicts someone real) results in greater arousal and better learning (Lim & Reeves, 2010; Okita, Bailenson, & Schwartz, 2007).

It is also noteworthy that interventions designed to reduce the negative consequences of media often teach people, and particularly children, how primitive responses are triggered and how media professionals use them to control attention. This can reduce negative effects by teaching people to *choose* a symbolic experience rather than a natural one (Jeong, Cho, & Hwang, 2012).

### Sampling robots

There is likely far more variance between media characters (robots included) than variance between human responses to any one of them (Reeves, Yeykelis, & Cummings, 2016). Several years ago, we cataloged 342 social robots that were used in over 1,000 studies in the last decade (available at <https://goo.gl/Gqpzcx>). Variance between the robots is impressive and the catalog doesn’t even include some of the most interesting current products (e.g., there are no experimental studies that use sex robots as stimulus material; Döring, Mohseni, & Walter, 2020)

The point of the catalog is to show that any selection of a single or few robots can be easily biased. Consequently, stimulus sampling (Clark, 1973; Cummings & Reeves, 2022; Judd, Westfall, & Kenny, 2012; Yarkoni, 2022) is critical for social robots. A discussion based on robots that are dinosaurs, sex companions, soccer competitors, or dance partners, will be different than one based on characters from Michelangelo and Shakespeare and the most common (and least exciting?) robots in health care and children’s learning.

### Conclusion

Symbolic and natural experiences are both important for understanding how media characters are experienced. One is not more important than the other, and one cannot be explained by the other. The switching that occurs between these frames offers a different answer to the author’s *social artifact puzzle* about how people can think media characters are real, and at the same time realize that they are mere mechanical artifacts. They are both. And what matters most is how either experience works, which is applied when, and how they might interact over time.

**Financial support.** The author reports no funding related to this commentary

**Competing interest.** None.

### References

Bailenson, J. (2018). *Experience on demand: What virtual reality is, how it works, and what it can do*. WW Norton.



- Bartsch, A., Kalch, A., & Oliver, M. B. (2014). Moved to think: The role of emotional media experiences in stimulating reflective thoughts. *Journal of Media Psychology: Theories, Methods, and Applications*, 26(3), 125.
- Bolls, P. D., Lang, A., & Potter, R. F. (2001). The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research*, 28(5), 627–651.
- Chen, C., Hensel, L. B., Duan, Y., Ince, R. A., Garrod, O. G., Beskow, J., ... Schyns, P. G. (2019). Equipping Social Robots with Culturally-Sensitive Facial Expressions of Emotion using Data-Driven Methods. 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1–8). IEEE.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- Cummings, J. J., & Reeves, B. (2022). Stimulus sampling and research integrity. In L. J. Jussim, J. A. Krosnick, & S. T. Stevens (Eds.), *Research integrity: Best practices for the social and behavioral sciences* (pp. 203–223). Oxford University Press.
- Döring, N., Mohseni, M. R., & Walter, R. (2020). Design, use, and effects of sex dolls and sex robots: Scoping review. *Journal of Medical Internet Research*, 22(7), e18551.
- Hoffman, G., & Ju, W. (2014). Designing robots with movement in mind. *Journal of Human–Robot Interaction*, 3(1), 91–122.
- Hu, Y., & Hoffman, G. (2019). Using Skin Texture Change to Design Emotion Expression in Social Robots. 2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI) (pp. 2–10). IEEE.
- Jeong, S. H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of Communication*, 62(3), 454–472.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kiilavuori, H., Sariola, V., Peltola, M. J., & Hietanen, J. K. (2021). Making eye contact with a robot: Psychophysiological responses to eye contact with a human and with a humanoid robot. *Biological Psychology*, 158, 107989.
- Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790.
- Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication* 50(1), 46–70.
- Lang, A., Dhillon, K., & Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *Journal of Broadcasting & Electronic Media*, 39(3), 313–332.
- Li, J. J., Ju, W., & Reeves, B. (2017). Touching a mechanical body: Tactile contact with body parts of a humanoid robot is physiologically arousing. *Journal of Human–Robot Interaction*, 6(3), 118–130.
- Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human–Computer Studies*, 68(1–2), 57–68.
- Marmpena, M., Lim, A., & Dahl, T. S. (2018). How does the robot feel? Perception of valence and arousal in emotional body language. *Paladyn, Journal of Behavioral Robotics*, 9(1), 168–182.
- Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in Cognitive Sciences*, 19(12), 771–782.
- Okita, S. Y., Bailenson, J., & Schwartz, D. L. (2007). The mere belief of social interaction improves learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29, p. 29). Lawrence Erlbaum Associates.
- Reeves, B., Hancock, J., & Liu, S. (2020). How do people perceive social robots and what makes them effective. *Technology, Mind and Behavior*, 1(1), 1–37.
- Reeves, B., Lang, A., Kim, E. Y., & Tatar, D. (1999). The effects of screen size and message content on attention and arousal. *Media Psychology*, 1(1), 49–67.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people* (Vol. 10, p. 236605). .
- Reeves, B., Yeykelis, L., & Cummings, J. J. (2016). The use of media in media psychology. *Media Psychology*, 19(1), 49–71.
- Shepard, R. N. (1990). *Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art*. WH Freeman/ Times Books/Henry Holt.
- Willemsse, C. J., & Van Erp, J. B. (2019). Social touch in human–robot interaction: Robot-initiated touches can induce positive responses without extensive prior bonding. *International Journal of Social Robotics*, 11(2), 285–304.
- Worth, S., & Gross, L. (1974). Symbolic strategies. *Journal of Communication*, 24(4), 27–39.
- Worth, S., & Gross, L. (2017). Symbolic strategies. In *Communication theory* (pp. 121–136). Routledge.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, 1–78.

## Meta-cognition about social robots could be difficult, making self-reports about some cognitive processes less useful

Matthew Rueben 

Shiley School of Engineering, University of Portland, Portland, OR 97203, USA  
 rueben@up.edu  
<https://matthewrueben.github.io/>

doi:10.1017/S0140525X22001480, e44

### Abstract

There are reasons to suspect that meta-cognition about construing social robots as depictions would be more difficult – or absent – than Clark and Fischer discuss. Self-reports about the cognitive processes involved might therefore tend to be incomplete or inaccurate, limiting their usefulness as measures.

Clark and Fischer's (C&F's) central claim is that “people construe social robots ... as *depictions* of social agents” (target article, sect. 1, para. 3). This process involves interacting with three “scenes”: “They perceive the raw machinery of a robot, construe it as a depiction of a character, and, using the depiction as a guide, engage in the pretense that they are interacting with the character depicted” (target article, Abstract). But as C&F note in section 4.5, “It is one thing to tacitly distinguish the three perspectives on a robot (a matter of cognition) and quite another to answer questions about them (a matter of meta-cognition)” (target article, sect. 4.5, para. 1). This distinction between cognition and meta-cognition is important, partly because it determines the usefulness of self-reports as measures of cognitive processes, but C&F are vague about the extent to which people reflect on this process of construing social robots as depictions, and whether they are able to put their reflections into words. In the same section they cite the study by Kahn et al. in which participants aged 9–15 “clearly struggled” to answer questions about the nature of a Robovie robot. Assuming C&F are correct that these participants construed the Robovie robot as a depiction of a character, these participants' responses – and difficulty responding – seem to suggest that they did not understand this clearly, or were unable to put it into words. C&F do not say this outright or explore its implications, instead highlighting that the questions in the study were not clear about which of the three scenes from their framework were being asked about.

There are reasons to suspect that meta-cognition (Dunlosky & Metcalfe, 2008) about construing social robots as depictions would be more difficult – or absent – than C&F discuss. First, there could be difficulties from the nature of the measurement. A survey item or interview question might prompt the first time the participant has reflected on how they think about the robot. The amount of time and effort that participants give to this reflection could greatly affect their responses. Also, this meta-cognition is vulnerable to memory biases because participants must remember their experiences of the cognitive process. Finally, as C&F note in section 4.5, the survey item or interview

question might be ambiguous about whether it refers to the robot's physical mechanism or, to use their terminology, the character it depicts. It would be similarly problematic if participants interpreted a question as inviting them to "play along" with imagining the robot to be a character (see target article, sect. 7.2), as some participants might indeed play along in their responses while others might not, instead answering about the robot as a mere mechanical artifact.

There could also be meta-cognitive difficulties from the process itself (i.e., of construing a social robot as a depiction). For one thing, robots do not fit neatly into our existing categories. For example, C&F mention in section 2.2 the study by Gray et al. in which robots were rated low in "experience" (e.g., hunger, pain, fear)" but moderate in "agency" (e.g., self-control, morality, memory)" (target article, sect. 2.2, para. 2). These two characteristics usually occur together in animals and not at all in inanimate objects. Also, the human origins of robots' actions can be difficult to keep in mind. First of all, robots often perform actions without any direct, visible indication that a human caused that action: There is not a puppeteer with their hand inside the robot or manipulating it via strings, and robots often lack signs of remote control such as wires leading around the corner or a nearby human holding a controller (Rueben et al., 2021). Second, as C&F argue in section 7.3, people in an interaction with a robot are under time pressure to process the robot's actions as they occur so they (the person) can respond appropriately. In the language of section 6.3, this might require people to mostly do "engagement" to the exclusion of "appreciation," perhaps making it difficult to produce an account of C&F's three scenes upon reflection.

Finally, the "social artifact puzzle" is puzzling: Even if someone can articulate that they have interacted with a robot as if it were a social agent while also knowing that it is a mechanical artifact, they might not be able to reconcile those two facts in a verbal description. Even human-robot interaction (HRI) theorists who think about this puzzle professionally find it difficult, and continue to disagree about whether the correct framework is depiction or image perception (Remmers, 2020), stance taking (Thellman, 2021), a dual process theory (Złotowski et al., 2018), or something else. The reflections of laypeople on this theoretical puzzle might therefore be fragmentary, self-contradictory, or vague. Many people might simply give up.

C&F's theory might prove to explain how people "know that the robots are mechanical artifacts" and yet "interact with them as if they were actual agents" (target article, Abstract), but the process and results of people's meta-cognition about this is not much described. Additional empirical and theoretical work is needed here, especially inasmuch as meta-cognitive accounts of these cognitive processes might tend to be incomplete or inaccurate, as this commentary has suggested. One reason this is important is that HRI researchers often use self-report measures such as surveys and interviews to study anthropomorphism (Złotowski, Proudfoot, Yogeewaran, & Bartneck, 2015), mental state attribution (Thellman, de Graaf, & Ziemke, 2021), and related phenomena. Future work should study what valid inferences about cognitive processes can be made from self-reports, and when other types of measures should be used instead.

**Acknowledgments.** I am grateful to Sam Thellman for discussing the paper and commenting on my first draft, and to Peter Remmers also for discussing the paper.


**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage.
- Remmers, P. (2020). The artificial nature of social robots: A phenomenological interpretation of two conflicting tendencies in human-robot interaction. In *Culturally sustainable social robotics* (pp. 78–85). IOS Press.
- Rueben, M., Klow, J., Duer, M., Zimmerman, E., Piacentini, J., Browning, M., ... Smart, W. D. (2021). Mental models of a mobile shoe rack: Exploratory findings from a long-term in-the-wild study. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2), 1–36.
- Thellman, S. (2021). Social robots as intentional agents. Doctoral dissertation, Linköping University Electronic Press.
- Thellman, S., de Graaf, M., & Ziemke, T. (2021). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(4), 1–51.
- Złotowski, J., Proudfoot, D., Yogeewaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics*, 7(3), 347–360.
- Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, 10(5), 701–714.

## Depiction as possible phase in the dynamics of sociomorphing

Johanna Seibt 

Research Unit for Robophilosophy and Integrative Social Robotics, Aarhus University, Beder 8330, Denmark

[filseibt@cas.au.dk](mailto:filseibt@cas.au.dk)

[www.robophilosophy.org](http://www.robophilosophy.org)

doi:10.1017/S0140525X22001431, e45

### Abstract

The depiction model presents a major advance in our theoretical conceptualization of how humans experience and understand social robots. But the scope of the model is, I suggest, more limited: It pertains to one possible phase in a more comprehensive cognitive-practical dynamics of sense-making ("sociomorphing") as conceptualized in the OASIS framework. According to the OASIS framework, some basic social actions can be realized by robots, while others may be depicted in the way described by the model.

The article is an excellent illustration for the progress we can make in unraveling the "social artifact puzzle" once human-robot interaction (HRI) integrates humanities expertise pertaining to the analysis of the *symbolic space* of human social interaction. The depiction model presents a vast advance over coarse-grained taxonomies for human experiences in HRI (see e.g., Barak, Alves-Oliveira, & Ribeiro, 2020; Onnasch & Roesler, 2021), and unlike "relational" or "postphenomenological" pointers to "social construction" it provides concrete guiding concepts for analysis and design of HRI.

The authors' core assumption, however, that social robots are *always* and *only* experienced as depictions of social agents, rather than as social agents proper, seems problematic. Cognitive science and neuroscience research on "implicit" (pre-conscious) phases of social cognition provide ample counterevidence: Robot motions trigger many of the same perceptual "implicit mechanisms of social cognition" as human motions. There is thus no reason to assume that, at the level of implicit social perception, human motions are processed as socially coordinated movements but robotic motions only as depictions thereof, unless one wishes to keep the traditional assumption that sociality requires human consciousness.

The OASIS framework, another descriptive framework for human experiences in HRI developed by HRI researchers with humanities background (Seibt, 2018; Seibt, Vestergaard, & Damholdt, 2020) relinquishes this traditional assumption about sociality and allows for precise descriptions of various forms of asymmetric sociality, which has been found useful for the description of participatory sense-making in social robotics and AI (Zebrowski & McGraw, 2022). I quickly want to set some pointers for how one could combine the important insights of the depiction model with the analytical concepts of the OASIS framework, because both approaches seem to complement each other.

