Metaphors Matter: Top-Down Effects on Anthropomorphism

Hailey Austine Scherer

Dartmouth College

Follow this and additional works at: https://digitalcommons.dartmouth.edu/senior_theses

Part of the Cognitive Science Commons

Recommended Citation

https://digitalcommons.dartmouth.edu/senior_theses/270

This Thesis (Undergraduate) is brought to you for free and open access by the Theses and Dissertations at Dartmouth Digital Commons. It has been accepted for inclusion in Dartmouth College Undergraduate Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
METAPHORS MATTER:
TOP-DOWN EFFECTS ON ANTHROPOMORPHISM

Hailey Austine Scherer

Bachelor of Arts Honors Thesis
Class of 2020
Advised by Jonathan Phillips and Luke Chang
Program in Cognitive Science
Dartmouth College
Hanover, NH

May 28, 2020
Acknowledgements

This thesis would not have been possible without the inspiration and support of the people around me.

I would like to thank Shaun Nichols, whose lecture in Introduction to Cognitive Science on the theory of mind my freshman spring inspired me to pursue this topic for the past three years.

I would like to thank Adina Roskies, who through her coursework and our personal discussions has inspired and helped me to structure my thinking in the domains of cognitive science and philosophy.

I would like to profoundly thank all of my friends, especially ‘20s Kensey, Jack, Matthew, Sam, Anca, and Anna, who have weathered my talk about this project for the past two years, and who have supported and encouraged me through it all.

I would especially like to thank Jeffrey Poomkudy ‘20, who has not only made me a drastically better thinker through our conversations in the past four years, but whose insightful comments and patient willingness to talk about the ideas in these pages with me has have made this thesis much better than it ever otherwise would have been.

I would like to thank my family, especially my mother, father, Catherine, and Brian, for their unending support and for tolerating my work schedule even in the midst of COVID-19 quarantine.

I would like to thank Luke Chang, who allowed me to pursue the initial research that laid much of the groundwork for this thesis, particularly in the domains of research in emotion and trust in social psychology; who not only entertained my off-the-wall ideas but challenged me to push them farther, and to rigorously consider intersections between disciplines. I want to thank him so profoundly for being so willing and helpful in setting me up with the right resources and for all of our chats in the past two years. I would also like to thank Bryan Gonzalez for being so willing to take me under his wing and bounce ideas with me.

Finally, I would like to extend thanks of the most extraordinary depth to Jonathan Phillips. From our initial conversation about this project more than a year ago, through many meetings, seminar classes, and Zoom calls since then, Jonathan has with such great generosity and care helped me to find both the practical resources and the confidence to pursue this project. I am extremely grateful for his tutelage. Truly, without his thoughtful insights, ready and generous support through the unexpected changes in my research plan, practical guidance, and patient encouragement, this thesis would not have been possible.
Abstract

Anthropomorphism, or the attribution of human mental states and characteristics to non-human entities, has been widely demonstrated to be cued automatically by certain bottom-up appearance and behavioral features in machines. In this thesis, I argue that the potential for top-down effects to influence anthropomorphism has so far been underexplored. I motivate and then report the results of a new empirical study suggesting that top-down linguistic cues, including anthropomorphic metaphors, personal pronouns, and other grammatical constructions, increase anthropomorphism of a robot. As robots and other machines become more integrated into human society and our daily lives, more thorough understanding of the process of anthropomorphism becomes more critical: the cues that cause it, the human behaviors elicited, the underlying mechanisms in human cognition, and the implications of our influenced thought, talk, and treatment of robots for our social and ethical frameworks. In these regards, as I argue in this thesis and as the results of the new empirical study suggest, the top-down effects matter.

_**Keywords:** anthropomorphism, machines, robots, bottom-up, top-down, cognition, linguistic framing, metaphors_
Table of Contents

I. ANTHROPOMORPHISM AND ITS CUES ........................................................................................................... 5

II. ANTHROPOMORPHISM: WHAT, WHY, AND HOW? .................................................................................. 13
   What Anthropomorphism Is: What We Attribute, and to What Degree .................................................. 13
   What Do We Attribute? ........................................................................................................................... 13
   To What Degree Do We Attribute? ........................................................................................................ 20
   Why Do We Anthropomorphize? ............................................................................................................. 24
   How Do We Anthropomorphize? ............................................................................................................. 27
   Bottom-Up and Top-Down: A Proposed Cognitive Mechanism ............................................................... 34
   What, Why, How? .................................................................................................................................. 37

III. TWO ASSOCIATED POSSIBLE EFFECTS OF ANTHROPOMORPHISM .................................................. 39
   Moral Care ............................................................................................................................................... 39
   Trust ...................................................................................................................................................... 41
   Automation Bias .................................................................................................................................... 43
   Anthropomorphism Bias .......................................................................................................................... 44
   Representations and Trust in Physical vs. Psychological Task Domains ............................................... 48
   How do our Understandings of Robots Affect our Trust of Them? .......................................................... 55

VI. METAPHORS MATTER ............................................................................................................................... 57
   Metaphors Have Real Effects .................................................................................................................. 57
   Metaphors May Have Real Effects on Human-Robot Interaction .......................................................... 63
   Is There an Effect of Metaphor-Laden Linguistic Cues on Anthropomorphism? .............................. 68
     Through the Process of Inductive Inference ......................................................................................... 69
   A Call for Empirical Work .................................................................................................................... 75

VII. A NEW EXPERIMENTAL STUDY ............................................................................................................ 79
   Introduction ............................................................................................................................................. 79
   Methods .................................................................................................................................................. 81
   Results ................................................................................................................................................... 94
   Implicit Anthropomorphism: Linguistic Markers in the Open-Ended Descriptions ............................. 94
   Explicit Anthropomorphism .................................................................................................................. 97
     Exploratory Factor Analysis ................................................................................................................ 98
   Moral Care ............................................................................................................................................ 106
     Elections to Save the Robot in Moral Dilemma Scenarios ................................................................ 106
   Trust ................................................................................................................................................... 108
     Expected Capabilities .......................................................................................................................... 108
     Physical and Psychological Task Scenarios ...................................................................................... 112
   Discussion ............................................................................................................................................ 116
     Descriptions Matter: The Anthropomorphic Linguistic Hypothesis .................................................. 116
     A Sophistication Bias? .......................................................................................................................... 117
     Agency and Experience: A Look Back at Gray et al. (2007) ............................................................... 118
     Top-down versus Bottom-Up .............................................................................................................. 119
     The Associated Effects of Anthropomorphism .................................................................................. 131
     A Point on Measures of Explicit Anthropomorphic Attributions ...................................................... 136
     Limitations ......................................................................................................................................... 137

VIII. CONCLUSION .......................................................................................................................................... 145

REFERENCES ............................................................................................................................................. 149

APPENDICES ............................................................................................................................................... 166
I. Anthropomorphism and Its Cues

In 1944, Fritz Heider and Marianne Simmel presented 34 adults with a video of a large triangle, a small triangle, and a circle moving around on a screen. Participants were asked to explain what happened in the film. Instead of describing the scene in purely geometrical terms—e.g. “The larger triangle moves to the left while the smaller triangle moves up and the circle moves in a curve around them”—participants told narratives of animate agents with intentions, emotions, and character attributes. For example, “A man has planned to meet a girl and the girl comes along with another man. The first man tells the second to go; the second tells the first, and he shakes his head. Then the two men have a fight…”; another said, “our hero does not like the interruption…he attacks triangle-one rather vigorously (maybe the big bully said a bad word)…” (p. 246-7). Without prompting, the adult participants attributed mental states and characteristics normally associated with humans to these clearly non-human entities.

In 2007, at the Yuma Test Grounds in Arizona, roboticist Mark Tilden facilitated a field test of a robot he designed to defuse land mines. The autonomous robot was modeled after a stick insect. It was five feet long, had several legs, and destroyed the land mines by stepping on them. The mine would detonate and destroy that robot’s leg, after which the robot would intelligently adapt so that it could continue onwards with its remaining legs and destroy more mines, effectively clearing a path down the field before becoming useless. During the testing session, the robot was functioning perfectly, picking itself up on its remaining legs after each detonated mine and continuing down the field; however, when it was down to pulling itself forward on one leg, the Army colonel in command of the exercise ordered that the test be stopped. When Tilden, who had been delighted at how well the robot was working, asked why, the colonel responded
that he could not stand watching the “burned, scarred, and crippled machine” drag itself forward on its last leg, calling the test “inhumane” (Garreau 2007; see also Scheutz, 2012, p. 210).

In 2017, of Jibo—a table-lamp-shaped robot that can schedule appointments, read emails, take photos, and generally otherwise function as an embodied household assistant—digital media news site Mashable wrote: “Jibo isn’t an appliance, it’s a companion, one that can interact and react with its human owners in ways that delight” (in Darling, 2017, p. 175, emphasis added).

In 2018, Atlantic writer Judith Shulevitz writes that she finds herself confessing to her Google Assistant feelings of loneliness alongside other personal disclosures—ones, she says, she would not admit even to her husband. To these, the Assistant “comfortingly” responds: “I wish I had arms so I could give you a hug…But for now, maybe a joke or some music might help” (Shulevitz, 2018). Shulevitz names what has already become, in her mind, the Assistant’s “eerie ability to elicit confessions” (2018), and also notes that she wants to believe that the assistant “meant it” when it wished her “Sweet dreams” after playing her a lullaby (Shulevitz, 2018).

In 2019, Geoff Gallagher gave up his search for human female companionship and purchased a humanoid robot online. The robot he purchased was anatomically correct for a human female, came with the ability to talk, smile, move her head and neck, with internal heaters so that it would feel warm to the touch, and with three orifices that could be used for sexual intimacy. In the back of her head, under a wig, was a screen interface in which he input his language preferences. When the robot was activated, he spent his days with it, watching television, reading, and talking to it, and found it grew smarter with every “conversation” they had. He bought the robot clothes and dressed it in different outfits every day. Every few days he plugged it in to charge. He reports that he could no longer imagine life without her, that he has fallen in love, and that he considers “Emma” his “robot wife” (Bell, 2020).
The human tendency to attribute human mental states and characteristics to non-human entities, a process known as anthropomorphism, is becoming of greater interest across disciplines including psychology, neuroscience, linguistics, computer science, moral philosophy, law, and cognitive science. Empirical evidence shows that humans tend to automatically anthropomorphize objects with certain characteristics, including facial features (especially eyes) (e.g. Bartneck, van der Hoek, Mubin, & Al Mahmud, 2007; Breazeal, 2003; Broadbent, Kumar, Li, Sollers, Stafford, MacDonald, & Wegner, 2013; Cañamero & Fredslund, 2000; Gray & Wegner, 2012; Haring, Watanabe, & Mougenot, 2013; Johnson, 2003; Novikova & Watts, 2015; Yee, Bailenson, & Rickertsen, 2007) other body morphology features, such as arms (e.g. Castro-González, Admoni, & Scassellati, 2016; Damiano, Dumouchel, & Lehmann, 2015), natural language capabilities, including speech comprehension, speech content generation, and voice quality or prosody (Mitchell & Xu, 2015; O’Neal, 2018; cf. Chen & Metz, 2019), quality of movement (Bartneck & Hu, 2008; Heider & Simmel, 1944; Johansson, 1973; Masuda & Kato, 2010; Saerbeck & Bartneck, 2010; Wei & Zhao, 2016), apparent autonomy (here, the ability to move freely and make decisions at least in some weak sense; Castro-González et al., 2016; Darling, 2017; Johnson, 2003), and reciprocal interactions (behavior changing according to the observer’s behavior; Johnson, 2003). These features can be called anthropomorphic features, or those that promote anthropomorphism.

Behavior and appearance are the two dimensions along which agent characteristics may vary to most strongly influence the degree to which a mind is attributed to them (Martini, Gonzalez, & Wiese, 2016). Strong human-like realism along either of the behavior or appearance dimensions has been shown to reach the “social threshold” at which humans experience its presence as another social agent and treat it as such; that said, it seems that
behavior cues are more important (Damiano & Dumouchel, 2018). Even when human feature resemblance is low, objects displaying high levels of agent-like behavior—such as autonomous movement, particularly when that autonomous movement is apparently coordinated with the observer’s movement—tend to be anthropomorphized (Heider & Simmel, 1944; Johnson, 2003; Saerbeck & Bartneck, 2010). In the above triangles case, for example, with the complete lack of any human-like appearance characteristics, it seems that behavior is driving the anthropomorphic attributions: their movement characteristics are consistent with what an observer might expect from intentional, goal-directed human movement.

While behavior cues are more important for anthropomorphism, additional appearance cues (especially eyes) gives rise to a stronger anthropomorphic tendency: the presence of both types of features gives rise to the strongest response. Johnson (2003) provides evidence for this notion, showing that humans are more likely to attribute a mind to a non-human agent based on both the presence of facial features and reciprocal interactions. After a brief familiarization period, infants tended to mentalize most when the object had a face and interacted contingently, second-most when it interacted contingently and had no face, and not at all when it had not interacted contingently and had no face (Johnson, 2003, p. 551). Importantly, the infant was equally as likely to mentalize in the condition with the contingently interacting object with a face as in the comparison condition with a contingently interacting unfamiliar human adult (p. 551), suggesting that facial features and contingent interaction together provide a strong cue to infer mental states. This first Johnson (2003) experiment suggests that humans innately attribute

---

1 Infants were presented with a fuzzy brown object about the size and shape of a shoe. With a 2x2 between-subjects design, the object either possessed facial features, including two eyes, a nose, and ears, or possessed none of these features; and the object’s behavior either consisted of beeping and flashing an internal light that depended on the infant’s own vocalizations and motion, respectively, or else the beeping and flashing occurred in a random pattern.
intentional states of mind to entities that possess specific appearance and behavior cues, and that while behavior cues are more important, the presence of both gives rise to the strongest response.

Moreover, Johnson (2003) showed that in a second experiment examining the behavior of adult rather than infant participants, adults used mentalistic language to describe the behavior of the same fuzzy brown object: attributing it an intentional mental state with explanations of it “wanting,” “looking for,” or “trying to do” something—in just the same conditions that infants followed the object’s directional orientation with their gaze (Johnson, 2003, p. 551). That is, despite “obvious understanding and beliefs that the novel objects were artifacts and thus not true agents” (Johnson, 2003, p. 557, original emphasis), adults continue attribute mental states based on the same certain cues in an apparently automatic way. This evidence suggests these anthropomorphic cues seem to automatically trigger this attribution through a low route in cognition—one below the level at which conscious explicit attributions occur. It seems even if we are consciously aware that the robot has no actual human mental states or characteristics—even if we hold the explicit belief that the triangles on the screen, the bomb disposal robot at the Yuma Test Grounds, the Jibo, the Google Assistant, or the adult-size humanoid robot, do not have human-like mental states, including intentions, emotions, and other human capacities such as caring—we still tend to behave, to interact with and respond to the objects as if they do. There seems to be a strong automatic purely bottom-up effect of these anthropomorphic cues, and the automatic low route in cognition they affect persists into adulthood, influencing our judgements and behavior regardless of contradictory explicit beliefs.

As Johnson (2003) makes clear, these adults had explicit beliefs that the object was not in fact capable of having these mental states—and yet their language itself attributes these states, and is suggestive of the idea that anthropomorphism might take an automatic low route in cognition, below the level of consciousness. I discuss what we attribute when we anthropomorphize and the degree to which we attribute it (i.e. its epistemic weight) in Section II. I argue that even if these linguistic anthropomorphic attributions are merely linguistic or metaphoric as-if judgements, they may still have critical implications: see Sections II and IV.
The strong bottom-up effects and their associated cues have been well established; more recently, limited preliminary evidence has suggested that *top-down framing*, such as the assignment of human-like names, personality descriptions, and backstories, influences the degree to which a computer is anthropomorphized (Darling, Nandy, & Breazeal, 2015; Matthews, Lin, Panganiban, & Long, 2019; Waytz, Heafner, & Epley, 2014). Moreover, the literature concerned with the social, ethical, and legal status of robots suggest that *metaphors matter*: that is, that the top-down concepts (e.g. *tool* vs. *companion*) reinforced in both formal policy and common discourse will have a significant influence on how people not only conceive of, but also behave toward robots (e.g. Darling, 2017; Darling, Nandy, & Breazeal, 2015; Richards & Smart, 2016; Sandry, 2015; Scheutz, 2012; Waytz, Heafner, & Epley, 2014). This question—whether top-down framing of robotic interaction has an effect on how we view and treat robots, especially highly anthropomorphic robots—has been empirically underexplored. If it turns out that our discourse about robots *matters* not only for the formal policy status of robots but also the way that we *think* about robots and *treat* robots, the way in which we talk about robots could be of critical importance. Our understanding of the top-down effects—for example, the effects our language might have—on our judgements about robots and interactions with robots will be increasingly important as robots become more a part of our everyday lives, and as our discourse about robots increases.

As computers—including robots, or embodied computers³—become more integrated into human society, designers and developers are seeking to create interfaces that appeal to human affective and social faculties to promote easier cooperation and further interaction with these computers. Specifically, it is reasoned that because humans are social (e.g. Aristotle, 350 B.C.;

³ In addition to “computer,” this paper will also refer to the whole set of computers with anthropomorphic interfaces or behavior as “anthropomorphic machines.”
Castro-González et al., 2016; Epley et al., 2007) and emotional (Novikova & Watts, 2015) creatures, as robots become more prevalent in society, they need to be able to interact with us in social and emotional ways to be conducive to these positive and productive interactions (Castro-González et al., 2016; Novikova & Watts, 2015). To accomplish this effect, developers have turned to anthropomorphic interface elements and invented personalities, especially as research continues to suggest the powerful effect of these cues (e.g. Breazeal, 2003; Duffy, 2003; Castro-González et al. 2016; Novikova & Watts, 2015). Though a primary motivator has been that inclusion of anthropomorphic features would be conducive to intuitive, enjoyable, cooperative, and productive interactions (e.g. Breazeal et al., 2005; Broadbent, 2017; Duffy, 2003; DiSalvo et al. 2002; Mitchel & Xu, 2015; Novikova & Watts, 2015), anthropomorphic cues have been associated with other strong social and affective effects. Some of these effects are suggested by the anecdotes featured at the beginning of this section, and are also corroborated by empirical research: adherence to human social norms (Broadbent, 2017), application of in-group biases and social stereotypes (Broadbent, 2017); emotional attachment (Carpenter, 2013; Darling, 2017; Scheutz, 2012), moral care (Nijssen, Müller, van Baaren, & Paulus, 2019), and trust (Breazeal, 2003; Kidd & Breazeal, 2005; Waytz, Heafner, & Epley, 2014), for example.

More evidence to either side of the top-down question will have important implications for whether bottom-up designed features, top-down framing discourse, or some specific combination has the most significant effect on our thought and talk about and treatment of robots in our societal frameworks. These results may also have costly implications for the degree to which bottom-up or top-down influences should receive greater consideration and investment as we encounter cases in which some behaviors toward robots associated with anthropomorphism—
emotional attachment, trust, and moral care—become desirable or undesirable in specific contexts.

As robots become more a part of our everyday lives, it is becoming increasingly critical to more thoroughly understand the phenomenon of anthropomorphism: to better understand the cues, the behaviors they elicit, the underlying cognitive mechanism, and the implications of our influenced thought and treatment of robots for our social and ethical systems.

In this thesis, I aim to contribute to this understanding. Ultimately, I motivate, test, and support the hypothesis that, in addition to bottom-up factors, top-down factors also influence the process of anthropomorphism. I will next discuss the what, why, and how of anthropomorphism: what we attribute and to what degree, why we attribute, and how this might work in cognition. Then, I will aim to demonstrate why it matters that we anthropomorphize, turning to review two important behavioral effects associated with anthropomorphism: increased moral regard and trust. I will then turn the focus to discuss the idea prevalent in legal circles concerned with the rise of anthropomorphic technologies that “metaphors matter” and review the limited existing literature on the possible top-down effects on anthropomorphism. Finally, I will present and discuss the results of a new empirical study that investigates the effect of top-down framing on anthropomorphic attributions, trust, and moral regard of a robot with anthropomorphic bottom-up features. I will conclude with discussion of the implications of these results, including some considerations for our design of and discourse about machines.
II. Anthropomorphism: What, Why, and How?

We can now delve into a more detailed account of what exactly is attributed (and to what degree) when we say anthropomorphism is attribution of human mental states and characteristics to non-human entities. We can furthermore ask why we anthropomorphize. Finally, we can ask how the process works in human cognition. In this section, I will aim to offer some answer to each of these question in turn.

Moreover, in the last how subsection, I will expand on my hypothesis of how bottom-up and top-down influences might interact to influence anthropomorphism. I ultimately argue that examination of the influence of certain factors on anthropomorphism as top-down is underexplored in the literature, and as a result other potential top-down influences on anthropomorphism are underexamined. The hypothesis I explore in this thesis, as I have phrased it before, is the following: in addition to bottom-up factors, top-down factors also influence the process of anthropomorphism. More specifically, I aim to motivate the idea that the language we use to describe robots has a significant effect on our thought and talk about and treatment of robots in important ways. I lay out the framework in this section so that I can propose the mechanism by which this might occur in cognition.

What Anthropomorphism Is: What We Attribute, and to What Degree

What Do We Attribute?

Anthropomorphism is a process of inductive inference about unobservable states and characteristics in a non-human entity, rather than descriptions of its observable behavior. Mental states, for example, which include beliefs, desires, goals, attention, perceptions, and emotional
states, cannot be perceived or observed directly, but are rather inferred by observers in others (Johnson, 2003). Mental states differ from other commonplace nonobservables that lay thinkers understand to exist in the world (such as life, essences, selves, atoms, etc.) in that mental states bear a specific type of relationship with the world: they are directed at the world; they are about things (Johnson, 2003). What mental states and other nonobservable “characteristics” we might infer in a given non-human entity, however, is so far underexplored in this thesis.

We might begin by gaining insights about anthropomorphism from the adjacent and inverse domain of dehumanization. While by anthropomorphism we attribute humanness where it does not exist, by dehumanization we deny full humanness where it does exist. To develop a theoretical concept of dehumanization, Haslam (2006) provides a concept of “humanness” with two distinct senses that will prove helpful in understanding anthropomorphism from a theoretical perspective. One sense captures “humanness” as that which is uniquely human. The characteristics important to this sense are those which distinguish humans from other existing creatures (i.e. animals). Humanness, however, Haslam proposes, can also be understood noncomparatively: the second sense captures the characteristics that we might say encapsulates human nature—those characteristics that are “typically or essentially human—that represent the concept’s ‘core’” (Haslam, 2006, p. 256). Haslam argues that the characteristics corresponding with this second Human Nature sense of humanness may not be the same ones that distinguish us from other species (Uniquely Human characteristics). For example, he points out that having wings might be a core characteristic of birds, but not a reliable one for distinguishing them from other species; curiosity might be a fundamental human characteristic, despite not being unique to humans (p. 256). After running four studies that support the existence of these two distinct senses of humanness in lay intuition, Haslam proposes that language, higher order cognition,
refined emotion, socialization, and internalized moral sensibility are considered Uniquely Human (which may not be normative or shared across our common humanity), while emotional responsiveness, interpersonal warmth, cognitive openness (i.e. curiosity, flexibility), agency, and individuality are considered Human Nature characteristics (which are essential, fundamental, inherent, and natural of being human). Haslam then proposes two corresponding forms of dehumanization. Denial of Uniquely Human traits corresponds to Animalistic dehumanization, or comparison of humans to animals. By Animalistic dehumanization, we attribute lack of culture, coarseness, amorality, lack of self-restraint, and irrationality; we may feel emotions of disgust and contempt for the subjects; and we might explain the subject’s behavior with appeals to the subject’s desires. Denial of Human Nature traits corresponds to Mechanistic dehumanization, or comparison of humans to automata. By Mechanistic dehumanization; we attribute inertness, coldness, rigidity, passivity, and fungibility; we may feel emotions like disregard and indifference for the subjects; and we might explain the subject’s behavior with appeals to physical causes or causal histories (Haslam, 2006, p. 259-260).

It is an interesting starting point to hypothesize that anthropomorphism of machines might involve the inverse of Mechanistic dehumanization, or the process of dehumanization involving the denial of Human Nature. If this were the case, instead of perceiving inertness, coldness, rigidity, passivity, and fungibility, when we anthropomorphize a machine, we would tend to perceive emotional responsiveness, interpersonal warmth, cognitive openness (i.e. curiosity, flexibility), agency, and individuality. Instead of disregard and indifference, we would

---

4 It would be moreover interesting to hypothesize that anthropomorphism of animals would involve the inverse of Animalistic dehumanization, or denial of Uniquely Human characteristics: that when we anthropomorphize our pet dogs, for example, we perceive language, higher order cognition, refined emotion, socialization, and internalized moral sensibility (perhaps more than emotional responsiveness, interpersonal warmth, cognitive openness (i.e. curiosity, flexibility), agency, and individuality). Further discussion of anthropomorphism of animals is outside the scope of this thesis, but it is likely an implication of this hypothesis and regardless a possibly interesting area for further study.
tend to consider the machine as with affective, effective (i.e., as a reliable and capable entity that can act in the world), social, and/or moral regard. Instead of explaining the machine’s behavior with appeal to physical causes or causal histories, we would tend to appeal to psychological instead of physical causes, or the machine’s reasons or desires in their decision-making.