In the OASIS framework, human experiences of social robots are taken to involve complex cognitive-practical processes of "sociomorphing." Sociomorphing is currently a theoretical construct – a dynamics with various phases and feedback, which typically first engages preconscious "mechanisms" of social cognition and subsequently more or less routinized, tacit, or actively searching perceptual interpretation. While the latter phases may include the establishment of what the authors call the "scene depicted," the initial perceptual mechanisms effect that a robot's motion is understood as a socially coordinated bodily movement (e.g., keeping critical distance in spatial navigation). Thus, unlike in the depiction model, here it is assumed that already the "base scene" can involve *bona fide* social agents, because robots can *realize* certain basic capacities of social coordination.

OASIS recognizes 10 levels of social coordination based on capacities ranging from socio-biological automatisms to empathic coordination to various forms of collective intentionality. While robots *realize* some low-level social coordination capacities, they currently can only *simulate* more involved coordination capacities, such as the capacity to coordinate based on affective empathy or the understanding of social norms. OASIS distinguishes five degrees of simulation (defined in terms of similarity relations among [human vs. robotic] processes). If a robot simulates a high-level capacity poorly, that is, at a low degree, human responses to robots often include active interpretatory sense-making processes of the kind that the authors describe as the transition from a "base scene" to a "depiction": Kismet's (poor) simulation of *smiling-at-X* requires much interpretatory effort and thus is consciously understood as a mere depiction of *smiling-at-X*. However, this seems less plausible in the case of sophisticated simulations of coordinative capacities – *vide* the smiles of the robots Ameca or Sophia, which we may experience as insincere smiles rather than as depictions. (Unless exhibited in a museum, a three-dimensional pipe made of wood imitation is a pipe with restricted functionality.) In general, one might wonder whether the authors' thesis that all robotic gestures are experienced as depictions rides on the fact the authors' illustrations involve robots (Aibo, the Smooth robot, Asimo) with low-degree simulations of high-level coordinative capacities.

While the authors focus on the robotic *object* as artifact, prop, and character, in OASIS it is robotic actions (and parts of actions) that are the primary target of human sense-making. This allows us to differentiate between robotic actions that are low-degree simulations and thus experienced as depictions, such as Asimo's ceremonial bow, and those that we understand as such genuine social actions, without symbolic reference, such as Asimo's pointing to the right, because they are high-fidelity simulations.

Furthermore, on the OASIS approach, any social interaction requires at least seven perspectives: first-, second-, and (internal) third-person perspectives for each of the (here: two) agents, plus the external third person perspective of an observer (e.g. society at large). The cognitive activity of sociomorphing begins with implicit phases of social cognition but largely takes place in more or less tacit sense-making processes that arise when a human agent takes the second-person perspective onto her or his own action: "how will the other understand what I do?" The authors' fine-grained description of the parameters of interpretatory processes (e.g., in target article, sect. 8) offers valuable tools especially for these later phases of sociomorphing where human agents try to anticipate coordinative capacities of their interaction partner. The dimension of depiction may or may not loom large in such anticipations, depending on the degree of simulation and on whether human agents include the external third-person perspective of (in Clark and Fischer's terminology) the robot's "principal" or creator. Besides the principal, however, there are many other external third-person perspectives that might influence how we anticipate, in more or less tacit sense-making, the coordinative capacities of a robot. The taking and changing of perspectives figures centrally both in the depiction model and in OASIS, and by combining the respective perspectival differentiations we receive a more differentiated description of how people understand robotic actions.

While the OASIS account of sociomorphing could complement the idea that human understanding of robotic actions may involve that we understand them as depictions of social actions, the authors' assimilation of social robots to fictional characters strikes me as unhelpful: Social interactions cannot straddle the actual-fictional divide – a rescue robot can issue commands as rep-agent, but not as Hamlet.

**Financial support.** This work was supported by a grant from the Danish Research Council for the Project "Robot-Mediated Learning and Socratic Robotics: New Forms of Experienced Sociality for Tutoring, Self-Edification, and Coaching (ROLES)," Grant no. 9131-00136B.



**Competing interest.** None.

## References

- Barak, A. K., Alves-Oliveira, P., & Ribeiro, T. (2020). An extended framework for characterizing social robots. In C. Jost, B. Pedevic & M. Grandgeorge (Eds.), *Methods in human-robot interaction research* (pp. 21–64). Springer.
- Onnash, L., & Roesler, E. (2021). A taxonomy to structure and analyze human-robot interaction. *International Journal of Social Robotics*, 13(4), 833–849. doi: 10.1007/s12369-020-00666-5
- Seibt, J. (2018). Classifying forms and modes of co-working in the ontology of asymmetric social interactions (OASIS). In M. Coeckelbergh, J. Loh, M. Funk, J. Seibt & M. Nørskov (Eds.), *Envisioning robots in society – Proceedings of robophilosophy 2018* (pp. 134–146). IOS Press.
- Seibt, J., Vestergaard, C., & Damholdt, M. F. (2020). Sociomorphing, not anthropomorphizing: Towards a typology of experienced sociality. In M. Nørskov, J. Seibt, O. Quick (Eds.), *Culturally sustainable social robotics – Proceedings of robophilosophy 2020* (pp. 51–67). IOS Press.
- Zebrowski, R. L., & McGraw, E. B. (2022). Carving up participation: Sense-making and sociomorphing for artificial minds. *Frontiers in Neurobotics*, 16, 815850. doi: 10.3389/fnbot.2022.815850. PMID: 35774354; PMCID: PMC9239697.



## How cultural framing can bias our beliefs about robots and artificial intelligence

Jeff M. Stibel<sup>a</sup>  and H. Clark Barrett<sup>b</sup> 

<sup>a</sup>Natural History Museum, Los Angeles, CA 90007, USA and <sup>b</sup>Center for Behavior, Evolution and Culture, Department of Anthropology, University of California, Los Angeles, Los Angeles, CA 90095, USA  
[Jeff@BryantStibel.com](mailto:Jeff@BryantStibel.com)  
[Barrett@ucla.edu](mailto:Barrett@ucla.edu)

doi:10.1017/S0140525X22001686, e46

### Abstract

Clark and Fischer argue that humans treat social artifacts as depictions. In contrast, theories of distributed cognition suggest that there is no clear line separating artifacts from agents, and artifacts can possess agency. The difference is likely a result of cultural framing. As technology and artificial intelligence grow more sophisticated, the distinction between depiction and agency will blur.

Imagine a human living on earth 1,000 years ago, before the discovery of electricity or before anyone could have imagined a computer let alone an autonomous robot. Suppose this person suddenly encountered a robot transported back from the year 2022 (or, perhaps, an alien life form or probe from another planet). How would this person construe the mechanical entity? Would they treat it as a “depiction” of a real agent or simply as an agent, full stop?

Lacking any cultural, personal, or historical concept of the idea of a “robot,” it seems unlikely that a twelfth-century human would take the object before them as a human-made artifact designed to “depict” authentic agency. More likely, they would construe this unknown entity as a real agent of some kind.

Agency distinctions are not just limited to prehistoric analogies: Even well-informed individuals can perceive artifacts as having agency and intelligence. Indeed, one could imagine scenarios in which many people today, upon encountering a robot, artificial intelligence (AI), or deepfake, would not take certain artifacts as “depictions,” but as real agents. There is evidence that contemporary humans do perceive artificial agents as real. To take just one example, an AI researcher at Google has recently been suspended for arguing that a program they were interacting with had achieved sentience; this was followed by an MIT research professor who argued that Amazon’s Alexa could also become sentient (Kaplan, 2022).

Clark and Fischer (C&F) do an outstanding job outlining a particular cultural framing, or schema, of robots. Crucially, however, their theory is not and cannot be a universal theory of how all humans can, do, or will perceive and interact with artificial kinds. What is missing from C&F’s theory is an anthropological viewpoint. Through such a lens, one can see that the notion of robots as “depictions” of real agents requires expectations – a mental model of what a robot is – that are not shared by all humans.

C&F cite individuals such as Danish theatergoers who bring *a priori* assumptions about robots from films, popular media,

science, and school. Such prior expectations about robots allow people to act within a culturally delineated frame. They interact with a robot as if it had agency while knowing that the robot is a mere artifact, with no agency beyond that extended by its author.

From an anthropological perspective, it seems clear that this is a culturally provided mode of interaction, not one that has been available to all or even most humans across the span of history. Indeed, we suggest that this may not be the way that all or most humans currently perceive or will perceive robots in the future. Robots-as-depictions might be a category of robots that will always exist, but it is unlikely to be the only category of robots or artificial agents.

In evaluating C&F’s proposal, it is important to distinguish between *real* and *perceived* agency. The question of what makes something a real, or actual, agent is largely a philosophical question. The question of when people perceive, or construe, an entity as a real agent is a question for psychology and anthropology (Barrett, Todd, Miller, & Blythe, 2005; Gergely & Csibra, 2003). C&F’s article is primarily concerned with the second question and assumes that robots are not real agents. However, we argue that we should not take this for granted. It is possible for *artificial* agents to have *real* agency. As the technological sophistication of robotics and AI grows, this becomes increasingly likely.

While AI is still in its infancy, we can look to how humans interact with artifacts as a guide to how we will ultimately treat artificial agency. Consider for instance a blind person and how she interacts with her cane: Studies have demonstrated that the cane is treated as a part of the body (Malafouris, 2008; Maravita & Iriki, 2004). The effect is even more pronounced with artificial limbs (van den Heiligenberg et al., 2017, 2018), and this was likely true with stone tools as they became integral to the livelihood of prehistoric *Homo* (Haggard, Martin, Taylor-Clarke, Jeannerod, & Franck, 2003; Malafouris, 2020).

There is also evidence to support the direct impact of artifacts on our biology. As *Homo* increased its reliance on physical artifacts, our genus’ bodies grew less muscular and robust as a result (Ruff, 2005). The same may be true of cognitive tools: *Homo sapiens* have experienced more than a 5% reduction in brain mass throughout the Late Pleistocene and Holocene (Stibel, 2021) and that loss of brain mass has been linked to an increased use of cognitive tools (DeSilva et al., 2021). Modern technology, such as the internet and cell phones, have been shown to supplant thinking more broadly (Barr, Pennycook, Stolz, & Fugelsang, 2015; Sparrow, Liu, & Wegner, 2011). Cognitive tools enable thought to move to and from the brain by offloading cognition from biological wetware to artificial hardware. Just as physical artifacts offload physical exertion, cognitive offloading may allow our expensive brain tissue to be selected against while enabling our intelligence to increase (DeSilva et al., 2021; Stibel, 2021).

When an artifact is used in the thinking process, it is as much a part of the process as are the neurons in the brain (Clark & Chalmers, 1998; Malafouris, 2020; Maravita & Iriki, 2004). In that respect, artifacts are already a part of human agency so it seems reasonable to believe that, as AI gains in sophistication, we will perceive artificially intelligent agents as real and not depictions. Part of the problem may be that the term “artificial intelligence” is loaded. The technology humans create is artificial, but the intelligence created is real: *artificial* minds can have *real* intelligence. At present, most social robots are not yet sophisticated

enough to arouse any response beyond a depiction, an imitation of something that has agency. But as artificial agents gain in sophistication and intelligence, it is likely that humans will treat them as having real agency.

**Competing interest.** None.

## References

- Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). The brain in your pocket: Evidence that smartphones are used to supplant thinking. *Computers in Human Behavior*, 48, 473–480. <https://doi.org/10.1016/j.chb.2015.02.029>
- Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, 26(4), 313–331. <https://doi.org/10.1016/j.evolhumbehav.2004.08.015>
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. <http://www.jstor.org/stable/3328150>
- DeSilva, J. M., Traniello, J. F., Claxton, A. G., & Fannin, L. D. (2021). When & why did human brains decrease in size? A new change-point analysis & insights from brain evolution in ants. *Frontiers in Ecology and Evolution*, 9, 742639. <https://doi.org/10.3389/fevo.2021.742639>
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287–292. [https://doi.org/10.1016/S1364-6613\(03\)00128-1](https://doi.org/10.1016/S1364-6613(03)00128-1)
- Haggard, P., Martin, F., Taylor-Clarke, M., Jeannerod, M., & Franck, N. (2003). Awareness of action in schizophrenia. *Neuroreport*, 14(7), 1081–1085. <https://doi.org/10.1097/01.wnr.0000073684.00308.c0>
- Kaplan, M. (2022). After Google chatbot becomes “sentient,” MIT professor says Alexa could too. *New York Post*, Retrieved from <https://nypost.com/2022/06/13/mit-prof-says-alexa-could-become-sentient-like-google-chatbot/>
- Malafouris, L. (2008). Beads for a plastic mind: The “blind man’s stick” (BMS) hypothesis & the active nature of material culture. *Cambridge Archaeological Journal*, 18(3), 401–414. <https://doi.org/10.1017/S0959774308000449>
- Malafouris, L. (2020). Thinking as “thinging”: Psychology with things. *Current Directions in Psychological Science*, 29(1), 3–8. <https://doi.org/10.1177/0963721419873349>
- Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8(2), 79–86. <https://doi.org/10.1016/j.tics.2003.12.008>
- Ruff, C. B. (2005). Mechanical determinants of bone form: Insights from skeletal remains. *Journal of Musculoskeletal & Neuronal Interactions*, 5(3), 202–212.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science (New York, N.Y.)*, 333(6043), 776–778. <https://doi.org/10.1126/science.1207745>
- Stibel, J. M. (2021). Decreases in brain size & encephalization in anatomically modern humans. *Brain, Behavior and Evolution*, 96(2), 64–77. <https://doi.org/10.1159/000519504>
- van den Heiligenberg, F. M., Orlov, T., Macdonald, S. N., Duff, E. P., Henderson Slater, D., Beckmann, C. F., ... Makin, T. R. (2018). Artificial limb representation in amputees. *Brain*, 141(5), 1422–1433. <https://doi.org/10.1093/brain/awy054>
- van den Heiligenberg, F. M., Yeung, N., Brugger, P., Culham, J. C., & Makin, T. R. (2017). Adaptable categorization of hands and tools in prosthesis users. *Psychological Science*, 28(3), 395–398. <https://doi.org/10.1177/0956797616685869>

## Social robots and the intentional stance

Walter Veit<sup>a</sup>  and Heather Browning<sup>b</sup> 

<sup>a</sup>School of History and Philosophy of Science, The University of Sydney, Sydney, NSW 2006, Australia and <sup>b</sup>London School of Economics and Political Science, Centre for Philosophy of Natural and Social Science, Houghton Street, London WC2A 2AE, UK

[wvweit@gmail.com](mailto:wvweit@gmail.com); <https://walterveit.com/>

[DrHeatherBrowning@gmail.com](mailto:DrHeatherBrowning@gmail.com); <https://www.heatherbrowning.net/>

doi:10.1017/S0140525X22001595, e47

## Abstract

Why is it that people simultaneously treat social robots as mere designed artefacts, yet show willingness to interact with them as if they were real agents? Here, we argue that Dennett’s distinction between the intentional stance and the design stance can help us to resolve this puzzle, allowing us to further our understanding of social robots as interactive depictions.