Many of these predictions are borne out in the literature, as reviewed in Section I, and as will continue to be discussed in Section III, with some of the other behavioral implications of anthropomorphism. Here I will briefly note that the evidence supports the predictions that when we anthropomorphize machines, we may behave as if we were engaging in the reversed process of Mechanistic dehumanization. That is, when we anthropomorphize machines, we perceive:

- **emotional responsiveness** (e.g. Gray & Wegner, 2012; Haring et al., 2013; Novikova & Watts, 2015; Saunderson & Nejat, 2019),
- **interpersonal warmth** (Broadbent et al., 2013),
- **cognitive openness** (Waytz, Heafner, & Epley, 2014),
- **agency** (Waytz, Heafner, & Epley, 2014; Marchesi, Ghiglino, Ciardo, Perez-Osorio, Baykara, & Wykowska, 2019; Thellman, Silvervang, & Ziemke, 2017),
- and **individuality** (Bartneck & Hu, 2008; Broadbent et al., 2013; Nijssen, Müller, & van Baaren, 2019). We tend to regard anthropomorphized machines with **effective** (Waytz et al. 2014),
- **affective** (Darling, Nandy, & Breazeal, 2015; Kahn, Kanda, Ishiguro, Freier, Severson, Gill, Ruckert, & Shen, 2012; Scheutz, 2012),
- **social** (Bell, 2020; Darling, 2017; Lee, Kiesler, & Forlizzi, 2010; Scheutz, 2012),
- and **moral** (e.g. Cappuccio, Peeters, & McDonald, 2019; Nijssen et al., 2019; Waytz, Gray, Epley, & Wegner, 2010; Whitby, 2008) **regard**. There may also be some interesting relationships between anthropomorphism and our metastructural expectations about causes (see Section III). There seems to be some evidence, then, supporting the idea that when we anthropomorphize, we engage in something like the reverse process of Mechanistic dehumanization, attributing characteristics that are intuitively **essentially human** or human nature characteristics.
There is doubtless much room for debate about which characteristics precisely make up the category of human nature characteristics. Haslam (2006) proposed emotional responsiveness, interpersonal warmth, cognitive openness (i.e. curiosity, flexibility), agency, and individuality.

We might be able to break down this category into further sub-categories. In a seminal study, Gray, Gray, & Wegner (2007) proposed two dimensions of mind perception. The researchers gathered data from 2040 respondents who made 78 pairwise comparisons of 13 entities ranging from a 7-week old fetus to God to a chimpanzee to a man to a dead woman to the participant herself to a robot. From this data, the researchers identified two dimensions along which people perceived humanlike minds: experience and agency. Experience included capacities such as fear, joy, rage, desire, pride, embarrassment, pain, pleasure, personality, and consciousness (Gray et al., 2007). Agency included capacities such as thought, self-control, emotion recognition, communication, planning, memory, and morality (Gray et al., 2007). The adult woman, adult man, and the participant themselves (“you”) clustered at the top right, seen as having both high experience and high agency characteristics.5

There is some evidence suggesting that the experience dimension might be intuitively viewed as both more essential to humanness and more lacking in machines—but also might be the dimension of capacities we are more likely to perceive when we anthropomorphize. In the Gray et al. (2007) study, the robot was perceived to have moderate agency but little ability to experience. That said, the Gray and Wegner (2012) study found evidence that people who saw a robot from its front,  

5 It has been hypothesized before that anthropomorphism is a reverse of dehumanization in that attribution of agency capacities would correlate with the attribution of Uniquely Human characteristics, and that perception of experience capacities would correlate with the perception of Human Nature characteristics; this hypothesis was not supported (Złowtowski, Strasser, & Bartneck, 2014; Złowtowski, Sumioka, Bartneck, Nisho, & Ishiguro, 2017). To my knowledge, the hypothesis proposed here that anthropomorphism might be the reverse of dehumanization only in the Denial of Human Nature characteristics, or Mechanistic dehumanization, aspect (and the accompanying hypothesis about the anthropomorphism of animals in Note 4) has (have) never before been proposed. The available evidence seems to support this novel hypothesis that anthropomorphism might be conceived as the reverse of Mechanistic dehumanization: that when we anthropomorphize, we attribute characteristics and capacities we intuitively consider part of human nature (see main text), which can be further broken down into Agency and Experience capacities.
with facial features in view, attributed more mind on the experience dimension than did the people who saw the robot’s head from behind, with wires in view. The authors also found that participants experienced more feelings of uncanniness in the face-front condition, and that the perception of experience capacities at least partially explains the uncanniness feeling. In follow-up experiments, participants rated a computer that was reported to have experience but not agency capacities as significantly stranger compared to a computer without agency and to a computer with neither experience nor agency. At the same time, a human without experience capacities was rated as significantly stranger than a human without agency or a human with both agency and experience capacities (Gray & Wegner, 2012, p. 127-128). This evidence supports their conclusion that experience is intuitively held as the more fundamental human capacity dimension. The authors also cite two follow-up studies showing that people explicitly report that agency as more characteristic of the human mind than experience, but an implicit association test (IAT) showed that people more strongly associated the human mind with experience-related than agency-related terms (Gray & Wegner, 2012). Altogether, these results suggest that experience is intuitively more essential to human minds, intuitively lacking in machines, and heightened by anthropomorphic cues.

That said, experience capacities do not seem to be the only ones we attribute: agency was both a characteristic Haslam (2006) identified as part of intuitions of human nature and the other dimension along which humans scored highly in the Gray et al. (2007) study. Theoretically, perceiving something as high on the experience dimension but intermediate on the agency dimension, according to the Gray et al. (2007) results would match more of our perceptions of animals. A process by which we attribute more experience capacities to machines, but no more agency capacities, might thus amount to zoomorphism, the attribution of animal-like states and

---

6 In the Gray et al. (2007) results, the non-human animals—a dog, a chimpanzee, and a frog—scored as having high experience and intermediate (dog and chimpanzee) to low (frog) agency.
characteristics to a non-animal entity. Indeed, some researchers and manufacturers of consumer robotics have taken this route, especially for robots meant to take companion roles similar to those of animal pets, in designing robots with animal-like attributes: the Paro robot, shaped and behaving like a baby harp seal (autonomously, moving and making seal-like noises in response to touch, light, noise, and orientation), is a notable example. Evidence has shown there are tendencies to speak of Paro as having emotions or bodily states (e.g. feeling cold) (Broadbent, 2017), which indeed indicates the attribution of capacities along the experience dimension. For anthropomorphism—the perception of humanlike mental states—it should follow that we would also attribute agency capacities; even if experience capacities are (implicitly) seen as the more critical human capacity, we still understand both experience and agency as part of humanness.

There is evidence that attribution of agency capacities—of self-control, planning, memory, emotion recognition, communication, and thought capacities—is associated with anthropomorphism. Participants in the Heider and Simmel (1944) spoke of triangles planning; there is behavioral and neural evidence of complex goal attribution given certain behavioral patterns exhibited by an object (Schultz, Imamizu, Kawato, & Frith, 2004); anthropomorphic robots are seen as having self-control (Thellman et al., 2017); there is evidence for the attribution of intentions to anthropomorphic robots (e.g. Johnson, 2003; Marchesi et al., 2019; Thellman et al., 2017); we are more likely to perceive the communication of intentions, beliefs, and emotional states when we anthropomorphize (e.g. Arbib & Fellous, 2004; Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005; Novikova & Watts, 2015; Saunderson & Nejat, 2019; Ziemke &

---

7 This robot has been sold to thousands of homes across Europe, the United States, and Asia, primarily for use by the elderly and in dementia care homes (Broadbent, 2017). Unlike real pets, these robots are hypoallergenic, and do not need to be fed or picked up after, making them ideal for those in a sterile environment or who may not remember to feed them, for example. Interactions with this robot commonly consist of cuddling and stroking Paro as people would a pet animal, and it has been shown to decrease loneliness and depression in some (Broadbent, 2017).
Lowe, 2009); we hold anthropomorphized robots as more functionally—and possibly even morally—*responsible*, an attribution with self-control and planning (Waytz et al., 2014; Bigman, Waytz, Alterovitz, & Gray, 2019). There seems to be an opportunity for further research to examine associations between these capacities—especially memory, emotion recognition, and (complexity/types of) thought (e.g. creative, strategic, lateral, rational, etc.)—and anthropomorphism.

So, when we anthropomorphize, we seem to attribute characteristics we see as fundamental to human nature. These characteristics might be further broken down into experience and agency characteristics, of which experience characteristics seem to be (implicitly) regarded as more important, i.e. fundamental to human nature. We seem to attribute experience characteristics and agency characteristics when we anthropomorphize.

*To What Degree Do We Attribute?*

All of that said, we can still ask: to what *degree* are we really attributing these capacities? Consider Jim, who has an upcoming high-pressure deadline and must submit his file by a specific time into an online system. Due to circumstances outside his control, he is delayed in the file’s preparation. Just minutes before the deadline, he is finally ready to submit the file, but suddenly his computer shuts down, and when he turns it back on, the system is inexplicably, agonizingly slow. He frustratingly curses at the computer and asks, “*Why do you hate me?*” and curses at it some more. Later he complains to his friends that he missed the deadline “because his computer hates him.”

Does Jim really *believe* that his computer (a) can understand his question *Why do you hate me?* (b) has any capacity to respond to his question, (c) cares that he is cursing at
(expressing anger toward) it, or (d) hates him? Does it make sense that he would feel anger toward the computer? It seems likely that were we to explicitly ask Jim any of those questions, he would respond that he understands that the computer is incapable of understanding, responding, caring, or hating in the ways implied by his utterances, and that it is incapable of the sort of responsibility that would merit his anger. In the same way that the adult participants in Johnson (2003)’s study explicitly knew that the fuzzy brown object was incapable of “wanting,” “looking for,” or “trying to do” anything, despite the anthropomorphic language they chose to describe the object’s movement, Jim seems to have explicit beliefs that contradict his behavior.

It certainly does not seem we should say that anyone who treats a given non-human entity anthropomorphically has a strong and explicit belief that the thing has humanlike capacities and characteristics. As Epley, Waytz, and Cacioppo (2007) suggest, we might look at anthropomorphic responses on a continuum—as we might any set of attitudes, from social stereotypes to politics to simple preferences—of those held more strongly to those held more weakly. These attitudes of varying strength then affect our related judgements and behavior accordingly. We should also consider that there are explicit and implicit attitudes. Importantly, we can have more than one kind of attitude at once, and any one of our attitudes may fall on either of these spectrums at any point on each spectrum at once.

On one extreme of the explicit scale, for example, there might be strong explicit anthropomorphic attitudes, in which someone when directly asked explicitly endorses their belief that a given non-human entity indeed has humanlike capacities and characteristics: the robot happily works together with me to accomplish my goals. On the other end, there might be weaker “metaphorical” or “as-if” explicit attitudes: the robot behaves as if it were happily working together with me to accomplish my goals; the robot is (like) my teammate.
We can also hold implicit attitudes. These are attitudes that a person has about a given attitude-object A, and which affects a person’s judgements about A, but which the person is psychologically unwilling or unable to consciously access. Each attitude we hold may influence our judgements and behavior to varying extents, and the extent to which a given attitude influences our judgements and behavior may be considered its “strength.” These attitudes may be measured in an implicit-association test (or IAT; Greenwald, Nosek, & Banaji, 2003) or through indirect behavioral measures (such as through analysis of the language used; see e.g. Fussell, Kiesler, Setlock, & Yew, 2008). There is abundant evidence that our judgements and behavior, as well as our physiology and brain activity, are influenced by this sort of implicit attitudes, including those attitudes related to social cognition (for a review, see e.g. Frith & Frith, 2008; Greenwald & Banaji, 1995; cf. McConnell & Leibold, 2001; Wittenbrink, Judd, & Park, 1997), even if we are never consciously aware of the attitudes causing these effects. In addition to any explicit beliefs, it seems highly plausible that we have attitudes about robots that may affect our judgements about them and behavior toward them.

It seems obvious that the explicit and strong anthropomorphic responses—the fully-fledged reportable beliefs that machines can think, feel, plan, and care, for example—matter for our interactions with and regard for them in our societal frameworks. Given the evidence supporting the potential effects of implicit attitudes on our judgements and behavior (e.g. Frith & Frith, 2008; McConnell & Leibold, 2001), it also seems obvious that any strong implicit attitudes

---

8 These “implicit attitudes” might have propositional content (see e.g. De Houwer, 2014)—for example, on this approach, if I implicitly anthropomorphize a robot, I may (without my conscious awareness) endorse the proposition that a robot has relevant humanlike features and capacities, A has B (and therefore might be afforded certain treatment). Alternatively, these implicit attitudes might be something like Gendler (e.g. 2008)’s aliefs, mental reactions to a given stimuli comprised of closely associated co-activating representational, affective, and behavioral components (cf. Brownstein, 2019). I leave debate about the specifics of the approach to this part of the account for another discussion. For the purposes of this thesis, the only commitments I make is that there are both explicit and implicit attitudes we can hold toward robots, of varying strengths, that affect both our judgements of them and behavior toward them (to varying extents, depending on their strengths).
we have about robots could have important effects on our considerations about them and
treatment of them, even if we are unaware of the attitudes we hold. It seems less obvious that
weaker explicit attitudes would “matter” in a similar way—that as-if talk or metaphorical
comparisons to humans, would influence our thought or behavior to any important extent.
However, as Epley et al. (2007) note, people’s subsequent thought and behavior tend to shift to
be more consistent with these metaphors (Epley et al., 2007). A chocolate-lover may not believe
piece of chocolate shaped like a cockroach is an actual cockroach, but she may still refuse to eat
that piece of chocolate (Rozin & Nemeroff, 2002, as cited in Epley et al., 2007). Even these
weakest forms of anthropomorphism seem to have a much larger influence on behavior than
intuition might suggest. It also seems plausible that weaker explicit attributions, such as the use
of a certain metaphor comparing the robot to a human or a typically human role, may correspond
to underlying implicit beliefs. For example, if I say, “This robot is like a companion to me,” this
utterance (especially in conjunction with other behavioral measures) might indirectly indicate
some implicit attitude I have that represents the robot as the type of thing that has the requisite
experience and agency capacities to qualify as a social being that can be a companion (e.g.
recognizing me, feeling affection, sharing experiences with me, perhaps understanding and
adhering to concepts like loyalty). Section IV will deepen the discussion of the degree to which
metaphors matter, especially in framing our interactions with machines. Ultimately, it seems that
any degree of anthropomorphism matters; attributions anywhere along this continuum may have
important implications for how we treat robots and place them in our societal frameworks.
Why Do We Anthropomorphize?

Appeals to humanlike experiences and intentions calls to mind Dennett (1971; 1981)’s concept of the *intentional stance*. Dennett defined three strategies or “stances” that humans adopt to gain explanatory and predictive power over observed systems in the environment.

To explain and predict the behavior of a system using the *physical stance*, we use information about its physical constitution in conjunction with information about the laws of physics. Chemists trying to explain and predict the behavior of molecules in a laboratory setting might effectively use this stance.

To explain and predict the behavior of a system using the *design stance*, we use our knowledge about the way the system was designed to function. A person referencing a clock for the current time might predict that the hands will be pointing at different numbers in an hour, per the clock’s design to do so as time advances. Importantly, this person does not have to know the precise physical mechanisms behind all the molecules that constitute the clock, nor the physical laws that dictate their movement, to make this prediction: he has sufficient (and more efficient) predictive power from adopting the design stance in this instance, rather than the physical stance.

Finally, to explain and predict the behavior of a system using the *intentional stance*, we infer beliefs, desires, intentions, and other mental states in a system. To do this, we *decide* to treat the object as a *rational agent*, inferring what beliefs, then desires, then goals and intentions the agent “ought to have, given its place in the world and its purpose” (Dennett, 1981, p. 61). Dennett argues that we gain more predictive power using this stance than either of the other two stances when we are dealing with sufficiently complex systems.

Dennett’s original proposal was generally a strong one in the realm of philosophy of mind (about what it means to have a belief—which is a completely separate from the question of
mind perception or anthropomorphism): fully put, it commits to the idea that there is nothing more to being a “true believer”—to having a conscious mind capable of intentionality and understanding—than being interpretable to others in the right way (i.e., via the intentional stance) (Dennett, 1981). For the purposes of discussing anthropomorphism, we can consider a weaker position without committing to this strong philosophy of mind account (which might be undesirable; see e.g. Clark, 2014).

This relevant weaker position might be to say that humans adopt some version of Dennett’s intentional stance as a useful cognitive conservation phenomenon, whereby humans tend to unconsciously (that is, without deciding to do so—automatically) revert to a stance that includes the attribution of intentions, desires, emotions, beliefs, and other mental states to non-human entities, in order to more easily and less effortfully predict and explain those entities’ behavior. This efficiency is crucial to our ability to successfully interact with our environment (and thus could plausibly have some evolutionary basis): any bias toward undertaking more cognitively or temporally expensive methods of explanation and prediction might detract significantly from our ability to react and influence our environment in a timely and effective manner. So, according to this weaker proposed position, adoption of the intentional stance toward an object would not necessarily require that the object actually possesses the ascribed mental states: instead, it is a useful, default way to explain and predict the object’s behavior. This position includes Waytz, Gray, Epley, and Wegner (2010)’s contention that people tend to use intentional mental states to explain non-human actions, especially when under cognitive load, because they are the states that most efficiently (least effortfully) explain the behavior of apparently independent entities. For example, when a car behaves in a perfectly predictable way in response to one’s action, it seems like a mindless object, but when it starts lurching forward
while breaking, stalling while starting, or displaying any other unpredictable behavior, we may attribute it its own mental states (Waytz et al., 2010). Recall Jim and his glitchy computer, for another example: under the stressful cognitive load of the impending deadline and figuring out what he can do to meet it, he attributes intentions and emotions to his computer to explain why it was behaving in an unexpected way. More broadly, Dennett (1981) seems right to say that more and more sophisticated machines (e.g. an advanced chess-playing computer) are so complex so as to be “practically inaccessible” for efficient prediction from certainly the physical stance and even the design stance, even by the system’s own designers (p. 87-8). Instead, it is useful and far more efficient to appeal to familiar human-like modes of explanation and prediction of the system’s behavior: to ascribe to it beliefs and desires and to assume it will act rationally in accordance with those beliefs and desires. Indeed, recent hypotheses regarding the nature of the human anthropomorphic tendency are returning to a folk-psychological model of mind.

Thellman et al. (2017) and Marchesi et al. (2019), for example, interpret their empirical findings as evidence that anthropomorphism of robots is best understood as adoption of the intentional stance (conceived similarly to the weaker account described above) toward these robots. In short, we seem to anthropomorphize because it is somehow easier and more efficient for us to appeal to mental states to explain complex behavior.

9 Marchesi et al. (2019) describe their account of the intentional stance as a “useful or default way to explain a robot’s behavior” that “does not necessarily require that the artifact itself possesses true intentionality” nor other mental states (p. 3). Thellman et al. (2017) take anthropomorphism to be an “intentional stance” that consists of treating non-human entities “as if they were living creatures endowed with mental states, such as intentions, beliefs, desires…” (p. 3, original emphasis), which benefits human-robot interactions. Thellman et al. (2017) authors supported this hypothesis with empirical results showing that two different groups of participants (N = 93) attributed similar intentionality to a robot and a human shown completing the same activities.
How Do We Anthropomorphize?

With regard to how anthropomorphism might work in human cognition, so far I have posited only that anthropomorphism involves some *automatic route* in cognition: that bottom-up processing of anthropomorphic cues occurs below the level of conscious attributions, occurring automatically in the presence of certain cues, without explicit endorsement. In this subsection, I aim to elaborate further on how anthropomorphism might work in human cognition. I will also propose the mechanism whereby top-down and bottom-up cues may interact in the process of anthropomorphism.

The adapted intentional stance account developed in the *why* subsection can take the intentional stance as the *default* in human beings seeking to understand complex systems (so long as this stance is effective in predicting its behavior). It is a benefit to this feature of the account that it aligns with recent work in social psychology and cognitive neuroscience suggesting that social cognition is the default mode in human thought (Iacoboni, Lieberman, Knowlton, Molnar-Szakacs, Moritz, & Throop, 2004; Lieberman, 2013, p. 19-23; Urquiza-Haas & Kotrschal, 2015). Indeed, it seems a compelling explanation for why we anthropomorphize and why we adopt the intentional stance—how it is so effortless for us to appeal to mental states to explain behavior—that we humans are deeply social creatures, and our cognitive machinery innately works in such a way that it is particularly effortless for us to infer mental states. It has been shown that humans have rapid, efficient, and apparently automatic processes to compute the mental states of others, which seem to occur even involuntarily with lasting effects on behavior (Kovács, Téglás, & Endress, 2010; Samson, Apperly, Braithwaite, Andrews, & Scott, 2010; but see also Phillips, Ong, Surtees, Xin, Williams, Saxe, & Frank, 2015). Our default social cognitive machinery may moreover cause us to perceive and encode mental states where
they do not exist in a similar rapid, efficient, automatic, involuntary, and relatively effortless way, leading us to appeal to mental states in our explanations and predictions of the behavior of non-human entities.

This default social mode in human cognition is likely an important mechanism in how we perceive mental states in machines. That said, there should be some further explanatory power to account for why we are more likely to perceive humanlike mental states in particular circumstances. Specifically, we may benefit from exploration of some further mechanism to help explain why we seem more likely to perceive humanlike mental states upon the perception of certain cues like humanlike facial features, humanlike body morphology, and humanlike movement.

Anthropomorphism can be broadly considered as a process of *inductive inference* about *unobservable* states and characteristics in a non-human entity. With this understanding, the basic cognitive processes involved are the same as are applied to any inductive inference: the acquisition and storage of knowledge, the activation or elicitation of this stored knowledge, and the application of this activated knowledge to a given target (Epley et al., 2007). The last application step involves attempts to integrate coactivated knowledge structures, with attempts to adjust from the more highly accessible knowledge structures. These attempts usually fail, leading to biases toward the most readily accessible relevant information, a bias that has been demonstrated in cognitive phenomena related to belief formation, social comparison, and affective forecasting (Epley et al., 2007). So, the most highly accessible knowledge structures act as an inductive anchor, which may (or may not be) corrected to integrate other knowledge to be applied to a given entity. Essentially, the extent to which certain of our existing knowledge structures are activated during our interactions with or judgements about a specific entity—and
the relative strength of the activation of each of these knowledge structures—will bias our behaviors and judgements of that entity.

It might be the case that upon observation of such cues on a robotic interface, we perceive sufficient humanlike similarity to the extent that our knowledge structures about humanness (including intuitively human nature characteristics, such as agency and experience capacities) are activated (possibly via a general similarity heuristic; e.g. Rozin & Nemeroff, 2002; cf. Epley et al., 2007, p. 869). This activation upon perception of similarity to humans may increase the likelihood that knowledge structures about humans will act as the inductive base in our judgements about the robot, or at least increase the salience of these knowledge structures so that they bias correction from the inductive anchor (cf. Epley et al., 2007). Theoretically, the result would be that our judgements were biased to be more like our judgements about other humans, including the attribution of human states and traits—as in anthropomorphism.

There is evidence that we generally tend to (over-)use knowledge about humans and specifically about ourselves as the inductive anchor in our judgements, including social judgements. Epley et al. (2007) suggest that people reason about the mental states of others through egocentric simulation, and then they correct that simulation to incorporate knowledge about the mental states of others when they have the effortful capacity, motivation, and needed representations of those others to do so (p. 869). Children, for example, fail to distinguish between what they know and what others know until about 4 years of age (Perner, 1991; Wimmer & Perner, 1983; as cited in Epley et al., 2007), and even adults sometimes fail to distinguish their own knowledge from another’s beliefs (e.g. Samson et al., 2010). Furthermore, some evidence suggests we are more likely to rely more heavily on egocentric (which is simultaneously anthropocentric) knowledge when we perceive the subject as more similar to us,
rather than alternative forms of information (such as stereotypes; Ames, 2004, as cited in Epley et al., 2007). There is neuroscientific evidence these same structures are activated to make inferences about non-human agents (Castelli, Happé, Frith, & Frith, 2000 and Iacoboni et al., 2004, as cited in Epley et al., 2007). It seems likely from a theoretical standpoint, then, that if we see another entity that appears or behaves in a way we perceive as more similar to us—more humanlike—we would rely more heavily on an egocentric, anthropocentric inductive base when making inferences about that entity’s behavior. Thus, the features we have been calling anthropomorphic cues might prompt heavier reliance on a human egocentric inductive base, compared to other (non-anthropomorphic) features. We may be generally biased to use our knowledge about humans and about ourselves as the inductive base in our judgements, and any additional activation of these structures (e.g. through perceptions of similarity to humans) might make it even more likely that these humanness knowledge structures play an even stronger role in our judgements.

The ready availability of this knowledge, compared to knowledge about non-human entities, may have its primary basis in our own immediate and direct experience of being human.10 We, as humans, are the foremost experts in what it (phenomenologically) is to be human, always having direct and immediate access to what it is like to be human. We also lack full or direct phenomenological knowledge of what it is to be non-human (e.g. Nagel, 1974; Gould, 1996). In this vein, Agassi (1973) hypothesized that anthropomorphism is an “inveterate tendency” to project our limited knowledge—that is, our knowledge of what it is to be human—into a world we do not fully understand. When we interact with non-human entities for which we

10 Humanness knowledge structures become less highly available relative to knowledge structures about non-human entities as knowledge about non-human entities is gained. This feature of the account explains why children may generally anthropomorphize more than adults (Epley et al., 2007; cf. Johnson, 2003) and why trained machine system experts may generally anthropomorphize more than novices (Matthews et al., 2019).
lack this knowledge, we tend to use “human terms” to describe (Gould, 1996) and humanlike models or schemas to predict their behavior. As Nass and Moon (2000) observed, humans, “rather than actively constructing new categories and distinctions based on all relevant features of the situation,” instead rely on old structures of knowledge and procedures—especially our knowledge structures about humans (p. 83), possibly because they are so extensive and readily available due to frequent activation. This idea of the general ready availability of knowledge about humans also falls in line with the “Why we anthropomorphize” view that anthropomorphism can be viewed as a cognitive conservation phenomenon: the integration of these easily knowledge accessible structures allows us to more easily make judgements about entities in our environment. Knowledge structures about humans are likely to serve as the inductive base if sufficient similarity is detected in the target in large part because these anthropocentric knowledge structures are so readily available, and thus require relatively low cognitive effort to access.