Clark and Fischer (C&F) offer an excellent analysis of what they call the *social artefact puzzle*, that is, why it is that people simultaneously (1) hold the view that social robots – whether in the shape of animals or humans – are merely designed mechanical artefacts, and (2) show willingness to interact with them as if they were real agents. Their solution to this apparent inconsistency is to suggest that people do not inherently treat social robots as real agents, but rather treat them as interactive depictions (i.e., analogues) to real agents. To our surprise, however, in their discussion the authors did not mention Daniel Dennett’s (1987, 1988) distinction between the intentional stance and the design stance – two attitudes that humans routinely take in their engagement with the world. Yet we think that it is precisely this distinction that can help to address some of the unresolved issues the authors raise as currently lacking from the alternative perspectives: Why (i) people differ in their *willingness* to interact with social robots, (ii) why people can *rapidly change their perspective* of social robots, from agents to artefacts, and (iii) why people seem to only *selectively* treat social robots as agents.

The intentional stance, according to Dennett, involves treating “the system whose behavior is to be predicted as a rational agent; one attributes to the system the beliefs and desires it ought to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs” (Dennett, 1988, p. 496). This stance can be applied to other agents as well as to oneself (Veit, 2022; Veit et al., 2019). On the other hand, when one takes the *design stance* “one predicts the behavior of a system by assuming that it has a certain design (is composed of elements with functions) and that it will behave as it is designed to behave under various circumstances” (Dennett, 1988, p. 496).

When humans are faced with a social robot, both stances are useful for predicting how the robot is going to behave, so people are faced with a choice of how to treat it. Which stance they choose to adopt may depend on a range of factors, including individual differences, and the particular goals of the interaction. For instance, people will differ in their social personality traits, and their prior experience with social robots or similar artificial agents, which makes it unsurprising that they will then also differ in their willingness to adopt the intentional stance and interact with them as if they were real agents with beliefs and desires; as opposed to adopting the design stance and treating them in a more pragmatic manner, as useful objects but nothing more (though we note that Marchesi et al. [2019] did not find any differences within the demographic groups they screened for).

Thinking about these perspectives as conditional and changing stances, rather than strong ontological and normative commitments about the status of social robots and how they should be treated, removes the mystery regarding why and how people can rapidly change their perspectives of social robots, treating them as artefacts at one point in time and as agents at another. It can now be regarded as a fairly simple switch from one stance

to another. This also provides a solution to the question of why people show selectivity in their interpretation of the capacities and abilities of social robots. People can adopt one stance or the other, depending on the context and goals of the particular interaction.

It is important to keep in mind that both stances are ultimately meant to be useful within different contexts. Our interactions with social robots will occur within a range of contexts, and people will have vastly different goals depending both on their own aims and values, and the situation they are encountered in. In some cases it will be useful for someone, with reference to their goals, to ignore the nonhuman-like features of a social robot and treat them as another social agent. Particularly, in light of the evidence the authors discuss, of people's strong emotional responses to some social robots (e.g., companion "animals"), there may here be psychological and social benefits in adopting the intentional stance and treating the robot as a social agent (indeed, this would appear to be the very purpose of these robots in the first place). It may also assist in rapid and flexible predictions of behaviour, supported by the fact that people more readily adopt the intentional stance when viewing social robots interacting with other humans, than when viewing them acting alone (Spatola, Marchesi, & Wykowska, 2021). In other cases, often even within the same interaction, it will be more useful to ignore the human-like features and focus on the more mechanical properties, shifting to a treatment of the robot as an artefact instead. This is more likely in cases where interaction with the robot is more instrumental, in service of some other goal.

We want to emphasise that one doesn't have to see Dennett's account as a competitor to C&F's. Indeed, we think they are complementary. Our suggestion here is that the authors could include this distinction within their proposal, drawing more links between their account and some of the existing studies that explore the intentional and design stances in relation to people's responses to robots (e.g., Marchesi et al., 2019; Perez-Osorio & Wykowska, 2019; Spatola et al., 2021). In particular, we see benefit in more empirical research on people's interactions with and attitudes towards social robots, to test these ideas and see which may apply more strongly within different contexts. As the current evidence base is small, and underdetermines the current available theories, if we want to advance our understanding of when, how, and why ordinary people treat social robots as agents, we will ultimately need further empirical work and we think that Dennett's distinction provides an additional useful framework from which to build this.

**Financial support.** WV's research was supported under Australian Research Council's Discovery Projects funding scheme (project number FL170100160).

**Competing interest.** None.

## References

- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences*, 11(3), 495–505.
- Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots?. *Frontiers in Psychology*, 10, 450.
- Perez-Osorio, J., & Wykowska, A. (2019). Adopting the intentional stance towards humanoid robots. In *Wording robotics* (pp. 119–136). Springer.

Spatola, N., Marchesi, S., & Wykowska, A. (2021). The intentional stance test-2: How to measure the tendency to adopt intentional stance towards robots. *Frontiers in Robotics and AI*, 8, 666586.

Veit, W. (2022). Revisiting the intentionality all-stars. *Review of Analytic Philosophy*, 2(1), 1–24. <https://doi.org/10.18494/SAM.RAP.2022.0009>

Veit, W., Dewhurst, J., Dolega, K., Jones, M., Stanley, S., Frankish, K., & Dennett, D. C. (2019). The rationale of rationalization. *Behavioral and Brain Sciences*, 43, e53. <https://doi.org/10.1017/S0140525X19002164>

## Binding paradox in artificial social realities

Kai Vogeley<sup>a,b</sup> 

<sup>a</sup>Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, 59037 Cologne, Germany and <sup>b</sup>Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), 52428 Jülich, Germany  
[kai.vogele@uk-koeln.de](mailto:kai.vogele@uk-koeln.de)  
[k.vogele@fz-juelich.de](mailto:k.vogele@fz-juelich.de)  
<https://psychiatrie-psychotherapie.uk-koeln.de/forschung/ag-soziale-kognition/>  
<https://www.fz-juelich.de/de/inm/inm-3/forschung/soziale-kognition>

doi:10.1017/S0140525X22001467, e48

### Abstract

The relation between communication partners is crucial for the success of their interaction. This is also true for artificial social agents. However, the more we engage in artificial relationships, the more we are forced to regulate and control them. I refer to this as binding paradox. This deserves attention during technological developments and requires professional supervision during ongoing interactions.

Complementary to the technological development of artificial social agents, the question of how we can understand and conceptualize them in order to successfully communicate must be answered at the same time. This is the well-chosen focus of the target article by Clark and Fischer (C&F). They provide many examples for the different realizations of such agents (target article, sect. 3.2). That the relationship between two communication partners is crucial has been emphasized since the beginnings of modern social psychology (Watzlawick, Beavin, & Jackson, 1967).

In communication, we exchange information by conveying meaningful messages. According to symbolic interactionism, we interact on the basis of interpretable meanings that develop during the interaction between persons and can change over time (Blumer, 1969; Carey, 2009; Mead, 1963). However, content can only be transmitted if the communication partner is experienced as reliable and trustworthy. The "connectedness" or "attunement" between both partners is also referred to as rapport based on mutual attentiveness, reciprocal exchange of positivity cues, and coordination of nonverbal behaviors (Bernieri et al., 1996; Tickle-Degnen & Rosenthal, 1990). The relationship is the primary aspect of communication, while the content is secondary. For this reason, we tend to constantly interpret even unintended signals as meaningful: "we can not not communicate" (Watzlawick et al., 1967). These processes of communication do not always and necessarily occur unconsciously and



automatically, and their full understanding requires thoughtful consideration (C&F, target article, sect. 7).

To have a similar experience with artificial social agents, we are forced to treat them as if he or she was human or “as if they were actual agents” (C&F, target article, long abstract). We can then “respond socially and naturally” and refer to the “media equation” (C&F, target article, sect. 2.1). It is one of the earliest insights in the study of fiction that we temporarily accept fiction as reality. This “willing suspension of disbelief” was already proposed by Samuel Taylor Coleridge (1722–1834), the English critic and poet (Coleridge, 1817/1907). This early concept already contains the key components of “willingness” and “changes of perspective” that allow us to treat an artificial social actor as human at one time and as an artifact at another (C&F, target article, sect. 2.4, para. 2). This temporary suspension of disbelief depends on different dimensions (C&F, target article, sect. 3.2). It can be suggested that the more we are confronted with artificial social agents who appear and behave as “persons,” the more pronounced the suspension is (Kasap & Magnenat-Thalmann, 2007; Swartout et al., 2006; Vogeley & Bente, 2010). Even the instruction to interact with another person and plausible gaze behavior of a virtual character lead persons to believe that they are interacting with real humans (Pfeiffer et al., 2014; Vogel et al., 2021).

These socially enriched realities create experiences of “presence” or “social presence” (Bente et al., 2008), the other can become a “social hallucination” (Madary & Metzinger, 2016). This implies that this powerful technology is capable of blurring the boundaries between reality and virtuality, much like classical thought experiments of “brains in a vat” (Putnam, 1981), the “experience machine” (Nozick, 1974), or the invention of “phantomology” and “phantomatics” (Lem, 1964/2014). In a completely transformed virtual life world, we would no longer be able to distinguish between simulation and reality (Lem, 1964/2014).

It is the tension between real and artificial social agents that creates the “social artifact puzzle” that frames the target article: We communicate and interact with putative social agents even though we know they are artifacts (C&F, sects. 1 and 10). This raises ethical concerns (Marloth et al., 2020). Blurred boundaries bear the potential to be stressful (Pan & Hamilton, 2018) or become even traumatic (Ramirez & LaBarge, 2018). Legally, too, the foreseeable infliction of harm or even trauma can raise challenging questions regarding responsibility (Lemley & Volokh, 2018), which are addressed by conceptualizing “authorities” and asking for “principals” behind the agents (C&F, target article, sect. 7.3). The more realistic social artificial agents become and the more seducing it is to interact with them, the more we need to be reminded of their artificial nature and the more we need to control and regulate the depth of such a relation.

Probably the most reflective area dealing with a very similar conflict is the practice of psychotherapy. Effective psychotherapy requires the psychotherapist and the patient enter into a relationship, but the psychotherapist must maintain a professional distance and cannot simultaneously become a close friend or even a lover of the patient. Even Sigmund Freud commented on a case of a patient falling in love with the therapist as “counter-transference love” (“Übertragungsliebe”; Freud, 1914/1982). When it occurs, it requires a very careful interaction in which the relationship established must be controlled to avoid going “too deep.”

In conclusion, the relationship between humans and artificial social agents requires careful thought and reflection about the

nature of their relationship as outlined in many important aspects of C&F’s target article. Some level of rapport must be established in order to effectively interact with an artificial human, but the human partner must be protected from confusion about the quality and depth of the initiated relationships while being forced to control the relationship. This is what I call the “binding paradox.” It is related to the “social artifact puzzle” (C&F, target article, sect. 1), but extends it by conceptualizing this tension in the relationship between communication partners as more universal including also human–human relations, and opening an ethical debate. There is only a small corridor within which we can establish a functionally relevant relationship without being affected by an illusionary relationship that can become potentially harmful. This must be considered in any kind of empirical research or technological development of artificial social realities. During ongoing communication, it requires careful monitoring of people communicating with artificial agents, much like psychotherapy, which requires supervision.

**Financial support.** This work was supported by the European Commission (FET Proactive project consortium “VIRTUALTIMES,” project ID 824128) and the German Research Foundation (Collaborative Research Centre CRC 1252 “Prominence in Language,” project ID 281511265) and the German Ministry of Research and Education (SIMSUB: Simulating (inter)subjectivity, project ID 01GP2215).

**Competing interest.** None.

## References

- Bente, G., Rüggenberg, S., Krämer, N. C., & Eschenburg, F. (2008). Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human Communication Research*, 34(2), 287–318.
- Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1), 110–129.
- Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Prentice-Hall.
- Carey, J. W. (2009). *Communication as culture* (revised edn.). Routledge.
- Coleridge, S. T. (1907). *Biographia literaria* [original 1817]. Clarendon Press.
- Freud, S. (1982). *Bemerkungen zur Übertragungsliebe [original 1914] Studienausgabe Bd. I* (pp. 217–230). Fischer.
- Kasap, Z., & Magnenat-Thalmann, N. (2007). Intelligent virtual humans with autonomy and personality: State-of-the-art. *Intelligent Decision Technologies*, 1, 3–15.
- Lem, S. (2014). *Summa technologiae* [original 1964]. University of Minnesota Press (section 6).
- Lemley, M. A., & Volokh, E. (2018). Law, virtual reality, and augmented reality. *University of Pennsylvania Law Review*, 166, 1051–1138.
- Madary, M., & Metzinger, T. K. (2016). Real virtuality: A code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. *Frontiers in Robotics and AI*, 3, 1–23.
- Marloth, U., Chandler, J., & Vogeley, K. (2020). Psychiatric interventions in virtual reality: Why we need an ethical framework. *Cambridge Quarterly of Healthcare Ethics*, 29(4), 574–584.
- Mead, G. H. (1963). *Mind, self, and society* [original 1934]. University of Chicago Press.
- Nozick, R. (1974). *Anarchy, state and utopia*. Basic Books.
- Pan, X., & Hamilton, A. F. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109, 395–417.
- Pfeiffer, U., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A., Bente, G., & Vogeley, K. (2014). Why we interact: On the functional role of the Striatum in the subjective experience of social interaction. *NeuroImage*, 101C, 124–137.
- Putnam, H. (1981). *Reason, truth and history*. Cambridge University Press.
- Ramirez, E. J., & LaBarge, S. (2018). Real moral problems in the use of virtual reality. *Ethics Information Technology*, 20, 249–263.
- Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J., & Traum, D. (2006). Toward virtual humans. *AI Magazine*, 27, 96–108.
- Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4), 285–293.
- Vogel, D. H. V., Jording, M., Esser, C., Weiss, P. H., & Vogeley, K. (2021). Temporal binding is enhanced in social contexts. *Psychonomic Bulletin & Review*, 28, 1545–1555.

Vogeley, K., & Bente, G. (2010) "Artificial humans": Psychology and neuroscience perspectives on embodiment and nonverbal communication. *Neural Networks*, 23, 1077–1090.  
 Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. Norton.

## On the potentials of interaction breakdowns for HRI

Britta Wrede<sup>a</sup>, Anna-Lisa Vollmer<sup>b</sup> and Sören Krach<sup>c</sup> 

<sup>a</sup>Software Engineering for Cognitive Robots and Cognitive Systems, University of Bremen, 28359 Bremen, Germany; <sup>b</sup>Medical Assistive Systems, Bielefeld University, 33615 Bielefeld, Germany and <sup>c</sup>Department of Psychiatry and Psychotherapy, Social Neuroscience Lab (SNL), Lübeck University, Center of Brain, Behavior and Metabolism (CBBM), 23538 Lübeck, Germany  
[bwrede@techfak.uni-bielefeld.de](mailto:bwrede@techfak.uni-bielefeld.de)  
[anna-lisa.vollmer@uni-bielefeld.de](mailto:anna-lisa.vollmer@uni-bielefeld.de)  
[soeren.krach@uni-luebeck.de](mailto:soeren.krach@uni-luebeck.de)

doi:10.1017/S0140525X22001674, e49

### Abstract

How do we switch between “playing along” and treating robots as technical agents? We propose interaction breakdowns to help solve this “social artifact puzzle”: Breaks cause changes from fluid interaction to explicit reasoning and interaction with the raw artifact. These changes are closely linked to understanding the technical architecture and could be used to design better human–robot interaction (HRI).

Clark and Fischer (C&F) propose a new account for the “social artifact puzzle” as they call it: The observation that humans tend to interact with robots as if they were social agents framed in a specifically intended social situation while at the same time being aware of its technical nature and switching smoothly from “playing along” to treating it like a technical tool. C&F solve this riddle by proposing three levels at which a social robot is construed: The raw artifact, its depiction, and the scene depicted between which human interactants switch seemingly effortlessly. This approach elegantly explains the contradicting observations.

The switch from “playing along” to treating robots as technical agents is stated to happen “effortlessly,” “smoothly,” “implicitly,” “automatically,” “unconsciously,” and it has been proposed that “people are predisposed” or they use “natural rules” of communication. This assumption is in line with Nass and coworkers on stereotypes and research on anthropomorphism (e.g., Nass & Moon, 2000; Zlotowski et al., 2018).

Here, we argue that unexpected and difficult to interpret “breaks” or “interruptions” in the interaction, such as for example, when the robot is crashing, falling, or shutting down, provide a valuable source of information about the “social artifact puzzle.” When human partners are urged to deal with questions such as “Did the character fall asleep, or did the robot’s battery die?,” such “breaks” may shake up the human interaction partner to switch from an automatic interaction style to a more conscious process that requires more explicit strategies. From this we derive the following three assumptions:

1. *Breaks structure interaction into phases that require different processing approaches:* As C&F noticed: “As we noted at the beginning, when a robot stops moving, viewers must decide, ‘Did the character fall asleep, or did the robot’s battery die?’” (target article, introduction, para. 2). Thus, while the interaction at the level of the scene depicted seems to progress rather effortlessly, making use of intuitive human interaction strategies that are strengthened by the anthropomorphization of the robot, the interaction at the “raw artifact level” requires explicit reasoning processes in order to try to find an explanation of the (unexpected) robot behavior. In line with this, studies indicate that during human–robot interaction (HRI) the interaction with a robot is facilitated when the users had a better understanding of the architecture, that is, the raw artifact, and thus were better able to derive the reasons for interaction errors (Hindemith, Göpfert, Wiebel-Herboth, Wrede, & Vollmer, 2021). Moreover, higher anthropomorphism scores, that is, perceiving the robot as more human-like, were associated with a decreased understanding of interaction errors (Hindemith et al., 2021) and less interaction success (Hindemith et al., 2021), suggesting that a convincingly depicted scene, as indicated by high anthropomorphism scores, hindered the correct processing of the raw artifact. These findings are in line with neurobiological investigations of HRI showing that brain regions associated with theorizing about another agent’s putative intentions were increasingly engaged the more human-like the scene was depicted (Hegel, Krach, Kircher, Wrede, & Sagerer, 2008; Krach et al., 2008).

2. *How do prior experiences, expertise, or maturity affect these processes?:* Vollmer, Read, Trippas, and Belpaeme (2018) showed that children were more likely to “play along” in a social group pressure situation with a robot group than adults who were less affected by the social group pressure exerted by robots. This could indicate that adults, who have more experience with and thus stronger prior beliefs about robot behavior than children, were capable of guiding their attention more strongly to the raw artifact level, thus increasing the effect of the raw artifact on the depicted scene level. Thus, we assume that children will be less inclined to change levels in the interaction with a robot and that more “severe” breaks would be necessary to shake up children during HRI. It is unclear though, how expertise in robotics would affect this process. On the one hand, we would assume that more expertise allows the user to more easily spot when and why things go awry during the interaction with the robot. This would allow experts to switch into an interaction more smoothly at the raw artifact level as compared to more naïve interaction partners (see Fig. 1).

On the other hand, it could be that children more easily immerse in the scene. Why should children become more easily immersed? According to Schilbach and colleagues, for an immersive social interaction at least two factors are required: A dynamic interaction between two agents with high emotional engagement (Pfeiffer, Timmermans, Vogeley, Frith, & Schilbach, 2013; Schilbach et al., 2013). Studies indicate that children have higher engagement during HRI (Burdett, Ikari, & Nakawake, 2022) thus one may reason that emotional engagement modulates how easily children may get out of the scene and change to the “raw artifact” level.

These thoughts finally lead to the question:

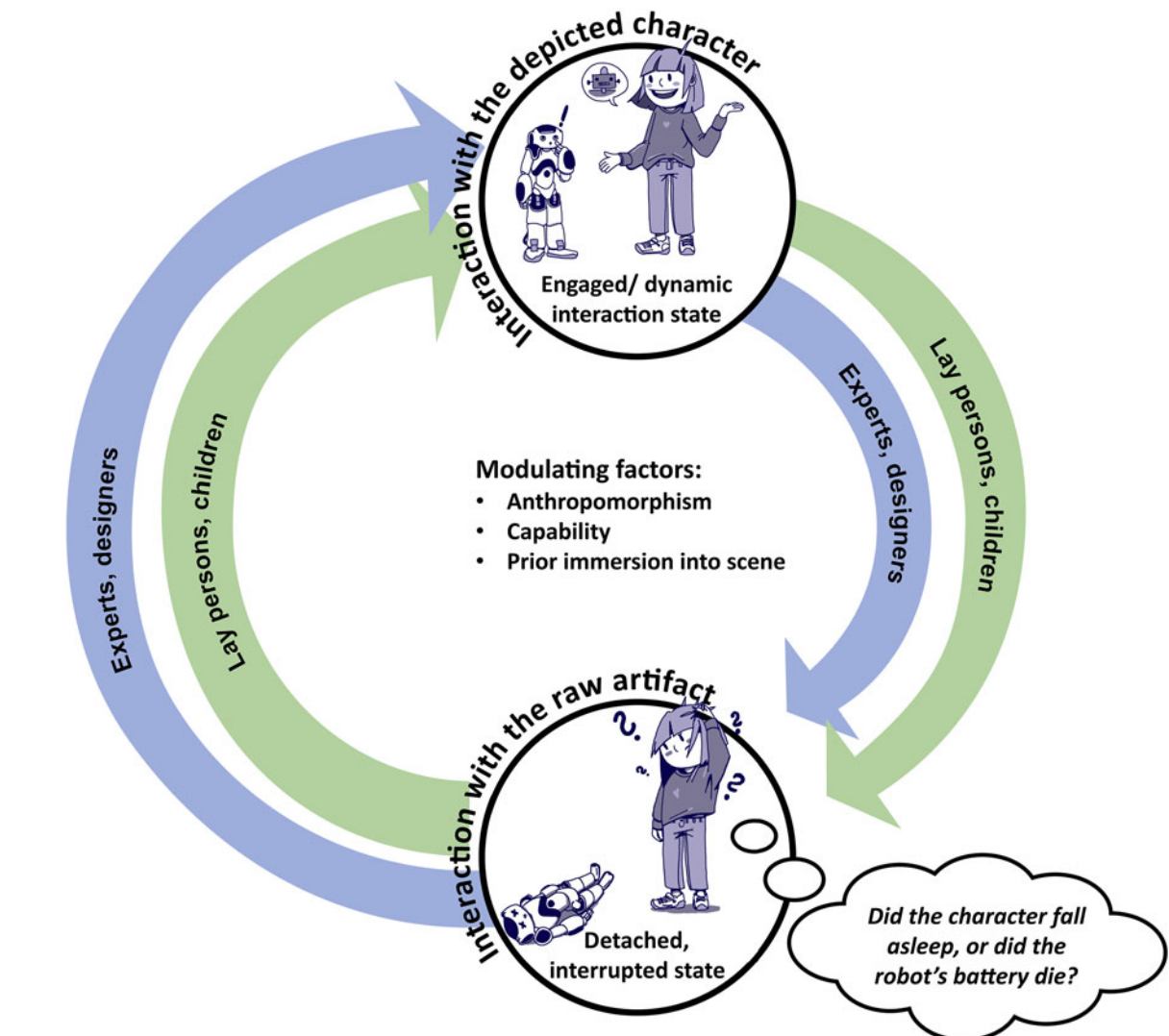
3. *What can roboticists and robot designers learn from these observations and how can insights be derived from these to improve HRI?:* As robots are based on fundamentally different

## Re-Establishment of immersive state

(relatively slow process)

## Loss of immersive state

(relatively fast process)



**Figure 1.** (Wrede et al.) The proposed process of change between “playing along” and treating robots as technical agents caused by an interaction breakdown and vice versa.

architectures than humans, their interaction – at least at the current state – is fundamentally different from human interaction even when developers try to mimic human-like behavior. Thus, human interaction partners need to be able to change from time to time to the “raw artifact” level in order to be able to understand the underlying rules of the artificial interaction with the robot. In didactics of computer science, the changes between function (i.e., depicted scene) and structure (i.e., raw artifact) are seen as an important strategy for learners to comprehend computational artifacts (Schulte, 2008). This leads to the question of how to use such breaking points for HRI? It may be useful, for example, to experimentally control and insert failures within the interaction to help humans learn and better understand how the robot works. On the other hand, what strategies can help to guide the user back to an implicit and smoother interaction?

Overall, these considerations indicate that breakdowns may serve an important role in HRI and deserve further research.

**Acknowledgment.** We thank Helen Beierling for the two illustrations of human-robot interactions.

**Financial support.** BW and ALV received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824.

**Competing interest.** None.


### References

- Burdett, E. R. R., Ikari, S., & Nakawake, Y. (2022). British children’s and adults’ perceptions of robots. *Human Behavior and Emerging Technologies*, 2022(January), 1–16.
- Hegel, F., Krach S., Kircher T., Wrede B., & Sagerer G. (2008). Understanding Social Robots: A User Study on Anthropomorphism. *RO-MAN 2008 – The 17th IEEE International Symposium on Robot and Human Interactive Communication*, August, pp. 574–579. IEEE.
- Hindemith, L., Göpfert, J. P., Wiebel-Herboth, C. B., Wrede, B., & Vollmer, A.-L. (2021). Why robots should be technical. *Interaction Studies*, 22(2), 244–279.



- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE*, 3(7), 11.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *The Journal of Social Issues*, 56(1), 81–103.
- Pfeiffer, U. J., Timmermans, B., Vogeley, K., Frith, C. D., & Schilbach, L. (2013). Towards a neuroscience of social interaction. *Frontiers in Human Neuroscience*, 7(February), 22.
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, 36(4), 393–414.
- Schulte, C. (2008). Duality reconstruction – Teaching digital artifacts from a socio-technical perspective. *ISSEP* (2008).
- Vollmer, A.-L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, 3(21), eaat7111.
- Zlotowski, J., Sumioka, H., Eyssele, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics*, 10(5), 701–714.

## How puzzling is the social artifact puzzle?

Tom Ziemke  and Sam Thellman 

Department of Computer and Information Science, Linköping University, 58183 Linköping, Sweden

[tom.ziemke@liu.se](mailto:tom.ziemke@liu.se); <https://liu.se/en/employee/tomzi64>

[sam.thellman@liu.se](mailto:sam.thellman@liu.se); <https://liu.se/en/employee/samth78>

doi:10.1017/S0140525X22001571, e50

### Abstract

In this commentary we would like to question (a) Clark and Fischer's characterization of the "social artifact puzzle" – which we consider less puzzling than the authors, and (b) their account of social robots as depictions involving three physical scenes – which to us seems unnecessarily complex. We contrast the authors' model with a more parsimonious account based on attributions.

We fully agree with Clark and Fischer's (C&F's) conclusion that no new ontological category is required for understanding people's interactions with social robots. What we would like to question in this commentary, however, is (a) the authors' characterization of the "social artifact puzzle" (target article, sect. 1, para. 2) – which we consider less puzzling than the authors describe it, and (b) their account of social robots as depictions involving three physical scenes – which to us seems unnecessarily complex.