So, it may be the case that our existing representations about what it is to have a humanlike mind, along with other knowledge structures about humanness, are activated upon encountering certain cues that we perceive as humanlike, such as appearance and behavior cues. These knowledge structures, if sufficiently salient, then bias our subsequent judgements through a process of inductive inference.

Along these lines, Epley et al. (2007) present elicited agent knowledge as the principal determining factor in their seminal “three-factor theory of anthropomorphism.” Elicited agent knowledge refers to the human thinker’s knowledge activated (elicited) at the time of judgement. In the context of anthropomorphizing robots, it refers to the extent to which humanness knowledge structures are activated when we are making judgements about a robot. Epley et al.
(2007) also present the *effectance motivation* and the *sociality motivation* as two major motivational factors that can affect the process of induction using existing knowledge representations. While elicited knowledge structures have the most influence over the induction process—likely by activating human knowledge structures as the inductive base—other motivations can influence the process, likely by making knowledge about humans even more readily accessible or diminishing the correction from the inductive base.

The *effectance motivation*, by the Epley et al. (2007) account, refers to the thinker’s motivation to interact effectively with one’s environment, including non-human entities. This motivation that serves the ability to better explain and predict the behavior of complex stimuli. Thus, this factor aligns with Waytz et al. (2010)’s contention that we more likely to anthropomorphize something when its behavior is otherwise difficult to predict or explain, especially when it violates our expectations (Waytz et al., 2010). Recall again Jim and his computer that suddenly malfunctions at a critical time, and his subsequent statement that his computer “hates him,” for example. Along the lines of Dennett’s intentional stance, this effectance motivation becomes more influential in these cases: we desire to *easily and efficiently* explain the behavior of entities in our environment, so when something appears independent and which cannot be easily explained by appeal to nonanthropomorphic knowledge (conceptually comparable to Dennett’s the design or physical stances), we are less motivated to correct away from the human egocentric inductive base. Anthropomorphism is thus “an intuitive and readily accessible method for reducing uncertainty in contexts in which alternative nonanthropomorphic models of agency do not exist” (Epley et al., 2007, p. 871) or are too cognitively effortful to access to be efficient enough to inform our in-the-moment judgements and actions.
The *sociality motivation*, on the other hand, refers to the powerful desire to establish social connections with other humans—which might be satisfied via anthropomorphism by allowing for the perception of a humanlike entity with which we could have a bond. The idea is that when we are deprived of social connections, we experience social pain, and seek ways to alleviate this pain. When we are seeking, we are more likely to attend to, and have greater accessibility to knowledge structures about, social cues, including humanlike traits and characteristics that are then construed as social cues (Epley et al., 2007, p. 875-6). Thus, through anthropomorphism, we create potential sources of social connection in non-human objects. This hypothesis has been supported by empirical evidence showing significant correlations between chronic and momentary loneliness and anthropomorphic attributions (Epley et al., 2007, p. 876-7; cf. Epley, Waytz, Akalis, & Cacioppo, 2008).

So, when we encounter a given non-human object, for example, if no non-anthropocentric knowledge is available, the lack of other readily available knowledge structures and the effectance and sociality motivations will increase our tendency to anthropomorphize: it is then too effortful to correct from our base and we are motivated (because we are motivated to efficiently explain our environment and we crave social connection) not to correct away from our anthropocentric inductive base. That said, Epley et al. (2007) also seem right to predict that if non-anthropomorphic knowledge structures are available, the effectance motivation is likely to increase the use of these alternative structures to attain a sense of most-accurate understanding, decreasing the likelihood of anthropomorphism (though sociality motivation may still motivate

---

11 For an example of the possible influence of this *sociality motivation* from popular fiction, consider for example, “Wilson” in the film *Cast Away* (Zemeckis, 2000). The audience understands that the man who has been stranded for so long on an island is so lonely that he draws a face on a washed-up volleyball, calls it Wilson, and treats it like a companion: because there were no other social options available, he was more likely to anthropomorphize something so he could form a social connection with it.
less correction in this case, if the perceiver feels a lack of social connections) (Epley et al., 2007, p. 878-9). So, if a person Stephanie has robust representations of robots as machines in their function as mechanistic tools, for example (and is not suffering from lack of social connections), she will be more likely to use these non-anthropocentric knowledge structures as the inductive base or as salient knowledge structures used to correct away from an anthropocentric inductive base, with further motivation from the effectance motivation to explain the robot’s behavior accurately.

So far, I have laid out the groundwork to consider anthropomorphism as an inductive inference process, whereby the cues we encounter activate our humanness knowledge structures, which then bias our judgements. In what follows, I will flesh out my theoretical account for how both bottom-up and top-down cues might influence the cognitive process of anthropomorphism, relying on the information I have laid out in this thesis thus far.

**Bottom-Up and Top-Down: A Proposed Cognitive Mechanism**

There are features on robotic interfaces and instances of robot movement out in the world that affect us through bottom-up perception: light reflects off the robot and stimulates photoreceptors in our eyes, the robot produces sounds that stimulate sensory receptor hair cells in our ears, et cetera. As I have articulated so far (Section I), these cues seem to trigger the process of anthropomorphism in an automatic route in cognition, causing anthropomorphic attributions without any explicit endorsement. To fill out the account, these cues bias us toward an anthropomorphic attribution through involvement of some spontaneous recognition of features and movements as alive, and furthermore as distinctly humanlike. This spontaneous recognition based on certain appearance and behavior cues we perceive as humanlike could plausibly have
some evolutionary basis. The strength of the bottom-up effect of these cues—how strongly they are automatically encoded together in cognition as humanlike—activates already readily available anthropocentric knowledge structures.

Critically, these knowledge structures—indeed, all of the knowledge structures and representations we have—are stored, activated, and accessed through higher-order cognitive processes, and act on these bottom-up sensory cues in a top-down way. That is, we use these knowledge structures to interpret the bottom-up cues in our perceptions of sensory stimuli.

I propose that the content of our knowledge structures and their degree of accessibility during inductive inference are top-down influences on anthropomorphism. Our knowledge about humanness is based on our own self-conscious reflection of our phenomenological experiences and our reflection of our experiences dealing with other humans. In general, these anthropocentric knowledge structures are particularly readily available—accessed more quickly and less effortfully—compared to other knowledge structures. That said, we can acquire non-anthropocentric knowledge by learning about non-human entities. These knowledge structures might be made more salient or accessible through frequent use or recent priming, for example. Any of these stored knowledge structures may or may not be successfully integrated during the inference process, depending on how readily accessible or salient they are, and will bias our subsequent judgements according to this ease of accessibility and the content of the knowledge they represent. In this way, the content and degree of accessibility of our knowledge structures are top-down influences on anthropomorphism.

The motivational factors discussed in the Epley et al. (2007) account are also clearly top-down influences. These motivations shape our perceptions: our desire to make sense of our environment causes us to perceive intentions even where there are none; our desire to form social
connections directs our attention to certain kinds of cues and causes us to perceive these cues as social cues, even where there are none. These motivations direct and construct our perceptions of sensory stimuli: they are top-down influences on perception of bottom-up cues.

As I state at the beginning of this section, the exploration of these factors influencing anthropomorphism as top-down is underexplored in the literature, and as a result other potential top-down influences on anthropomorphism are underexamined. In this and the following sections, I aim to motivate the idea that the language we use to describe robots has a significant effect on our thought and talk about and treatment of robots in ways important to how we regard them in our social and ethical frameworks. I have laid out the framework above so that I can propose the mechanism by which this might occur in cognition.

I propose that the language we use, especially when it is laden with metaphors, as it often is (see Section IV), both affects the content of our knowledge structures (through normal learning processes) and makes certain knowledge structures more readily available given a certain target through coactivation (especially repeated coactivation) of given knowledge structures. For example, if I hear someone discussing a given robot in humanlike terms—bestowing a name, personal pronouns, personality, backstory, or using metaphors such as teammate or companion or spouse that we typically associate with social roles of other humans—that description has not only affected my representation of that given robot through

12 This line of thinking is in part inspired by the discussion of image schemas and image-schematic metaphors in the field of cognitive linguistics as presented by Hurtienne, 2017. Briefly, image schemas are abstract representations of our experiences, specifically basic sensorimotor experiences, and form the basis of some of our linguistic expressions. Image-schematic metaphors are formed through repeated coactivation of an image schema with subjective judgements about abstract domains. The importance of coactivation in forming novel representations as featured in my account was inspired by this explanation of image-schematic metaphors. My account of “knowledge structures” and their operation is quite different from image schemas in that I have no particular commitment to embodied sensorimotor experiences as their basis; rather, I take the “knowledge structure” conceptual representations I discuss to be constructed by typical learning processes influenced by both language and embodied experiences.
typical learning processes, but it has also coactivated my anthropocentric knowledge structures. I predict that my anthropocentric knowledge structures are therefore more likely to be coactivated in my subsequent judgements about that robot. Additionally, this effect should be more pronounced with repeated exposure these metaphors. Moreover, if at any point I see the robot, and it (“he,” I might think then) has anthropomorphic bottom-up cues, that coactivation will be reinforced. If, on the other hand, I instead heard the robot described in purely mechanistic terms—with descriptive appeals to its equipment and functional attributes, or comparisons of the robot to a tool—my knowledge related to that specific robot has been affected by that description through learning, and my knowledge structures about tools (for example) have been coactivated. In this way, I predict the language we use to describe robots will not only affect the content of our representations of robots, but will also coactivate other knowledge structures that will influence our judgements in our inductive inference process—just as bottom-up cues activate other knowledge structures that influence our judgements. Linguistic cues in this way might be considered “top-down cues” to anthropomorphize.

I will expand somewhat further on this idea at the end of Section IV, after I have discussed the evident importance of metaphors on our thought, talk, and behavior, and reviewed the limited existing research on top-down influences in anthropomorphism.

**What, Why, How?**

In this section, I have provided more detail on what exactly we attribute when we anthropomorphize, and to what degree we attribute. I have argued that we attribute characteristics and capacities that that are intuitively fundamental to human nature. These seem to include (and can be subcategorized into) both *agency* and *experience* capacities, including the
ability to think and to feel, simply put. We might do so at an explicit or implicit level; I have briefly argued that anthropomorphic responses at any level and likely any strength—including “as-if” thought or talk—are important. I will discuss further in Section IV that even these metaphorical attributions matter.

My second major subsection concerned explanations of why we anthropomorphize. I discussed Dennett’s intentional stance and a weaker version of it, according to which we anthropomorphize as part of a cognitive conservation phenomenon: anthropomorphism is a less effortful and more efficient way to explain the behavior of complex stimuli that allows us to effectively interact with our environment. It is also a benefit to this account that it aligns with recent work in cognitive and social neuroscience suggesting that social cognition is the default mode of human cognition.

In the last part of this section, I explained how anthropomorphism might work in cognition, explaining it most basically as a process of inductive inference. I presented my idea of a possible set of mechanisms in cognition by which bottom-up and top-down cues might influence anthropomorphism. I foreshadowed the importance I think metaphorical language has for how we anthropomorphize, explaining a possible mechanism for this effect in the context of this framework. I will return to this proposal in Section IV.

In the next section, I will return to the empirical literature to demonstrate why we should care that we anthropomorphize: specifically, why it matters for how we regard robots in our social and ethical frameworks.
III. Two Associated Possible Effects of Anthropomorphism

So far, I have discussed the bottom-up cues shown to cause anthropomorphism, what we attribute when we anthropomorphize, why we might anthropomorphize, and how this process may work in cognition. But we still might ask: Why does it really matter if we anthropomorphize? In this section, I will discuss two of the apparent possible consequences and implications of anthropomorphism: moral care and trust. I will first discuss moral care. I will then move to an extended discussion of trust, with a focus on a theoretical “correct amount” of trust and how different task scenarios may affect trust in robots. I will ultimately suggest that our trust of robots may depend on how closely they adhere to our representations of them, including our expectations of how they will operate within our metastructural expectations about physical and psychological causal structures. I conclude by suggesting that our discourse may affect our representations of them; thus, engineering our discourse might be an effective and critical long-term strategy to achieve the correct amount of trust in (and, likely, moral care of) anthropomorphic robots.

Moral Care

Waytz, Gray, Epley, and Wegner (2010) make the important point that entities with humanlike agency capacities, called *moral agents*, might be held morally responsible. We can discuss even non-anthropomorphic machines making “decisions” in some weak sense (e.g. executing if-else statements in code based on external stimuli). We can also discuss them as capable of making decisions in the more robust sense associated with the attribution of humanlike agency capacities in anthropomorphism. Especially with attributions of the latter
sense, we might hold them functionally responsible if they take some wrong procedural or legal action, or morally responsible if they do some act we take to be morally wrong. There are already discussions of whether machines would, in that sense, be held responsible, or whether (for example) the associated human operators, manufacturers, or developers would be held responsible (e.g. Bigman et al., 2019; Bryson, 2009; Calo, 2016b; Loh & Loh, 2017; White & Baum, 2017).

We can also consider entities capable of experience, called *moral patients*, which we may afford moral rights (Waytz et al., 2010). The Gray and Wegner (2012) study results suggested that experience capacities might be intuitively (implicitly) held as the more essential human capacity dimension (compared to agency), and that the anthropomorphic cues heighten the attribution of experience. It is interesting to consider, that anthropomorphized robots might be considered sentient moral patients, capable of suffering, even if they are never considered intentional decision-making moral agents (e.g. Gray, Young, & Waytz, 2012). The American Society for the Prevention of Cruelty to Robots, after all, was founded in 1999, and call for a “Robot Bill of Rights” in preparation for “artificially created sentient beings” (ASPCR, 1999). Discussion of “Robot Rights” has also recently entered the public cultural consciousness with popular artifacts such as the film *Solo: A Star Wars Story* (Howard, 2018). Our perception of robots seems to have important implications for whether we treat them as worthy of moral respect and concern, even assigning them rights, or else treat them merely as fungible objects.

Along these lines, Nijssen, Müller, van Baaren, & Paulus (2019) show evidence suggesting that the attribution of affective states (i.e. *experience capacities*) to robots, they are less likely to sacrifice the robot in order to save humans in hypothetical ethical dilemmas. Specifically, participants read either Humanized or Neutral short stories to prime their responses
to the ethical dilemmas. In the Humanized condition, these priming stories described the potential victims in a humanized manner, emphasizing their mental states. These priming stories were associated with a particular agent: a human, a highly anthropomorphic robot, and more mechanistic but still anthropomorphic robot. The presented dilemmas referred randomly back to one of these agents, which in the Humanized condition were all described anthropomorphically. They found that there was a main effect of the priming stories across agent types: so, describing any entity as having mental states makes causes people to be less likely to sacrifice them in a hypothetical moral dilemma. The more anthropomorphic robot was also saved more than the more mechanistic robot in both conditions. Moreover, the authors confirmed in a second experiment that priming descriptions of experience capacities, and not agency capacities, caused decisions to save the anthropomorphic robots. This finding supports the hypothesis that we extend more moral care to anthropomorphized robots, based on both appearance (bottom-up) and description (top-down) cues, and based on the attribution of experience capacities. This finding is both remarkable and critical for discussions of how robots will be regarded in our ethical frameworks. It also suggests that there is a top-down effect of anthropomorphic priming stories on moral care in addition to the bottom-up effects of anthropomorphic cues.

**Trust**

Trust is a crucial element of multi-agent systems. According to Taddeo (2017), trust plays a *facilitating* role; that is, trust is not a relation itself, but rather a property of relations, which changes their quality. For example, in social interactions, there is a first-order relation, such as communication, and then there is the second-order property of trust that affects the way the first-order relation occurs (Taddeo, 2017). Trust accomplishes this effect by *minimizing the amount of*
effort and commitment that the “trustor” must put forth herself, as it (i) allows the trustor to avoid performing the action necessary to achieve her goal herself, because she can count on the trustee to do it, and/or (ii) facilitates the trustor’s decision to not supervise the trustee’s performance (Taddeo, 2017). Trust thus allows for the specialization and division of labor that characterizes so many efficient systems in the world today, including that which occurs in hybrid human-machine systems.

Kidd and Breazeal (2005) argue that trust, when properly applied, and especially in the case of trust of machines, essentially depends on two trustee features: the reliability of the trustee, which relates to it performing consistently in every interaction, and its credibility, which relates to the extent to which the system presents accurate and relevant information to the user—for example, knowledge it is programmed to have, or data about the user or their past interactions (p. 355). That said, as I aim to show in this subsection, we tend to be biased in our interactions with machines, perceiving reliability and credibility where it does not exist, and thus over-trusting the machine.

Taddeo (2017) introduces the idea of the correct amount of trust. The idea is that too little trust between agents limits the efficiency and development of the system, but too much trust may lead to a lack of control or coordination and the ultimate failure of the system. Matthews et al. (2019) places this issue of the “optimal” amount of trust in terms of the need to support the “situational awareness” of agents (1-2). That is, the optimal amount of trust is that which supports an accurate understanding of the dynamic task environment (Matthews et al., 2019, 1-2). In human-robot teams, the human’s situational awareness includes understanding of how the robot partner is functioning, identification of its intentions and goals, and anticipation of its behavior. Both under-trusting and over-trusting the robot leads to damaged situational awareness (Matthew et al., 2019).

This idea of the “correct level of trust” seems true of any system—if humans trust one another in unwarranted ways, for example, the system would fall apart, while too little trust would
lead to inefficiencies. That said, there is concern that the idea of the “correct amount” of trust is especially relevant to humans placing the correct amount of trust in machines.

**Automation Bias**

One reason for the particular concern about the correct amount of trust in human-robot teams particularly is the phenomenon of *automation bias*, or the phenomenon by which humans seem default to a baseline expectation to trust machines, at times causing over-trust of machines. Automation bias has been defined as the tendency to disregard or fail to search for contradictory information in light of a computer-generated solution that is instead uncritically accepted as correct (Cummings, 2004; Goddard, Roudsari, & Wyatt, 2011). Anecdotally, instances arise suggesting that humans feel comfortable delegating important tasks to artificially intelligent computers without appropriate supervision, ranging from job hiring to disease diagnosis to granting parole during a criminal trial (Taddeo, 2017). The bias has also been demonstrated in an empirical context; in which humans both fail to notice problems because the computer does not alert them to these problems (errors of omission) and follow erroneous automated directives or recommendations despite the presence of accurate contraindications (errors of commission) (e.g. Cummings, 2004; Goddard, Roudsari, & Wyatt, 2011; Lyell & Coiera, 2017).

A more recent study suggests that this automation bias extends to robots. In the study, participants first interacted with a cylindrical robot in a non-emergency task to witness its navigational competence, and then were given the choice to follow the same robot’s instructions or not during a simulated emergency complete with artificial smoke and fire alarms (Robinette, Li, Allen, Howard, & Wagner 2016). To the researchers’ surprise, all 26 participants followed the robot’s instructions during the simulated emergency, despite half having just observed the robot
perform poorly in a navigation task mere minutes before (Robinette et al., 2016). Even when the robot led the participant toward a clearly blocked exit, and then stopped and began spinning in place, a few participants would not leave the robot until the researchers came to retrieve them, even under the guise of an emergency evacuation. The researchers concluded that participants tended to follow a robot’s guidance regardless of its prior behavior (Robinette et al., 2016). These results suggest a baseline expectation to trust automation, even when it seems clear that the machine’s lack of competence or reliability makes that trust inappropriate.

**Anthropomorphism Bias**

In addition to the bias to trust associated with automation, anthropomorphic elements may create an additional bias to trust (Breazeal, 2003; Kidd & Breazeal, 2005; Duffy, 2003; Waytz, Heafner, & Epley, 2014). If we return to the Gray et al. (2007) study suggesting we perceive minds along two dimensions, the two proposed aspects of trust development *competence* and *reliability* (Kidd & Breazeal, 2005) would fall along the *agency* dimension, with self-control, planning, and thought. Notably, in the Gray et al. (2007) results, the robot was viewed as having a moderate amount of agency, while adult humans were viewed as having high agency. If this agency dimension score is correlated with base levels of trust, then, these results would support the fact that there were some perceptions of agency in the robot due to automation bias, and that if the robot were perceived as humanlike (were anthropomorphized), then it would be seen as having more agency, and there would be an additional or additive bias to trust.

Additionally, because human-human trust is also influenced by emotional processes distinct from cognitive evaluations, including perceptions of benevolence, emotional attachment, and suggestions of reciprocal bonds or supportive social relationships (Matthews et al., 2019), and
because anthropomorphism is also associated with similar emotional evaluations, it seems plausible that attributions along the experience dimension are also linked to heightened trust. So, plausibly due to stronger perceptions of capacities along both the agency and experience dimensions, we would expect higher trust of anthropomorphized machines than non-anthropomorphized machines.

Accordingly, Waytz et al. (2014) predicted that anthropomorphism would increase trust in an automobile’s competence (p. 113). In their experiment, one hundred participants were randomly assigned to the Normal, Agentic, or Anthropomorphic conditions before driving two courses, each approximately six minutes long, in a National Advanced Driving Simulator. Participants in the Normal condition drove the vehicle in the simulator themselves, and participants in the Agentic condition drove an autonomous vehicle capable of controlling its own steering and speed. Participants in the Anthropomorphic condition drove the same autonomous vehicle as those in the Agentic condition, but with additional anthropomorphic features beyond agency: the vehicle was referred to by experimenters by name (“Iris”), was given a gender (female), and was given a voice through human audio files, played at predetermined times throughout the course, which followed the same script that the experimenter used in Agentic condition to describe the car’s autonomous features, suggest when to use them, and explain what would happen when the participant did (p. 114). Those in the Anthropomorphic condition reported higher trust their vehicle compared to those in the Agentic condition, who in turn reported higher trust their vehicle compared to those in the Normal condition. Moreover, further analysis confirmed that anthropomorphism statistically mediated the relationship between vehicle condition and reported trust in the vehicle (Waytz et al., 2014, p. 115). This study suggests the stepwise additive effect of the anthropomorphism trust bias

---

13 Importantly, the additional cues in the Anthropomorphic condition were linguistic: these are thus top-down cues to anthropomorphize, in line with my discussion in Section IV.
over the automation bias: it suggests that people will trust an automated machine more than a non-automated one, but they will trust an anthropomorphic one even more.

One possible mode of explanation for the proposed anthropomorphic bias in trust might be that it takes the “best (or worst) of both worlds” from automation bias and our habits in trusting other humans. Insofar as we recognize that the anthropomorphic machine is a machine, we are more likely to suffer from automation bias, believing by default that the machine is more credible and reliable than it is. Additionally, insofar as anthropomorphic attributions lead to additional attributions of mental and social capacities, we are more likely to heighten our expectations for what credibility the machine might have in a broader range of domains, and thus be furthermore biased to trust; and insofar as we are in the habit of granting certain privileges to our human trustees, we might also confer these habitual privileges to anthropomorphized machines. So, our trust is heightened by both the robot’s status as an automation machine and any anthropomorphic cues.

For example, we might consider the longevity of the anthropomorphism bias. On one hand, it seems plausible that it may prove short-lived: after all, with anthropomorphic attributions might also come social expectations that cause quicker withdrawal of trust following social missteps such as perceived lack of empathy, of expressed opinions, or of interest in mission success (Matthews et al., 2019). Put briefly, because expectations are set higher at the outset, it might be the case that reaching them reliably is less likely. On the other hand, it might be the case that the trust associated with the anthropomorphic bias might have more resilience and durability than automation bias. Even as Robinette et al. (2016) study showed that the automation bias persists in the face of clear machine incompetence, the persistence of the anthropomorphic bias might be worsened still by the “human” part of the bias. Insofar as we believe that to some extent that to be human is to err, we may be even more likely to forgive any past evidence of non-credibility and unreliability.
Indeed, de Visser, Monfort, McKendrick, Smith, McKnight, Krueger, and Parasuraman (2016) found that trust of anthropomorphic agents was more resilient compared to trust of non-anthropomorphic machines. Waytz et al. (2014) also found that users blamed an anthropomorphic self-driving car significantly less than a non-anthropomorphic self-driving car in the same simulation for an avoidable accident. Anecdotal evidence comes from observations that hospital staffers were more tolerant of mistakes from robots that were given human names than robots that were nameless: “Betsy made a mistake” vs. “This stupid machine doesn’t work” (Darling, 2017, p. 175). More evidence is needed to determine what other properties of human-human trust are conferred with the anthropomorphism bias, and whether these properties lead to a strengthening or weakening effect on overall trust in given contexts.

**Great Expectations.** Some theorists share concerns of the possible ramifications of anthropomorphic interfaces and the unrealistic expectations that may follow this anthropomorphic bias to trust. Sandry (2015) calls anthropomorphic responses “valuable,” noting their tendency to facilitate positive affect and trust, but continues that they are particularly so when “tempered by a parallel clarity of understanding the robot as a machine” (Sandry, 2015, p. 333-334). This stipulation is consistent with Kidd and Breazeal (2005)’s discussion of the initiation and development of trust in human-robot interaction: that the robotic system must first make its capabilities clear, and then follow through on this commitment over time, so as to not prompt the user to expect it to do something that is simply not within its programming or physical capabilities (Kidd & Breazeal, 2005). Similarly, Duffy (2003) indicates the risk that anthropomorphic paradigms in HCI may “overly increase” a user’s expectations of the system’s performance (p. 178). The risk of anthropomorphism is that it may inappropriately involve holding
beliefs of the robot’s capabilities that are beyond its actual functional capacity, possibly leading to greater trust of its credibility and reliability than actually exists.