Our own perspective is roughly in line with what C&F characterize as the trait attribution approach. We have recently published a systematic review of 155 empirical studies of mental state attribution to robots (Thellman, de Graaf, & Ziemke, 2022), which shows that most research so far has been concerned with determinants (causes) and consequences, that is, the questions *when* do people attribute mental states to robots, and *why*? Known determinants include robot factors, such as appearance and behavior, and human factors, such as age and motivation. Known consequences include increased predictability, explainability, and trust, but also increases in cognitive drain and moral concern. However, relatively little is known about the

*how*, that is, the mechanisms underlying such attributions – and this is of course where C&F's account of social robots as depictions involving three physical scenes could potentially make an important contribution.

We think that the three-physical-scenes account works best in cases where there is a clear difference between the depiction and the depicted. When, for example, you see the actor Mark Hamill portraying Luke Skywalker in *The Empire Strikes Back*, it is easy for viewers to understand Luke's physical and psychological pain when he gets his hand chopped off by Darth Vader, who then also turns out to be Luke's father, although it is of course more or less clear to everybody that the actor experiences neither of those pains. Things are less clear, we think, in C&F's example of Kermit the frog depicting "a ranarian creature named Kermit" (target article, sect. 10, para. 2). It seems to us that in this case the distinction between the Kermit that does the depicting and the Kermit that is being depicted might not be particularly useful. One might also ask what motivates the limitation to exactly three physical scenes? Is not Kermit (the depicted) himself also a depiction of a certain type of human personality, rather than just a ranarian creature? Is not the fact that Kermit and Piggy are depictions of very different human personality types part of the reason why their relationship is funny to us? Are these examples of a possible fourth level in C&F's model, or alternative third scenes, or maybe a blended third scene? In cases like this, in our opinion, the attribution account is preferable, because it seems relatively straightforward to view people as attributing any number and combination of human, ranarian, and possibly other traits to Kermit.

To get back to socially interactive artifacts, let us take a concrete example (cf. Thellman et al., 2022; Ziemke, 2020): As a pedestrian encountering a driverless car at a crosswalk, you might be asking yourself: Has that car seen me? Does it understand I want to cross the road? Does it intend to stop for me? This would be an example of Dennett's (1988) *intentional stance*, that is, an interpretation of the car's behavior in terms of attributed mental states, such as beliefs and intentions. C&F's analysis in terms of three types of agents is clearly also applicable here: We have the self-driving car, the pedestrian, and the authorities responsible for the car (maker, owner, etc.). If we look at this in terms of C&F's three physical scenes, though, we are again (as in Kermit's case) not quite sure who or what is the character depicted. Is the software controlling the car a depiction of a human driver? That seems unlikely, given that the software as such usually remains invisible to the pedestrian. Or is the self-driving car as a whole a depiction of a normal, human-driven car? This might be in line with arguments that people should not even need to know whether a car is self-driving or not. Or is the car as such a depiction of a self-driving car? It is not clear to us why one would want to distinguish between the depiction and the depicted here. Instead of interpreting this case in terms of three physical scenes, it seems more straightforward to distinguish between the physical car and people's attributions to that car. Moreover, from the perspective of situated and embodied cognition, it would also seem more straightforward to view the pedestrian as interacting with the car in front of it – rather than interacting with some internal representation or an imagined depicted character. In other words, we think the attribution account is more parsimonious, and therefore preferable.

To get back to C&F's notion of the "social artifact puzzle," we do not agree that there is something "self-contradictory, even irrational" about the fact that "people are willing to interact with a

robot as if it was a social agent when they know it is a mechanical artifact” (target article, sect. 1, para. 2). In the example above, instead of the intentional interpretation, the pedestrian could of course take what Dennett refers to as the design stance and predict the car’s behavior based on the general assumption that such vehicles are designed to detect people and not harm them. That might seem safer or more appropriate to some pedestrians (and readers) but note that this would still require you to make additional, more situation-specific assumptions about whether the car has actually detected you (Thellman & Ziemke, 2021; Ziemke, 2020). This brings us back to what we said earlier about the consequences of mental state attribution to robots: In a nutshell, such attributions have been found to increase predictability and trust, which means that treating such artifacts as intentional, social agents might simply make them easier to interact with. In that sense, C&F’s “social artifact puzzle” is less puzzling than it might seem.

**Financial support.** Both authors are supported by ELLIIT, the Excellence Center at Linköping-Lund in Information Technology (<https://elliit.se>).



**Competing interest.** None.

## References

- Dennett, D. C. (1988). Précis of The Intentional Stance. *Behavioral and Brain Sciences*, 11(3), 495–505.
- Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human–Robot Interaction*, 11(4), article 41 (51 pages). <https://doi.org/10.1145/3526112>
- Thellman, S., & Ziemke, T. (2021). The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human–Robot Interaction*, 10(3), article 29 (15 pages). <https://doi.org/10.1145/3461781>
- Ziemke, T. (2020). Understanding robots. *Science Robotics*, 5(46), eabe2987. <https://doi.org/10.1126/scirobotics.abe2987>

## Authors’ Response

### On depicting social agents

Herbert H. Clark<sup>a</sup>  and Kerstin Fischer<sup>b</sup> 

<sup>a</sup>Department of Psychology, Stanford University, Stanford, CA 94305-2130, USA and <sup>b</sup>Department of Design and Communication, University of Southern Denmark, DK-6400 Sønderborg, Denmark  
[clark@stanford.edu](mailto:clark@stanford.edu); [web.stanford.edu/~clark/](http://web.stanford.edu/~clark/)  
[kerstin@sdu.dk](mailto:kerstin@sdu.dk); [www.sdu.dk/ansat/kerstin](http://www.sdu.dk/ansat/kerstin)

doi:10.1017/S0140525X22002825, e51

#### Abstract

We take up issues raised in the commentaries about our proposal that social robots are depictions of social agents. Among these issues are the realism of social agents, experiencing robots, communicating with robots, anthropomorphism, and attributing traits to robots. We end with comments about the future of social robots.

The two of us thank the commentators for their thoughtful reflections on our proposal. Although we cannot possibly address all of the issues they raised, we consider the ones most critical to our account.

Our proposal, briefly, is that social robots are designed to be depictions of social agents. People construe the humanoid robot Asimo, for example, as a depiction of a humanlike character, a social agent, who is able to engage with them in a genuine social interaction. We will call this the *depiction model of social robots*. In what follows, we have grouped the main issues raised in the commentaries into several categories – realism, experience with robots, communication with robots, anthropomorphism, and traits – and we take them up in turn. We end with a brief look at the past and future of social robots.

### R1. Real social agents

A theme running through many of the commentaries is captured in the title of Eng, Chi, & Gray’s (Eng et al.) commentary: *People treat social robots as real social agents*. Here are a few related claims (most italics are ours):

- (1) Eng et al.: “Robots are not human beings, but neither are they mere depictions of social agents. Instead, they are seen as *real* social agents, especially when people interact with them.”
- (2) Friedman & Tasimi: “While it might be difficult to confirm that social robots are viewed as depictions, it may be easier to confirm when they are viewed as *genuine agents*.”
- (3) Orgs & Cross: “The robot [Smooth] performs a *genuine* social interaction: one physically embodied, social agent offers an object to another physically embodied, social agent. The robot therefore does not *pose* as a social agent, it *is* a *genuine* social agent.”
- (4) Stibel & Barrett: “The question of what makes something a *real*, or *actual*, agent is largely a philosophical question. The question of when people perceive, or construe, an entity as a *real* agent is a question for psychology and anthropology.”

But what does it mean for a social agent to be real or genuine? About this there is much confusion.

#### R1.1 Real versus realistic

The bare nouns “tree,” “gun,” and “dog” ordinarily denote trees, guns, and dogs that are real or genuine. The phrases “artificial tree,” “fake gun,” and “toy dog,” on the other hand, denote *depictions* of trees, guns, and dogs as used, for example, in the theater or make-believe play. Artificial trees, fake guns, and toy dogs can be described as “realistic,” but not as “real” (*Oxford English Dictionary*). In these cases, “real” (or “genuine”) contrasts with “artificial.” For something to be construed as real or genuine, it needs to pass two tests:

*Reality test:* If an object or event is real, it “can be said to be real or actual, to be *really* or *actually* or *literally* occurring” (Goffman, 1974).

*Realism test:* If an object or event is a “real X,” it cannot be described as a “realistic X,” and vice versa.

The humanoid robot Nao, in our account, is therefore an *artificial* social agent that depicts a *genuine* social agent. And for us, a genuine social agent is a living being that is able to interact socially with humans. So, although Nao<sub>prop</sub> might be described as “realistic” for a social being, it would not be described as a “real” social being. Hortensius and Cross (2018, p. 93) agree: “We use the term artificial agents to refer to robots (including those that are machine-like, pet-like, or human-like).”

Many commentators, however, describe social robots such as Nao<sub>prop</sub> as genuine social agents. Orgs & Cross, contrary to Hortensius and Cross, insist “The robot [Smooth] does not pose as a social agent, it *is* a genuine social agent” even though it doesn’t pass either test for being a real social agent.

Eng et al. say “Robots are not human beings, but neither are they mere depictions of social agents. Instead, they are seen as *real* social agents.” But for an object to be “seen as” a real X, the object cannot *itself* be a real X. That is, to say that a social robot is *seen as* a real social agent is to imply that it is *not* a real social agent – that it only *looks like* one, that it *depicts* a real social agent. So, despite their objections, Eng et al. appear to agree that social robots are depictions of social agents.

## R1.2 Imagination

Still, people interacting with social robots *imagine* that they are interacting with real social agents. Owners of Sony’s robot dog Aibo, for example, offer spontaneous reports such as these (Friedman, Kahn, & Hagman, 2003): “I feel I care about him as a pal.” “He always makes me feel better when things aren’t so great.” “My emotional attachment to him ... is strong enough that I consider him to be part of my family, that he’s not just a ‘toy.’” Many of these “feelings,” we argue, are based on people’s engagement, engrossment, or immersion in the scenes depicted, a phenomenon that has long been recognized in novels, films, and plays.

Novels and films, according to Chatman (1980), divide into a *discourse* (the medium people process) and a *story* (the content they are to imagine), and people get engrossed in the content (Clark, 1996, p. 366; see Clark & Van Der Wege, 2015). As Gardner (1985, p. 132) puts it about novels, “The writer’s intent is that the reader fall through the printed page into the scene represented.”

Literary theorists, Gerrig notes, call this experience an *aesthetic illusion*. He cites Wolf (2009, p. 144), who said that the illusion “consists predominantly of a feeling, with variable intensity, of being imaginatively and emotionally immersed in a represented world and of experiencing this world in a way similar (but not identical) to real life.” In Wolf’s view, being immersed can range from “the disinterested observation of an artifact” to “the complete immersion (‘psychological participation’) in the represented world” (p. 144). Much the same feelings arise, we suggest, with ventriloquist dummies, hand puppets, and social robots.

## R1.3 Supporting imagination

Orgs & Cross assert that “Clark and Fischer link the quality of a social robot to its resemblance to a human agent.” This is a serious misreading. “All social robots,” we wrote, “represent *nonstandard* characters,” beings that one may never have met, seen, or thought about before. And the “quality of a social robot” is tied not to its literal resemblance to a character, but to the *depictive devices* by which the character is represented (see Clark & Van Der Wege, 2015). Here are a few such devices that film makers, play directors, and puppeteers have used to immerse people in the scenes depicted.

### R1.3.1 Perceptual illusions

A perceptual illusion is a perceptual experience that people *feel* is true even though they *know*, intellectually, that it is *not* true (see Gendler, 2008, on *aliefs*). Movies and plays are packed with them.

Some are visual (fake blood, fake knives, stunt actors, artificial scenery), and others are auditory (Foley effects, diegetic sounds, dubbing).

Social robots also rely on perceptual illusions. One of these, fittingly, is the *ventriloquist illusion*: We hear a ventriloquist’s voice as coming from the mouth of the dummy even though we know it is coming from the mouth of the ventriloquist. The robot Smooth’s voice, for example, comes from its ears, and the robot Asimo’s voice comes from its chest, and yet both voices are heard as coming from their mouths.

### R1.3.2 Concealment

Movie and stage directors try to conceal elements of scenes that are not depictive – such as the lighting, director, and stage crew. One reason is to distinguish outside elements from the depiction proper. Another is to avoid distractions that interfere with people’s immersion in a scene. With social robots, the machinery is generally concealed inside the body and head. Kismet the robot is an exception. It consists of eyes, ears, and lips hung from a visible metal frame, and sure enough, one child interpreted the metal frame as hair (Turkle, Breazeal, Dasté, & Scassellati, 2006, p. 324).

### R1.3.3 Disguise

In early performances of *Hamlet*, Ophelia was played by boys disguised as women, and recently *Hamlet* has been played by women disguised as men. In robots, camera lenses may be disguised as eyes, microphones as neckpieces, and loudspeakers as ears, yet people seem not to notice, or care.

### R1.3.4 Caricature

In animated cartoons, most characters are caricatures. Mickey Mouse’s head, ears, and feet are exaggeratedly large, and so is Porky Pig’s head. The actions depicted in cartoons are also caricatured (Thomas & Johnston, 1995). Objects in motion are squashed and stretched in unnatural ways, and characters exaggerate their starts, stops, and other movements.

The same is true of social robots. Nao, for example, has huge arms, legs, and shoulders, but very small hips, a caricature of an adult male. Despite its antirealism, caricature is often helpful. Drawings of faces are recognized more quickly when caricatured than when veridical (Rhodes, Brennan, & Carey, 1987). And people at times prefer abstractly designed social robots over more realistic ones (Hegel, 2012).

### R1.3.5 Feature selectivity

All depictions, we argued (target article, sect. 5.1), are selective about which features are depictive and which are not. Social robots are no exception. Nao, for example, has “eyes” and “ears” at the correct locations on its head, but it “sees” through cameras in its mouth and forehead and “hears” through microphones in its forehead. And Nao’s “ears” are loudspeakers that depend on the ventriloquist illusion, a fact they must ignore. Nao’s realistic “eyes” and “ears” help people see it as a depiction of a humanlike being even though these do not function as sense organs.