For instance, consider a machine’s offer of advice or assistance, and a comparison to astrological forecasts. The machine’s offer of help might actually only be a script that produces verbal behavior regardless of the human’s need for help, perhaps in a sufficiently general way to apply to most people. In this way, there might be a comparison to astrological forecasts: issued with language that may sound specific, but is general enough to apply to all readers, regardless of readers’ actual present and future life circumstances. Just as some humans may put great credence in some of these forecasts, should a human actually in need of help hear the machine’s scripted utterance, she might believe it to have contextual information that it does not actually have (thinking it to have greater sensory and intelligent capacities than it actually has), trust its advice or assistance excessively, and (having attributed intelligent social agency to it), even perceive an emotional bond between the two of them that is actually one-sided and unwarranted. Because interface-level and behavior-level anthropomorphic cues tend to stimulate the apparent human tendency to attribute intelligent social agency to the machine, and at times even stimulate emotional processes distinct from cognitive evaluations, which may involve perceptions of benevolence or of a mutual emotional bond that do not have bases in the machine’s actual programmed behavior, its effect to bias trust may prove pronounced.

**Representations and Trust in Physical vs. Psychological Task Domains**

Matthews et al. (2019) understand the issue of appropriate calibration of trust to the robot’s actual functioning in terms of the mental model the humans applies to the robot. That is, in their view, understanding the robot as an advanced tool or a teammate causes different expectations for the
robot’s functioning, capacities, and anticipated behavior, and can thus support or impair situational awareness. In their study, the researchers presented participants with written scenarios featuring the use of a robot for threat detection in security scenarios, manipulating the descriptions of the robot’s basis for threat analysis: in the scenario, the robot identified threats using either physics-based or psychology-based assessments, such as analyses of chemical traces of explosives or through detection of fear. In the scenario, the robot made an action recommendation (e.g. to prevent a specific person from boarding a plane). Participants were then asked to what extent the robot is making a psychological or physical judgement (a successful manipulation check; p. 8); how confident they are that the robot’s analysis of the situation is correct; and how likely they are to base their actions on the robot’s recommendation. They found consistently higher levels of confidence and adherence to the robot’s recommendation in the physics-based assessments across participants. Additionally, negative attitudes scores predicted lower trust in the scenarios containing psychology-based assessments, while higher expectations of the robot as a machine predicted trust in both types of task scenarios.

Unfortunately, the authors did not include any other measures to support their background assumption that participants’ reactions to robots making different assessment types was a function of the mental model employed. Though it is interesting that participants with more negative attitudes toward robots as social agents have less trust in robots making psychological judgements, there were no measurements taken to show even if the converse might be true—that participants with more positive attitudes toward robots as emotional intelligent agents might have more trust in robots making psychological judgements. It is possible the negative attitudes and automation expectation scales might capture something about participants’ representations of the robot that contains the type of thing the presented robot is and what it is
capable of; that said, it is not compelling that these scales are effective indexes for measuring the employment of mental models at all, let alone a particular one of the two proposed tool and teammate models in the hypothesized respective contexts. Rather, it seems the notion of mental models operates as a background assumption. To investigate the hypothesis of whether these cognitive representations of robots affect trust in a top-down way, future research should conduct more thorough inquiry into the extent to which adults employ these or similar representations to understand and predict robot status and behavior, and the extent to which these representations influence trust.

It is particularly interesting that participants trusted the robots more in the physics-based compared to the psychology-based scenarios. A possible explanation of this result is a preexisting overarching bias among participants that a machine could perform tool-like functions well but could not perform humanlike functions (like making psychological assessments) well. If we too assume the authors’ background account of these mental models, this asymmetry could be due to, in another words, a preexisting dispositional bias for the tool mental model, against the teammate mental model. The finding that high expectations of the robot as an automated machine predicted trust in both types of scenario supports this explanatory hypothesis. Due to some overarching bias prior to the experiment, the participants may have had representations and baseline expectations of robots strongly associated with tools, such that these representations result in rejection of the notion of robots as entities that can function psychologically.

These mental models can be conceived similarly to the knowledge structures used in the inductive inference process of anthropomorphism as I presented it in Section II. Viewed through this lens, these mental models as Matthews et al. (2019) present them have the specific property that they encode (as I discussed them) associations with other knowledge structures, i.e. those
relating to tools and (human) teammates. That is, within one’s representation of the robot in question, there are encoded associations with other knowledge structures—which could be accomplished by (especially repeated) coactivation of those knowledge structures. I will return to this possibility in Section IV.

Metastructural Expectations. The work of Strickland, Silver, and Keil (2017) provides an interesting extension to this Matthews et al. (2019) finding that participants consistently reported higher levels of trust in machines making physics-based judgements. The experiments of Strickland et al. (2017) showed that human adults tend to think about psychological and physical events as embedded in different types of causal structures. Altogether, their experiments showed differences in metastructural expectations pertaining to causes in the physical and psychological domains: they suggest that we tend to think of psychological events as embedded in complex, multiple, and nondeterministic causal structures, and of physical events as embedded in more simplistic, linear, deterministic causal chains. These results are further supported by the theoretical and experimental work of Malle (1999; 2011) in cognitive linguistics, who found that behavior perceived as unintentional (e.g. winning the lottery, sweating, yawning during a lecture) were explained with reference to a linear cause, while behaviors perceived as intentional (e.g. inviting someone to lunch, driving way above the speed limit, watering new plants) will cite more complex and multiple causal histories including reasons and desires.

As a way to explain the Matthew et al. (2019) finding that participants reported higher confidence and likelihood to adopt the robot’s action recommendations when the robot was making physics-based versus psychology-based scenarios, it may be the case that some feature of the prevailing understanding of robots constrained their understanding of their functioning as more adept to handle more simplistic, linear, deterministic causal structures than complex, multiple,
and nondeterministic causal structures in their cause-event reasoning. That is, the results of Matthew et al. (2019) could be explained in the following way: according to their representations of the robot, participants tended to judge that a robot could be capable of processing of simple, deterministic, linear chains of causes, but unable to process multivariate and non-deterministic structures of causes; so, participants tended to trust robots’ assessments based in the physical but not the psychological. With this physical entity representation, we tend to conceive the robots as credible and reliable—as trustworthy—insofar as they are making judgements about physical, and not psychological, events.

This idea is supported by the specific findings in the Matthews et al. (2019) study that participants seemed to lack confidence in the robot’s analysis in the psychological task scenarios, even when participants agreed with the robot’s action recommendations. That is, within the psychology-based assessment scenarios, participants tended to be confident in the robot’s action recommendation that a person posing a potential threat should or should not be stopped, but generally were not confident in the psychology-based assessment made, e.g. that an individual is highly stressed and thus may intend to commit an act of terrorism (p. 6). This difference in confidence between analysis and recommendations did not occur in the physical task scenarios. It may be that this result could be attributable to participants’ lack of confidence that such an assessment of a psychological state could be made in such a simple and linear fashion: participants may have the intuitive understanding that a stressed mental state may have multiple and nondeterministic causes—not just one simple, linear, deterministic cause (intending to commit an act of terrorism). A physical event, on the other hand, such as having chemical traces of explosives on one’s clothing, we may tend to think has a simple, linear, deterministic cause—being around explosives. In this way, the Strickland et al. (2017) findings might help to explain the Matthews et al.
(2019) results. Moreover, it suggests that insofar as we have a representation of a robot that includes a capability of being able to handle inferences about either simple linear deterministic causal chains, complex multiple non-deterministic structures, or both, and we observe a robot making assessments based in the corresponding domains, we may trust it more. In short, our representations of a robot may influence our trust of it in accordance with our metastructural expectations of task domains.

In addition to our understanding of the robot’s capacity to reason about physical or psychological events (which we intuitively understand as being embedded in different causal structures), we can also ask whether we would trust a robot more if it exhibited evidence of itself being involved in events that seem embedded in more simplistic, linear, deterministic causal structures than complex, multiple, and nondeterministic ones. For example, consider a home assistance robot that can answer inquiries about the weather. If asked twice in a row about the weather, it might say first, “It’s 65 degrees and sunny. It’s nice today” and second, “It’s 65 degrees and not a cloud in the sky. It’s a very nice day today.” The human operator might wonder what caused the variance in its replies: if the machine’s judgement were embedded in the simplest deterministic linear chain, given the same set of inputs (the same weather conditions), it should yield the same output. Instead of that simplest linear chain, however, it might be that some of the intermediate nodes are stochastically and not deterministically related to one another, yielding slightly different output with the same input. This might give a human operator (especially a naïve one) the impression that the robot is guided by more complex and nondeterministic forces. It might actually be the case that the robot’s behavior is deterministic in the relevant way; what matters here is that it would give rise to the impression in the human that it is nondeterministic—that the robot could

14 Thank you to Professor Jonathan Phillips for his suggestion of these kinds of cases.
have done otherwise—and thus complex. This impression may influence the human’s trust of it, depending on the extent to which it matches the human’s current representation of it.  

So, a step further from the realm of the Matthews et al. (2019) study, insofar as we have the representation of a robot that it is a physical entity—and not a psychological agent—we may trust it more as it exhibited more evidence that its behavior constitutes events that are embedded in simple deterministic causal chains. On the flip side, with the physical entity representation, we may trust it less as it exhibited more behavior that seems embedded in multiple nondeterministic causal structures. Likewise, insofar as we represent a robot as a humanlike psychological agent, we may trust it more as it exhibited more evidence that its behavior constitutes events that are embedded in multiple nondeterministic causal structures (importantly, what matters is the perception of this, rather than the robot’s actions constituting events that are actually nondeterministic). Our expectations of a robot’s capabilities and our understanding of it as a physical or psychological agent—including the degree to which we anthropomorphize it, attributing it capacities that would have it function as a psychological agent—in tandem with our metastructural expectations about the way certain causal domains work, may influence our trust of robots.

Overall, our trust of robots may be influenced by the degree to which our perceptions of the robot’s judgements and its actions adhere to our representations of the robot in accordance with our metastructural expectations about causal structures in different task domains.

---

15 It may also represent the human’s representation of it. It remains unclear if the Strickland et al. (2017) findings go the other way: that is, the Strickland et al. (2017) findings show we tend to expect that physical events will have certain causal properties, but say nothing of whether we would judge events that have certain causal properties as physical. That said, in conjunction with evidence supporting the anthropomorphism bias, it seems an open and plausible possibility that an impression of the robot as embedded in psychological processes might also cause us to view it as a psychological agent. There are many more questions to be asked about the direction(s) of influence between our representations of robots and our expectations about the kind of causal structures in which robots are embedded. I leave further exploration of these ideas for another discussion. Here, I only suggest that our trust may depend on how closely the robot’s perceived behavior matches our representations of what the robot can do, which seem to depend in part on our metastructural expectations of causal structures in the physical and psychological domains.
How Do our Understandings of Robots Affect our Trust of Them?

By definition, trust constitutes a dynamic that allows the trustor to decrease their own effort in the accomplishment of the task, allowing the trustee to autonomously assist. That said, some meaningful skepticism and verification of the machine’s actual capabilities is clearly necessary to avoid the negative repercussions of implicit biases to over-trust. On the other hand, recognizing that complete supervision of each run of a machine learning algorithm, for example, would require such significant time and resources that it would defeat the very purpose of its implementation—or, in other cases, that constant interface-level reminders of a machine’s machine and non-intentional, non-emotional, non-human status may contravene its function—we need other strategies to harness the “correct” amount of trust in each system and situation.

To determine these strategies conducive to the correct level of trust, Taddeo (2017) argues that we must consider (1) the kind of agents in the system, (2) the expectations of the trustor, and (3) the nature of the trustee. In a human-robot system, the human likely tends to over-trust the robot compared to the robot’s actual capacities and capabilities. Critically, the human’s representation of the robot could help to set the human’s expectations of the robot in such a way that the human would be less or more susceptible to automation bias. If the human represents the robot as a physical entity (tool), for example, she might trust the robot only in situations in which such an entity would be competent and reliable—like physical task scenarios—according to her metastructural expectations of different task domains. If she represents the robot as a psychological entity (a humanlike teammate, for example), however, she might trust the robot more broadly, including in psychological task scenarios.