The point is that perceptual illusions, concealment, and disguise add realism to depictions whereas caricature and feature selectivity do not. And yet all five devices help engage, engross, or immerse people in the scenes depicted. The same techniques are exploited in social robots. Nao is a good example.



## R.2. Real experiences and real accomplishments

Many commentaries (e.g., Eng et al., Stibel & Barrett, Orgs & Cross, and Vogeley) observe that people interacting with social robots have *real* experiences – real emotions, overt physical reactions, genuine feelings of responsibility – and that they accomplish *real* goals – from kicking balls back and forth to exchanging real information. Some of the commentaries take these as evidence *against* the depiction model, but that is a mistake.

The depiction model predicts just these phenomena. The character depicted by a social robot is selectively embodied in the robot's prop: The body of Nao's *character* coincides part-by-part with the body of Nao's *prop*, and the movements and speech of Nao's character coincide moment-by-moment with the motions and sounds of Nao's prop. People interact socially with Nao's character by engaging part-by-part and moment-by-moment with Nao's prop, and that leads to real experiences and real achievements. Several commentaries add evidence for this view.

### R2.1 Real experiences

Reeves argues that engaging with social robots includes not only imagined experiences “guided by pretense,” but “natural experiences that are direct, automatic and independent of any thoughtful mapping between what is real and depicted.” At an IMAX film, we are surrounded by an 18 × 24 meter screen, and when the camera goes over a mountain ridge, we feel our stomachs rise into our throats. Reeves calls experiences like this “natural responses.” (See also Förster, Broz & Neerincx [Förster et al.], Seibt, and Vogeley.)

Natural responses, Reeves argues, are a product of what Kahneman (2011, 2012) called system 1 thinking. System 1 is fast, intuitive, and involuntary, whereas system 2 is slow and “performs complex computations and intentional actions, mental as well as physical” (Kahneman, 2012, p. 57). The *time-locked* processes we described in section 7.3 belong to system 1. The audience at *Hamlet* must imagine Hamlet stabbing Polonius at precisely the same time as the actor playing Hamlet is “stabbing” the actor playing Polonius. The *percept-based* processes we discussed in section 7.3 also belong to system 1. People are usually able, without reflection, to recognize an apple as an apple. Both of these processes would be natural responses.

Reeves concludes with a significant insight: “Much of the history of media technology is about inventions that promote natural responses.” This is especially clear in depiction-based media, such as film, television, video, and telepresence technology. Reeves' point applies just as forcefully to social robots. People's experience with them seems real because it is based in part on natural responses.

### R2.2 Experiencing emotions

If system 1 “generates emotions” as Kahneman (2012, p. 57) argued, then emotions should be part of people's experience with depicted scenes. In research we cited (Gross, Fredrickson, & Levenson, 1994), students viewing a clip from the film *Steel Magnolias* often became so immersed in the story that they got sad and cried. Clips from other films reliably evoke emotions ranging from amusement, anger, and contentment to disgust, fear, and sadness (Gross & Levenson, 1995).

These emotions, Blatter & Weber-Guskar note, are “cases of what others have called *fictional emotions*.” They cite Gendler

and Kovakovich (2006), who contrast “real” emotions, which are about real situations, with “fictional” emotions, which are about fictional ones. But emotions require a finer analysis.

#### R2.2.1 Emotions proper

For emotion theorists like Gross and colleagues, an emotion is real regardless of its source. The sadness experienced in *Steel Magnolias* was real even though it was about a fictional scene. Blatter & Weber-Guskar seem to agree: “In all these cases, we know that these characters are fictional, but having followed their stories we feel emotions that are very similar to the ones we would feel for real people.”

#### R2.2.2 Sources

Many emotions have identifiable sources. People fear a gunman, worry about the weather, and feel compassion for an ailing sister. Emotions like these depend on whether the source is real or fictional (*à la* Gendler) and whether it is present or not. As we noted in section 9.2, owners of the robot dog Aibo become emotionally attached to it even though they recognize that it is an artificial agent.

#### R2.2.3 Motivated reactions

People's emotions often motivate further actions. At *Hamlet*, the audience experiences shock when the actor playing Hamlet suddenly “stabs” the actor playing Polonius. For an actual stabbing, people would intervene or call for help, but the audience at *Hamlet* does not do this (see Walton, 1978). People can also regulate or suppress their emotions; in horror films, they can cover their eyes or leave the building (Gross, 2008).

As Gerrig notes, people don't always suppress these reactions. When Clark watches crime films at home on television, he sometimes yells at characters “Watch out! Watch out!” despite frowns from his wife. Informal reports suggest that reactions like these are common. As Gerrig argues, “in the moment, the experience of an aesthetic illusion generates behavior that is real rather than pretense.” Clark construes his yelling as extensions of his emotional responses, which *are* real in the moment. So, when Aibo owners experience real emotional satisfaction in playing with their robots, that is in line with the depiction model.

### R2.3 Continuity of experience

Aesthetic illusions with novels and films tend to be continuous. Once people immerse themselves in a story, they stay immersed in it until they break out of it. The same should hold in people's engagement with social robots.

Rueben takes a different view. “There are reasons to suspect that meta-cognition about construing social robots as depictions would be more difficult – or absent – than Clark and Fischer discuss.” He goes on: “The amount of time and effort that participants give to this reflection could greatly affect their responses.” But in novels, people's immersion is continuous; they don't have to re-immersion themselves with each new sentence or paragraph. The same is true with social robots. People don't need extra “time and effort” for “reflection” at each new step of their interaction with a robot. Once engaged with a robot, people can stay engaged.

A final point is due to Wrede, Vollmer & Krach (Wrede et al.) (see also Healey, Howes, Kempson, Mills, Purver, Gregoromichalaki, Eshghi & Hough [Healey et al.]). People find it easy to stay immersed in an imagined scene as long as it goes smoothly. But once they notice an inconsistency in the evidence, the spell is

broken, they experience a breakdown, and the physical prop is foregrounded. The same happens in the theater when an actor forgets a line, the scenery falls over, or a stage light burns out. Breakdowns like these remind viewers of the base and depiction proper that lie behind the scene depicted. In our example, “When a robot stops moving, people must decide ‘Did the social agent fall asleep, or did the artifact’s battery die?’” And people may go for one interpretation one minute and another the next (Fischer, 2021).

### R3. Are social robots “mere depictions”?

In their commentary, Hortensius & Wiese say “[In] the framework put forward by Clark and Fischer ... people construe social robots as *mere depictions* of social agents,” and others make similar comments (our italics):

- (1) **Eng et al.**: “Research finds that – in real life – people also treat robots as actual social agents, not as *mere depictions* of social agents.” “[T]he more lifelike robots become, the more we treat them like social agents themselves, not *mere depictions*.”
- (2) **Förster et al.**: “Firstly, we argue that robots do constitute a separate category of beings in people’s minds rather than being *mere depictions* of nonrobotic characters.”
- (3) **Friedman & Tasimi**: “How can we tell if other people think they are dealing with a genuine social agent or a *mere depiction* of one?” “So rather than viewing robots as *mere depictions*, people might instead see them as genuine agents with limited moral worth and limited mental capacities.”
- (4) **Gillath, Abumusab, Ai, Branicky, Davison, Rulo, Symons & Thomas (Gillath et al.)**: “Even if Clark and Fischer are correct in suggesting that bots are *merely interactive depictions*, the interactions people have with them are inevitably embedded within social contexts and involve specific social roles.”
- (5) **Girouard-Hallam & Danovitch**: “A developmental and ontological perspective on social robots may move the conversation beyond *mere depiction* to a deeper understanding of the role social robots play in our daily lives and how we view them in turn.”
- (6) **Haber & Corriveau**: “Taken together, these data support the idea that children engage with social robots in much the same way as they do with other social informants – and importantly, *not simply as interactive depictions*.”
- (7) **Malle & Zhao**: “Most current social robots are *mere depictions*.” “[T]he more lifelike robots become, the more we treat them like social agents themselves, not *mere depictions*.”
- (8) **Seibt**: “The authors’ core assumption, however, that social robots are always and only experienced as depictions of social agents, *rather than as social agents proper*, seems problematic.”

To describe a depiction as a *mere* depiction, however, is to ignore its content – the scene people are to get engrossed in. It would be absurd to describe *King Lear*, *The Merchant of Venice*, and *Othello* as *mere* depictions. Shakespeare’s genius lay *first* in creating the stories about Lear, Shylock, and Othello and *then* in creating plays that immerse us in those stories. Yes, Shakespeare wrote magnificent dialogue, yet it is the stories that audiences get engrossed in and remember afterward. It is equally absurd to describe social robots as *mere* depictions. To do so devalues the thought and skill that engineers and social scientists put into their creations.

Comments like these reveal a misunderstanding of what it is to be a depiction, a concept characterized more fully in previous papers (Clark, 1996, 2016, 2019; Clark & Gerrig, 1990). Here we sort out some of those misunderstandings.

#### R3.1 Beyond “mere depictions”

At its heart, a depiction is a representation of something else – a *sign* that signifies an *object*. The philosopher Peirce (1932, 1974; Atkin, 2010) argued that signs come in three main types. (1) A *symbol* signifies an object by rule. The sound /hunt/ signifies “dog” for German speakers by a rule of German. (2) An *index* signifies an object by a physical connection with the object. An arrow is an index that signifies the thing it points at. And (3) an *icon* signifies an object by its perceptual resemblance to the object. A video of a dog barking at a squirrel is an icon that signifies the scene by its visual and auditory resemblance to that scene. Many signs, Peirce noted, are mixed signs – combinations of two or three of the basic types. (Petersen & Almor, alas, overlooked our citations in criticizing us for not tying our model to Peirce, signs, and icons.)

People communicate by producing signs for each other, and that leads to three *methods* of communicating: (1) *describing* things with symbols; (2) *indicating* things with indexes; and (3) *depicting* things with icons. Depicting is, therefore, a basic method of communication on a par with describing and indicating (Clark, 2016). Most acts of communication are composites of these methods.

Acts of communication, in turn, are based on the recognition of a producer’s intention in producing them (Grice, 1957, 1969). When Kate tells Lionel “I caught a fish this long (*holding up two hands, palms in, 30 cm apart*),” she intends him to recognize what she means by her gestural depiction – that the fish was 30 cm long – from two types of information: her perceptual display (the content, place, and timing of her gesture); and her intention, or purpose, in producing the display (as expressed in part in “I caught a fish this long”).

Kate and Lionel’s actions aren’t *unilateral* – separate and autonomous. They are *bilateral* – conditional on each other. Kate has to coordinate her display (its content, placement, and timing) with Lionel’s interpretation of her display. In Grice’s account, these two actions are conditional on each other even when they are displaced in space, time, or both. The same requirement holds for depictions such as social robots.

#### R3.2 On modern art

Depictions, in this view, have a purpose that recipients are intended to recognize. **Orgs & Cross** disagree. “Much of contemporary art,” they argue, “neither depicts nor represents.” As evidence, they cite delightful examples from performance art such as dance and theater that “dissolve the binary distinction between depicted and depictive scene, or acting and not-acting.”

**Orgs & Cross**’s argument, however, ignores purpose. The very point of much modern art is to have no practical point. Artists have license to entertain, divert, or fascinate however they like and often leave purpose indeterminate. When Andy Warhol painted “Campbell’s Soup Cans,” why did he depict Campbell’s soup cans, and why 48 of them? Why did Jackson Pollack drip paint on a canvas in the patterns he did? Why are so many works entitled “Untitled”? Other artists play with *trompe*

*l'oeil*, deceptive perspective, and visual illusions. Orgs & Cross's examples are of this ilk, and viewers appreciate them for what they are.

Everyday depictions, however, have a *practical* purpose. People base their interpretation of Michelangelo's *David*, Kate's depictive gesture, and social robots in part on what they believe the creators intended. Genuine depictions and artistic creations may live in the same world, but they are not all processed in the same way.

#### R4. Social robots must depict the way agents communicate

The depiction model holds that people construe social robots as depictions of *social agents*. But for an agent to be a *social agent*, it must be able to engage people in social interactions, and to do that, it must be able to communicate. As we put it in section 7, "it takes coordination for two individuals to interact with each other, and they cannot do that without communicating (Clark, 1996)." A social robot must therefore depict not only the agent's physical appearance and movements, but also its acts of communication – its speech, hand gestures, head nods, head shakes, eye gaze, facial gestures, body postures, and body placements (see target article, sect. 7.1). That is, for the depiction of a genuine social agent to be complete, it must include the agent's communicative acts.

To our surprise, acts of communication are not even mentioned in most of the commentaries. Worse yet, they are *in principle* impossible in the alternative models based on anthropomorphism, embodiment, mind perception, and trait attributions. Here we briefly review our own previous work on communication (e.g., Clark, 1996, 2005, 2021; Clark & Brennan, 1991; Clark & Henetz, 2014; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Fischer, 2016, 2021) and then show how it undercuts the alternative models.

##### R4.1 Joint activities

The basic idea is that whenever people interact socially, they do things together: They coordinate with each other in *joint activities* (Clark, 1996, 2005). And to coordinate with each other, they have to *agree* on their joint actions and positions, and that requires communication. Here is an example from two people assembling a TV stand (from Clark, 2005):

Ann: Should we put this (*holding up piece of wood*) in, this, this little like kinda cross bar (*pointing at a picture on the directions for the TV stand*), like the T? like the I bar?

Burton: Yes, we can do that.

In turn 1, Ann *proposes* a joint position for the two of them, and in turn 2, Burton *takes up* her proposal and *agrees* to it. People can also reach agreement with gestures, which are any "visible acts of communication" (Kendon, 2004):

Burton: (*extends hand with a peg to Ann*)

Ann: (*grasps the peg*)

The principle is this: "It takes coordination for people to do things together, no matter how simple, and it takes communication to achieve that coordination" (Clark, 2005, p. 507).

Social robots require the same techniques. In section 2.4 of our paper, we illustrated two exchanges between the robot Smooth and a woman named Beth:

Smooth: (*presenting water glasses to Beth*) Take your drink please.

Beth: (*takes a glass of water*)

Smooth: (*faces Beth*) Cheers!

Beth: (*lifting her glass slightly*) Cheers.

In the first pair of turns, Smooth offers Beth water, and she accepts his offer and takes a glass. In the second pair of turns, he makes a toast, and she takes it up and reciprocates. All four turns rely on both speech and gestures. In an example from Guo, Lenchner, Connell, Dholakia, and Muta (2017), a woman asked a robot concierge for the location of a bathroom. She posed the question in turn 1, and he took it up and answered it in turn 2:

Woman: Where is the bathroom.

Robot concierge: The bathroom is in aisle 13.

So, for humans and social robots to coordinate with each other, they must reach agreement on how to get things done together. A model of social robots unable to do this cannot be complete.

##### R4.2 Common ground

When two people communicate, they assume certain information to be part of their current common ground, and they add to that body of information with each new act of communication (Clark, 1996, Ch. 4; Stalnaker, 1978). The same goes for social robots. As Carroll argued, "Future robots must effectively coordinate common ground with humans."

Common ground comes in two main types, and social robots need to track both:

- (1) *Personal* common ground is information people establish based on their joint experiences – what they see, do, and communicate *with each other*. Suppose a woman named Jane asks an *actual* concierge, "Where is the bathroom?" and he answers "The bathroom is in aisle 13." With her question, the two of them would add her request to their current common ground, and with his answer, they would add the location of the bathroom.
- (2) *Communal* common ground is information people share as members of the same cultural communities, such as their nationality, occupation, language, gender, age cohort, or residence. Although Beth, for example, spoke to her friends in Danish, she took for granted that she and Smooth both knew English and spoke to him in English.

For face-to-face coordination to go smoothly, communication must also be reliable. That requires a process called *grounding*: People in joint activities try to establish, as they go along, the mutual belief that they have understood each other well enough for current purposes (Clark, 1996; Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986). Speakers monitor their conversations both for evidence of success (e.g., "uh huh," "good God," "oh," and nods from addressees) and for evidence of failure (e.g., misunderstandings that need



repairing). When one woman asked Guo et al.'s robot concierge something he couldn't understand, he cleared it up before going on:

Woman: I need to powder my nose. (*non-recognized question*)  
 Robot concierge: Can you rephrase the question?  
 Woman: Where is the bathroom.  
 Robot concierge: The bathroom is in aisle 13.

The side sequence in turns 2 and 3 is one of many strategies people use for repairs (Dingemanse et al., 2015; Schegloff, Jefferson, & Sacks, 1977). Social robots, then, also need strategies for tracking success and failure in their social interactions (see Healey et al. and Wrede et al.).

The problem, as we will show, is that alternative models of social robots have no means for accumulating common ground or for grounding what they do and say.

#### R4.3 Social agents are individuals

In communication, common ground is accumulated by *individuals* and not by *types* of individuals (Clark, 1996, 2021). When the actual concierge answered Jane's question, he tried to add new information to the common ground he shared with Jane the individual and not with some generalized person. He tried to *anchor* his references ("the bathroom" and "aisle 13") to an *actual* bathroom and an *actual* aisle he assumed was in his and her current common ground. It is no different for people communicating with a robot concierge.

Individual entities are fundamentally different from types of entities. Thoughts about "Jane" and "the concierge" are about the individuals they index. Thoughts about "woman" and "concierge," in contrast, are about the *types* of individuals they describe. Crucially, an indexical thought cannot be reduced to a set of descriptive thoughts (Perry, 1979, 1993; Recanati, 2012, 2013). An individual like Jane cannot be represented as merely a bundle of attributes. Yet that is the assumption behind many models of social robots based on anthropomorphism, mind perception, and trait attributions (e.g., Girouard-Hallam & Danovitch, Orgs & Cross, Ziemke & Thellman). The point may seem technical, but it is a significant strike against those models.

#### R4.4 Interim summary

In short, for a robot to be a *social* robot, it must represent a real social agent, an individual, able to engage humans in joint activities. The agent must be able to:

- (1) *Coordinate* with humans in joint activities, however restricted the activities,
- (2) *Communicate* with humans well enough to advance these activities,
- (3) *Accumulate* common ground with individual humans as these activities advance (as Carroll suggests),
- (4) *Ground* what gets said and done well enough for current purposes.

A number of proposals in the commentaries are incompatible with these features. In the next two sections, we take up four of these proposals.

## R5. Anthropomorphism

Petersen & Almor make a remarkable claim in the title of their commentary: "Anthropomorphism, not depiction, explains interaction with social robots." But in support of their position, they treat "social responses" and "social behaviors" as if they were genuine "social interactions," and they aren't. Citing Airenti (2018), Petersen & Almor write (with our italics):

For example, when a car engine fails to start, it is not uncommon for the would-be driver to engage in begging, chastising, or other *social behaviors* directed towards the car. It is difficult to argue that the car is a depiction of a social agent. Rather, Airenti argues that the interactive situation itself, in this case noncooperation, is sufficient to provoke a *social response*.

But anthropomorphizing a car doesn't turn the car into a social agent either artificial or real. The driver (call him Joe) directs actions toward the car, but the car just sits there. The two of them don't coordinate with each other. And when Joe "chastises" the car, he is not communicating with it. Whatever Petersen & Almor's "interactive situation" is, it is not a *social* interaction.

Ziemke & Thellman give a vivid example with driverless cars, but it, too, has problems.

As a pedestrian encountering a driverless car at a crosswalk, you might be asking yourself: Has that car seen me? Does it understand I want to cross the road? Does it intend to stop for me? ... [I]t would also seem straightforward to view the pedestrian as interacting with the car in front of it – rather than interacting with some internal representation or an imagined depicted character.

Here again, the interaction is not a *social* interaction. The car and pedestrian do not coordinate with each other as two *social* agents. The car is designed to predict what the pedestrian will do, and the pedestrian tries to predict what the car will do, but they do not communicate with each other about that. The situation is competitive, not cooperative. With social robots, people coordinate by communicating with the social agents the robots depict.

### R5.1 Problems with anthropomorphism

These two examples illustrate serious problems for anthropomorphizing as an account of social robots.

#### R5.1.1 Unilateral versus bilateral interpretations

Anthropomorphizing a thing – viewing it as human – is a *unilateral* action, which people perform on their own. Joe was free to imbue the car with any features he liked, and he chose human ones. But interpreting a depiction is a *bilateral* action, which also takes account of what the depiction is intended to represent. Viewers cannot anthropomorphize Michelangelo's *David* any way they like, and they know that. They try to interpret it as Michelangelo intended them to – as a depiction of the biblical David.

#### R5.1.2 Individual agents

Anthropomorphizing creates *types* of humans and not *individuals*. Yet, as we noted, it takes individual agents to coordinate with other humans, communicate with them, accumulate common ground with them, and ground what they say and do (Clark, 2021).

#### R5.1.3 Communication

Anthropomorphizing an entity does not specify *how* it coordinates with others with speech and gestures.

### R5.1.4 Nonstandard characters

To anthropomorphize an entity is to imbue it with *human* features. But many social robots are a mix of human, animal, and other features. “All social robots” we wrote (contra **Caruana & Cross**), “represent *nonstandard* characters.” They are “best viewed as *composite* characters – combinations of disparate physical and psychological attributes.” The species they belong to don’t come prefabricated. They have to be constructed. Anthropomorphizing simply cannot create the range of creatures that social robots represent.

These problems challenge *in principle* any proposal about social robots that relies on anthropomorphizing or mind perception (**Bigman, Surdel & Ferguson [Bigman et al.]**; **Blatter & Webber-Guskar**; **Carroll**; **Caruana & Cross**; **Doyle & Hodges**; **Eng et al.**, **Goldman**, **Baumann & Poulin-Dubois**; **Orgs & Cross**, and **Ravikumar, Bowen & Anderson [Ravikumar et al.]**).

### R5.2 Embodiment

**Ravikumar et al.** argue for treating social robots as *embodiments* of social robots: “[E]ven if social robots are interactive depictions, people need not mentally represent them as such. Rather, people can directly engage with the opportunities for action or *affordances* that such robots/depictions offer to them.” This position, however, also has problems.

Suppose Brigitte saw her old friend Alain and wanted to talk to him. She knew him well enough to assume they shared a great deal of common ground, such as how to approach, hug, kiss, and gossip with each other. She wouldn’t have known how to approach, hug, kiss, or gossip with an anonymous body. She needed to know it was Alain. Brigitte would have the same problem with the robot Asimo. Identifying an entity as a body is not enough to engage with it even in what Ravikumar et al. called “sociocultural settings.”

Embodiment also doesn’t distinguish features that allow affordances from features that do not. As we noted, “Observers of a depiction implicitly realize that only some of its features are depictive,” and only the depictive features afford the right inferences. Asimo<sub>prop</sub>’s hand depicts a real hand, which affords handshakes, but Asimo<sub>prop</sub>’s ears happen *not* to contain senses of hearing, so they do not afford headphones. Asimo<sub>prop</sub>’s lack of a mouth doesn’t afford speaking, yet Asimo<sub>char</sub> is able to speak. Discrepancies like these differ from robot to robot. If so, how can people “directly engage with the opportunities for action or *affordances* that such robots/depictions offer to them”?

Communicating with a robot is even more of a challenge. With Asimo, should Brigitte speak French, use French gestures, and kiss him on both cheeks, as she would with Alain? The affordances availed by Asimo<sub>prop</sub>’s body offer no answers.

### R5.3 Intentional stance

As the commentaries by **Veit and Browning** and **Ziemke and Thellman** noted, **Dennett (1987, 1988)** proposed two strategies, or stances, that attribute intentions to systems such as social robots:

The *intentional stance* is the strategy of prediction and explanation that attributes beliefs, desires, and other “intentional” states to systems – living and nonliving – and predicts future behavior from what it would be rational for an agent to do, given those beliefs and desires. (1987, p. 495)

In the *design stance*, one predicts the behavior of a system by assuming that it has a certain design (is composed of elements with functions) and that it will behave as it is designed to behave under various circumstances. (1988, p. 496)

There is merit in both stances. For Asimo, the intentional stance applies to the social agent it represents (Asimo<sub>char</sub>), and the design stance applies to Asimo’s physical design (Asimo<sub>prop</sub>).

These two stances, however, are *unilateral* interpretations and not the *bilateral* ones needed for social robots. It isn’t enough to attribute certain mental states to Asimo<sub>char</sub>. People must attribute the mental states they believe they were intended to attribute to Asimo<sub>char</sub>. More than that, their interpretation of Asimo<sub>char</sub> must be custom-built for the nonstandard individual that Asimo depicts.

## R6 Trait attributions

**Bigman et al.** entitle their commentary, “Trait attribution explains human–robot interactions,” and others agree with their claim (e.g., **Eng et al.**, **Ziemke & Thellman**). But as we noted earlier, individuals such as Asimo<sub>char</sub> cannot *in principle* be reduced to bundles of traits, so models based on trait attributions face problems from the start. Trait attributions may be useful in describing or designing social robots (see **Ziemke & Thellman**), but that doesn’t allow bundles of traits to count as models of social robots. Alas, trait attributions have other problems as well.

### R6.1 Measuring traits

Traits are often studied by asking people to rate how much human attributes apply to nonhuman entities (see **Epley, Waytz, & Cacioppo, 2007**; **Gray, Gray, & Wegner, 2007**; **Reeves, Hancock, & Liu, 2020**; **Weisman, Dweck, & Markman, 2017**; see **Thellman, de Graaf, & Ziemke, 2022**). In one study (**Gray et al., 2007**), participants rated 13 “characters,” which ranged from a baby, a fetus, a dead woman, and a frog to God, “you,” and a robot. All but the fetus were given proper names. People rated the entities on dimensions of “experience” (e.g., hunger, pain, fear, pride) and “agency” (e.g., self-control, morality, memory).

What people rated, however, weren’t objects they had interacted with. They were static photos (e.g., **Phillips, Ullman, de Graaf, & Malle, 2017**; **Reeves et al., 2020**; **Ruijten, 2015**), videos, labels (**Lencioni, Carpinella, Rabuffetti, Marzegan, & Ferrarin, 2019**), or descriptions of such objects. Here is **Gray et al.**’s description of their robot:

*Kismet*. Kismet is part of a new class of “sociable” robots that can engage people in natural interaction. To do this, Kismet observes a variety of natural social signals from sound and sight, and delivers his own signals back to the human partner through gaze direction, facial expression, body posture, and vocal babbles.

Clearly, this paragraph isn’t about Kismet<sub>base</sub> or Kismet<sub>prop</sub>, but about Kismet<sub>char</sub>. Only the character would have a proper name, be male, “engage people in natural interaction,” “observe ... social signals,” “deliver *his* own signals,” direct *his* gaze, and produce a “facial expression,” “body posture,” and “vocal babbles.” It is no surprise that Kismet<sub>char</sub> was judged to have agency.

What if people had been asked whether Kismet’s *machinery* experienced hunger, pain, or fear, or possessed pride or morality? Their ratings would surely have changed. Even if they thought Kismet’s machinery experienced “hunger” (as when its battery died), they would have based their judgment on a metaphorical, not literal, interpretation of *hunger*. Metaphors, indeed, are a widespread issue.

## R6.2 Metaphor problems

People seem willing to attribute “mental states” and “minds” to all sorts of artifacts. These include not only cars (Petersen & Almor, Ziemke & Thellman), but gadgets. A study by Epley, Akalis, Waytz, and Cacioppo (2008) examined five gadgets, including “Clocky (a wheeled alarm clock that ‘runs away’ so that you must get up to turn it off)” and “Pillow Mate (a torso-shaped pillow that can be programmed to give a ‘hug’).” People were asked to rate “the extent to which the gadget had ‘a mind of its own,’ had ‘intentions,’ had ‘free will,’ had ‘consciousness,’ and ‘experienced emotions.’”

But what were these people rating? When asked about the extent to which Pillow Mate had “a mind of its own” or “free will,” they were forced to interpret “mind” and “free will” as metaphors. If they had been asked whether Pillow Mate *really or actually or literally* had a mind of its own or free will, they, like us, would have said no (see Thellman et al., 2022). Likewise, if Joe the driver had been asked if he was *really or actually or literally* chastising the car, he, too, would have said no. And so would the pedestrian when asked whether the driverless car could *really or actually or literally* “see” him or her, “understand” things, or have “intentions”? (Did Romeo think that Juliet was *really or actually or literally* “the sun”?)

What we have here are metaphors: “hunger,” “a mind of its own,” “intentions,” “free will,” “consciousness,” “emotions,” “chastising,” “see,” and “understand.” It is a mistake to equate metaphorical attributions like these with their literal counterparts. They are *not* equivalent, and treating them as equivalent leads to misleading claims about both traits and social robots.

## R7. Depicting is universal

A theme running through many commentaries is that depictions are exotic – too complex for people to use and understand easily (cf. **Rueben**). Keeping track of two layers, the depiction proper and the scene depicted, takes too much metacognitive effort. But nothing could be further from the truth. Depicting is a basic method of communication, and it is everywhere.

To begin with, depictions that people perform in conversation, such as direct quotations, iconic gestures, facial gestures, and full-scale demonstrations, are part of all languages. Depicting is also the basis for ideophones such as *meow*, *cock-a-doodle-doo*, and *oink-oink*, and these, too, are part of all languages (Dingemans, 2013). And children begin to use performed depictions from as young as 18 months of age (Clark & Kelly, 2021). Conclusion: People everywhere use and understand performed depictions as part of everyday communication and from an early age.

Other types of depictions have been around since the Lower Paleolithic times. Cave paintings of horses, bulls, and hunters have been found on all continents (except Antarctica) dating from 25,000 to 10,000 BCE. More elaborate paintings, sculptures, and ceramic depictions have been found in Egypt, Greece, China, North America, and Meso-America dating from 2,500 to 1,000 BCE. Theater, puppet shows, and opera-like dramas have been documented in China, India, Greece, and Meso-America from as early as 1,500 BCE. There is nothing new about depictions like these, both static and staged.

Stibel & Barrett, two anthropologists, seem to challenge this view:

Lacking any cultural, personal, or historical concept of the idea of a “robot,” it seems unlikely that a twelfth-century human would take the

object before them as a human-made artifact designed to “depict” authentic agency. More likely, they would construe this unknown entity as a real agent of some kind.

And yet automata, the ancestors of modern robots, were developed in Europe, the Middle East, and China well before the Common Era (Foulkes, 2017). Heron of Alexandria (10–90 CE), for example, designed automata that depicted “a shepherd who gave water to his sheep, and even an articulated bird that could whistle” (Foulkes, 2017, p. 64). Heron in turn inspired the construction of automata throughout Europe and the Middle East, including tabletop marching and fighting armies, flying birds, singing birds, walking lions, a donkey driving a water wheel, and even people playing chess. Truitt (2015) called these “medieval robots.”

Social robots such as Asimo, Smooth, and Nao are introduced nowadays not only physically but with explicit interpretive frameworks. The same robots introduced in the same way should cause no more trouble to Stibel & Barrett’s twelfth-century human than they do to modern humans.

## R8. Other issues

Many issues raised in the commentaries deserve further discussion, but we can consider only a few.

### R8.1 Theory

Is the depiction model a theory (see **Bartneck**)? The answer is clearly yes. A theory, according to Dennis and Kintsch (2007), should satisfy certain criteria, and the depiction model does just that. It accords with empirical data; it is precise and interpretable; it is coherent and consistent; it predicts future applications; and it provides explanations that go beyond the model itself. We have cited evidence supporting each of these criteria.

We have also investigated alternative accounts. The media equation was one of the early inspirations for our work, and while the depiction model makes many of the same predictions, it also explains phenomena not covered by the media equation (see target article, sect. 2.4). In his commentary, **Reeves**, one of the progenitors of the media equation, appears to agree (contra **Bartneck**). Anthropomorphism and trait attributions are two other alternative accounts, but these suffer from the empirical and conceptual problems we discussed earlier. The point of the depiction model, in sum, is to explain social robots in terms of a broader theory, namely, how people engage with depictions, and we believe it succeeds at that.

### R8.2 Social roles

In our paper, we distinguished between *self-agents*, who “act on their own authority and are fully responsible for their actions,” and *rep-agents*, who “act on the authority of specified principals.” When Susan works as a server for Goldberg’s Bakery, she is a rep-agent for the bakery, but once off work, she is on her own, a self-agent. Healey et al. (see also **Carroll**) worried about the roles such agents take.

Each individual person, we assume, has an *individual role* that is continuous and enduring. Susan remains Susan whatever else she does. But individuals also take on additional roles, *social roles*, that change with the social situation. They may take the social role of sister, companion, playgoer, or bus rider for people



they interact with, and teacher, tutor, or concierge in working for others. Susan chose the particular role of server when she hired on at Goldberg's Bakery. Social robots could, in principle, take multiple social roles, but the robots we know of are able to take only one social role.

### R8.3 Future

Predicting the future is dangerous. In about 1970, Herbert A. Simon, one of the founders of artificial intelligence (AI), suggested to colleagues that people would be able to talk with computers in 10–20 years. Fifty years later there is still no such computer. People share limited facts with virtual assistants such as Siri, Alexa, and Google Assistant, but as conversations go, these exchanges are primitive (see Marge et al., 2022). Today's AI systems for conversation still cannot deal with such features as the timing of turns, the use of *uh* and *um*, performed depictions, pointing, anchoring, grounding, irony, sarcasm, and empathy. According to Bender and Koller (2020), current AI models of language cannot be complete *in principle* because they are based on the form of language alone.

So Malle & Zhao are brave souls. They are clear-sighted about today's robots when they say: “[C]urrent social robots are advertised to be much more capable than they really are – that is, they are largely a pretense, a fiction.” But they venture into Herbert Simon territory when they go on (see also Caruana & Cross; Franklin, Awad, Ashton, & Lagnado (Franklin et al.); and Stibel & Barrett):

Now consider what robots will be like in the future. They will not just be depictions; they will instantiate, as robots-proper, the actions that current robots only depict. Unlike dolls and dummies, they will not just be crafted and controlled by human programs. They will rapidly evolve through directing their own learning and devising their own programs. They will increasingly make autonomous decisions enabled by continuously updated and massively expanded algorithms.

As Simon's prediction shows, the future doesn't always work out the way we think – often for principled reasons. With social robots, who knows what those principles will be.

Still, no matter how humanlike social robots become, they will never be humans. They will always be artifacts intended to depict humanlike social agents. To frame them otherwise would be to engage in deception.

### R9. Coda

In 1919, film director Ernst Lubitsch made a silent comedy called “Die Puppe” (“The Doll”). A young man named Lancelot was informed by his rich uncle, a baron, that he had to get married by a certain date if he expected to inherit the family fortune. To trick his uncle, Lancelot arranged to marry an automaton – a beautiful life-sized mechanical doll. The maker of the doll, however, had created the doll in the image of his own beautiful daughter, and he managed to trick Lancelot into marrying his daughter instead of the doll. Happily, it all worked out in the end.

“Die Puppe” appeared the year before Karel Čapek's 1920 play “Rossum's Universal Robots.” Still, in the years that followed, writers needing a word for humanoid automata chose “robot” over “puppet.” What if they had chosen “puppet”? Social robots would now be called “social puppets,” and our claim that “social puppets are depictions of social agents” would be considered a

truism. We would have had no paper, and the commentators would have had nothing to comment on. Thanks to Čapek, but not Lubitsch, we all had lots to say.

### References

- Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, 9(2136), 1–13.
- Atkin, A. (2010). Peirce's theory of signs. In E. N. Zalta & U. Nodelman (Eds.), *Stanford encyclopedia of philosophy*. Stanford University.
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Chatman, S. B. (1980). *Story and discourse: Narrative structure in fiction and film*. Cornell University Press.
- Clark, E. V., & Kelly, B. F. (2021). Constructing a system of communication with gestures and words. In A. Morgenstern & S. Goldin-Meadow (Eds.), *Gesture in language: Development across the lifespan* (pp. 137–156). Walter de Gruyter.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, 7(4–5), 507–525.
- Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, 123(3), 324–347.
- Clark, H. H. (2019). Depicting in communication. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 235–247). MIT Press.
- Clark, H. H. (2021). Anchoring utterances. *Topics in Cognitive Science*, 13(2), 329–350.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.
- Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language*, 66(4), 764–805.
- Clark, H. H., & Henetz, T. (2014). Working together. In T. M. Holtgraves (Ed.), *The Oxford handbook of language and social psychology* (pp. 85–97). Oxford University Press.
- Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13(2), 259–294.
- Clark, H. H., & Van Der Wege, M. A. (2015). Imagination in narratives. In D. Tannen, H. E. Hamilton & D. Schiffrin (Eds.), *Handbook of discourse analysis* (2nd ed., pp. 406–421). John Wiley.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences*, 11(3), 495–505.
- Dennis, S., & Kintsch, W. (2007). Evaluating theories. In D. F. Halpern & R. J. Sternberg (Eds.), *Critical thinking in psychology* (pp. 143–159). Cambridge University Press.
- Dingemans, M. (2013). Ideophones and gesture in everyday speech. *Gesture*, 13(2), 143–165.
- Dingemans, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., ... Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLoS ONE*, 10(9), e0136100.
- Epley, N., Alakis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, 19(2), 114–120.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. doi: 10.1037/0033-295X.114.4.864
- Fischer, K. (2016). *Designing speech for a recipient: The roles of partner modeling, alignment and feedback in so-called “simplified registers”*. John Benjamins.
- Fischer, K. (2021). Tracking anthropomorphizing behavior in human-robot interaction. *ACM Transactions in Human-Robot Interaction*, 11(1), Article 4, 1–28. doi: 10.1145/3442677.
- Foulkes, N. (2017). *Automata*. Éditions Xavier Barral.
- Friedman, B., Kahn Jr, P. H., & Hagman, J. (2003). Hardware Companions? What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Ft. Lauderdale, FL.
- Gardner, J. (1985). *The art of fiction: Notes on craft for young writers*. Vintage Books.
- Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy*, 105(10), 634–663.
- Gendler, T. S., & Kovakovich, K. (2006). Genuine rational fictional emotions. In M. Kieran (Ed.), *Contemporary debates in aesthetics and the philosophy of art* (pp. 241–253). Blackwell.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science* (New York, N.Y.), 315(5812), 619. doi: 10.1126/science.1134475

- Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66, 377–388.
- Grice, H. P. (1969). Utterer's meaning and intention. *The Philosophical Review*, 78(2), 147–177.
- Gross, J. J. (2008). Emotion regulation. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 497–513). Guilford Press.
- Gross, J. J., Fredrickson, B. L., & Levenson, R. W. (1994). The psychophysiology of crying. *Psychophysiology*, 31, 460–468.
- Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9(1), 87–108.
- Guo, S., Lenchner, J., Connell, J., Dholakia, M., & Muta, H. (2017). Conversational bootstrapping and other tricks of a concierge robot. Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction, Vienna, Austria.
- Hegel, F. (2012). Effects of a Robot's Aesthetic Design on the Attribution of Social Capabilities. Paper presented at the 2012 IEEE RO-MAN: The 1st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France.
- Hortensius, R., & Cross, E. S. (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), 93–110.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kahneman, D. (2012). Two systems in the mind. *Bulletin of the American Academy of Arts and Sciences*, 65(2), 55–59.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Lencioni, T., Carpinella, I., Rabuffetti, M., Marzegan, A., & Ferrarin, M. (2019). Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks. *Scientific Data*, 6(1), 1–10.
- Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., ... Dey, D. (2022). Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71, 101255.
- Peirce, C. S. (1932). The icon, index, and symbol. In C. Hartshorne & P. Weiss (Eds.), *Collected papers of Charles Sanders Peirce* (Vol. 2, pp. 156–173). Harvard University Press.
- Peirce, C. S. (1974). *Collected papers of Charles Sanders Peirce* (Vol. 3). Harvard University Press.
- Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13(1), 3–21.
- Perry, J. (1993). *The problem of the essential indexical, and other essays*. Oxford University Press.
- Phillips, E., Ullman, D., de Graaf, M. M., & Malle, B. F. (2017). What does a robot look like?: A multi-site examination of user expectations about robot appearance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 1215–1219.
- Recanati, F. (2012). *Mental files*. Oxford University Press.
- Recanati, F. (2013). Mental files: Replies to my critics. *Disputatio*, 5(36), 205–240.
- Reeves, B., Hancock, J., & Liu, S. X. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior*, 1(1), 1–14.
- Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19(4), 473–497.
- Ruijten, P. A. M. (2015). *Responses to human-like artificial agents*. Uitgeverij BOXPress.
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.
- Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics 9: Pragmatics* (pp. 315–332). Academic Press.
- Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human–Robot Interaction (THRI)*, 11(4), 1–51.
- Thomas, F., & Johnston, O. (1995). *The illusion of life: Disney animation*. Hyperion.
- Truitt, E. R. (2015). *Medieval robots: Mechanism, magic, nature, and art*. University of Pennsylvania Press.
- Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006). Encounters with Kismet and Cog: Children respond to relational artifacts. In P. Messaris & L. Humphreys (Eds.), *Digital media: Transformations in human communication* (pp. 313–330). Peter Lang.
- Walton, K. L. (1978). Fearing fictions. *The Journal of Philosophy*, 75(1), 5–27.
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43), 11374–11379. doi: [10.1073/pnas.1704347114](https://doi.org/10.1073/pnas.1704347114)
- Wolf, W. (2009). Illusion (aesthetic). In P. Hühn, J. Pier, W. Schmid, & J. Schönert (Eds.), *Handbook of narratology* (Vol. 1, pp. 270–287). De Gruyter.