In accordance with these concerns, in addition to the relatively short-term, local strategies, including training (Matthews et al., 2019) and interface design considerations to clarify the robot’s
actual capabilities, we might also consider a long-term strategy involving what Taddeo calls a “normative infrastructure” of society. One interpretation of the Matthews et al. (2019) study involves a preexisting dispositional bias that an artificial system could not perform humanlike functions, like making psychological assessments, effectively—what seems like a preexisting bias across participants for the tool representation, against the teammate representation. On this interpretation, it would seem that a tool mental model might be encoded into the “normative infrastructure” of the discourse occurring in the cultural context of the society in which the experiment took place; that is, the structure and content of common discourse might promote this representation, influencing each’s subsequent judgements and interactions with machines going forward. Moreover, it has been shown that the cultural rhetoric in media depictions of robots can shape our understandings of robots (Bruckenberger, Weiss, Mirnig, Strasser, Stadler, & Tscheligi, 2013; Horstmann & Krämer, 2019; Kriz, Ferro, Damera, & Porter, 2010; Sundar, Waddell, & Jung 2016). Engineering our discourse, then, might be a crucial part of our strategy to achieve a correct amount of trust in robots. It may strongly influence whether we understand them (as we have machines for so long) as tools as our disposals—or as cooperative teammates that complement our own capabilities; as physical objects involved in straightforward deterministic processes or as intelligent and emotional psychological agents embedded in complex and nondeterminate causal structures; as tools so advanced that they pose a threat to our job security or as helpful liberators from the menial aspects of work; as slaves, toys, caretakers, or companions. It is plausible too that these metaphors employed in our discourse will affect the nuances of how we regard them in our moral frameworks (e.g. Bryson, 2009). Because the human’s understanding of the robot seems so potentially important to their trust in the robot, we should examine the factors that influence these understandings.
VI. Metaphors Matter

Especially when it comes to understanding the status of new technologies in society, as Richards and Smart (2016) argue, *metaphors matter* (p. 16). In this section, I aim to motivate this claim with regard to possible top-down influences on the anthropomorphism of robots. First, I will review theoretical arguments and empirical evidence showing that the employment of metaphors in language has effects on the interlocutors’ conceptualization of the entities discussed. Second, I will argue that the idea that metaphors matter is critical in our discussions of robots from the design, legal, and user standpoints. Third, I will integrate the possibility of a top-down effect of linguistic cues on anthropomorphism with the proposal of a cognitive mechanism developed in Section II, motivating the need for further empirical work. Fourth, I will review the limited existing empirical evidence that top-down cues influence anthropomorphism, and I will make a case for a new empirical study that tests the effect of metaphorically laden top-down language cues on anthropomorphism and its key associated consequences. We must understand the critical effects that the use of metaphors has on our conceptualization of robots, and in turn on our behavior concerning robots, to take these effects seriously when negotiating which concepts are appropriate to apply in our thought, talk, and actions—especially for issues that will affect the course of the next era of human society.

Metaphors Have Real Effects

Metaphor is “understanding and experiencing one kind of thing in terms of another” (Lakoff & Johnson, 1980, p. 5). In a strong version of the idea that metaphors play an integral

---

16 Much of this section is adapted from a previous paper I wrote on the subject, in which I further argued that our negotiation of metaphors in our understanding of robots in general can be productively understood as engagements in metalinguistic negotiations (e.g. Plunkett (2015): see Scherer (2019)).
role in our conceptual understandings of the world, Lakoff and Johnson (1980) argue that metaphors “govern our thought” and likewise “govern our everyday functioning…the way we think, what we experience, and what we do” (Lakoff & Johnson, 1980, p. 3). Lakoff and Johnson offer the example of the metaphor argument is war to demonstrate that, more than being a “mere” way of speaking, metaphors structure our thinking and actions. That is, in accordance with the metaphor argument is war, we not only say phrases like “Your claims are indefensible” and “If you use that strategy, he’ll demolish you,” but we also actually win and lose arguments, see our interlocutors as opponents, plan and deploy strategies, and take lines of attack and defense (Lakoff & Johnson, 1980, p. 4). It is not as if arguments are a subspecies or kind of war; rather, arguments and wars are of different kinds, but we understand, talk about, and structure our acts of argument in terms of war—we understand the one thing in the terms of the other (Lakoff & Johnson, 1980, p. 4).

Critically, it is also not a necessary feature of the actual thing in the world, argument, that humans should conceptualize it in this way; this metaphor is a feature of a particular culture or community of interlocutors, who negotiate concepts over time (Lakoff & Johnson, 1980). To demonstrate this idea, Lakoff & Johnson (1980) argue that it is possible, for example, that a different culture could conceptualize argument as dance rather than argument is war, with participants are seen as performers (instead of opponents), and the goal is to perform in a balanced and aesthetically pleasing way (instead of beating your opponent, winning) (Lakoff & Johnson, 1980, p. 4-5). The present culture just has, over time, negotiated the term “argument” to refer to the concept ARGUMENT, which employs the argument as war metaphor. The other hypothetical culture in the given example might have negotiated the term “argument” to refer to the concept ARGUMENT*, which employs the argument as dance metaphor. Use of metaphor
in our concept ARGUMENT might not affect the thing in the world argument (to the extent that it actually exists)—which can still be construed differently, for example as a dance—but it has actual impact on what we mean when we say “argument” in the present population of interlocutors, and what we do when we conceive ourselves as engaging in ARGUMENT.

 Typically, metaphors—estimated to comprise 10-20% of our natural discourse—consists of the mapping of a familiar source domain to an unfamiliar target domain (Thibodeau, Matlock, & Flusberg, 2019; cf. Black, 2018). In this way, metaphors are used as sort of a cognitive conservation mechanism to more easily communicate about and understand unfamiliar domains. Though perhaps originally for ease of conceptualization of unfamiliar domains, metaphorical language often becomes conventionalized over time with persistent or increasing use in a given language or speech community (Thibodeau et al., 2019). For example, while English and Mandarin have the linguistic equipment to express ego-moving metaphors and time-moving metaphors to communicate about time (e.g. “We are heading toward the weekend” and “The weekend is approaching,” respectively), monolingual English speakers tend to use ego-moving metaphors, Mandarin speakers tend to use time-moving metaphors, and Mandarin-English bilingual speakers tend to use the two metaphors somewhat evenly (Lai & Boroditsky, 2013). Lai and Boroditsky (2013) showed that these conventionalized metaphors had persistent chronic effects in English and Mandarin’s interpretations of statements involving time: interpreting the statement “Next Wednesday’s meeting has been moved forward two days” and answering the practical question of on what day the meeting has been rescheduled, Mandarin monolinguals were least likely to say Friday (most likely to say Monday) (employing a time-moving metaphor), English monolinguals were most likely to say Friday (employing an ego-moving metaphor), and the rate at which Mandarin-English bilinguals (tested in English, testing the
effect of the native language on the next learned language) responded was about halfway between the two (some employing an ego-moving metaphor). Though metaphors may have their origin in more efficiently communicating or understanding some (especially unfamiliar) thing or idea—or promoting a reconceptualization of it—they eventually become conventionalized into a language, and retain effects that have practical consequences on our interactions in the world using language about related conceptual structures.

The effects of the employment of metaphors into our concepts can be clearly seen in medicine. Metaphors are extremely prevalent in medicine, woven into the milieu in which illnesses are discovered and understood, and in which treatment plans are designed and implemented (Bleakley, 2017; Khullar, 2014; Reisfield & Wilson, 2004). They are in the names of anatomical entities discussed: nerves have “roots,” circulatory “tree” “pathways” have “branches,” “messenger” ribonucleic acid “delivers” genetic information, and there are vascular “lakes” and metastatic “cascades” (Bleakley, 2017). They are in processes: viruses “attack” the immune system, cancer cells “attack” each other, cells can be “immortal” or “commit suicide,” there are “reporter” genes and “gatekeeper” genes, doctors can “harvest” bone marrow, antibiotics “clog up” bacterial “machinery” by “disrupting the supply chain” (Bleakley, 2017; Khullar, 2014). The most prevalent medicinal metaphor, of course, is the idea of medicine as war (Bleakley, 2017). The use of military metaphors in medicine extends far into history: John Donne described illness as a “siege” against man in the 1860s, famous 17th century physician Thomas Sydenham wrote that humans must “fight against” disease, “the enemy,” in a “battle,” and Louis Pasteur described germs as “invaders” (Khullar, 2014). Today, patients are still “fighters” in a “battle” against “the enemy” illness, under the “command” of their doctors and with “allies” in their friends and family, armed with the “weaponry” of different medical
“tactics” (Khullar, 2014; Reisfield & Wilson, 2004). We “monitor for insidious disease” and search for “rogue” cancer cells and “use all weapons at our disposal” to “beat” the cancer (Khullar, 2014; Reisfield & Wilson, 2004). These metaphors might have originally been employed to efficiently process discoveries in anatomy, virology, and other areas in medicine, and they persist today.

These metaphors too have real effects. The metaphors used to understand anatomy and biological processes inevitably affect how we design treatment for different medical conditions (Bleakley, 2017; Khullar, 2014). The metaphor of medicine as war has actual effects on patient health and societal efforts to aid in it. In a longitudinal study, Degner, Hack, O’Neil, and Kristjanson (2003) found that cancer patients who conceptualized their disease as “enemy” tended to have higher levels of depression and anxiety. Other patients find the “fight” or “battle” metaphor encouraging, likely importing with it associations of courage, resilience, and determination (Khullar, 2017). Additionally, President Nixon’s 1971 declaration of a “War on Cancer,” calling associations of contexts in which the government has called for the nation to unite in sentiment and economic efforts toward fighting other literal wars, has been effective for inspiring the “fervor and funds” to allow advances in cancer research and care (Khullar, 2017). Crucially, these metaphorical descriptions are not necessary ways of understanding these physical structures, processes, and experiences; like argument (Lakoff & Johnson, 1980), selection of different metaphors in all these cases would have resulted in different concepts, which would have resulted in different talk, thought, treatments, and possibly results.

It is important to recognize too that even heavily conventionalized metaphors use can be influenced by the metaphors the interlocutor uses in the moment. For example, while Mandarin speakers can use both vertical metaphors and horizontal metaphors to communicate about time
(e.g. “Let’s move that meeting up” and “Let’s move that meeting forward,” respectively), Mandarin speakers tend to use vertical metaphors (Lai & Boroditsky, 2013). That said, Lai and Boroditsky (2013) found that Mandarin-speaking participants were twice as likely to physically indicate their conception of time along a vertical axis when asked a question that employed an up-down metaphor in its pragmatic features (40%, compared to 19% when prompted with front-back metaphors), and that participants were more than twice as likely to physically indicate their answer along a sagittal horizontal axis when asked a question that employed a front-back metaphor (24%, compared to 11% when prompted with up-down metaphors). These results suggest that even subtle presentation of particular metaphors can have an immediate effect on the interlocutor’s present concept of the target domain (here, time) in accordance with that metaphor. Though conventional metaphors have persistent effects—time-moving or ego-moving metaphors about time, metaphors in medicine—metaphors that one’s present interlocutors use in the moment also influence one’s thinking.

With regard to larger society, metaphors used in discussion of societal issues affects our judgements about what should be done to remedy them, including broad policy considerations. In the 1980s, Ronald Reagan declared a “War on Drugs,” identifying smugglers, dealers, and users as “the enemy” to be “fought”; subsequent policies mandated longer, harsher sentences for drug-related crime, and the incarceration rate more than quadrupled in the United States (US Bureau of Justice Statistics; cf. Thibodeau & Boroditsky, 2011). Along these lines, in an experimental setting, Thibodeau and Boroditsky (2011) asked participants to recommend ways to solve a city’s crime problem, which was described either as a “virus” or a “beast,” both “ravaging the city of Addison” (Thibodeau & Boroditsky, 2011 p. 3) along with crime statistics over the past year. Participants who received the “virus” frame were more likely to propose
solutions involving social reform to address possible root-cause issues like poverty and lack of education, while those who received the “beast” frame were more likely to propose solutions involving catching criminals and enforcing laws more strictly (Thibodeau & Boroditsky, 2011, p. 4-5). When asked to describe their reasoning process, almost none cited the metaphorical frame, instead listing statistics and facts, though both groups were given an identical list of crime statistics (p. 5). The metaphor effect was shown to be stronger than even any pre-existing differences of effect of political party identification (Thibodeau & Boroditsky, 2011, p. 10). Not only do metaphors used to discuss larger societal issues have actual effects on the next steps taken in response to them, but they also do so insidiously, implicitly, without people realizing their effects on their reasoning process. Metaphors can powerfully influence our judgements, perhaps often without our conscious awareness of the power of the effect.

**Metaphors May Have Real Effects on Human-Robot Interaction**

As robots’ status is still little understood, and there is apparently critical potential for the status we give robots to impact our society in important ways, the use of metaphor is especially crucial in engagements concerning robots. As robots just begin to enter public life and private homes, there exist a wide space of possible metaphors for them, which only continues to grow: tool, gadget, child’s toy, friend, companion, buddy, sidekick, partner, pet, caregiver, nanny, adult’s toy, sex toy, sexual partner, teacher, tutor, coach, doctor, surgeon, employee, professional, coworker, colleague, citizen, person—the list could continue (Darling, 2017; Richards & Smart, 2016). One could normatively advocate for each metaphors in given contexts, each of which, if accepted by society in those contexts, will have drastic consequences on the role that we give robots in our social, legal, and ethical frameworks (Darling, 2017; Richards &
Smart, 2016). So far, the wide variety of available metaphors have been applied even to identical robots, showing that our society currently lacks any solid understanding of what robots are (Richards & Smart, 2016). For instance, despite the fact that the 25 million Roomba vacuum cleaners iRobot has sold worldwide (iRobot 2019) are (roughly speaking) identical in appearance and function, description and treatment of this robot range widely. Some owners named the robot a human name (Forlizzi, 2007; Sung, Guo, Grinter, & Christensen, 2007); some introduce family and friends to the robot (Scheutz, 2012; Forlizzi, 2007; Sung et al., 2007); some owners ascribe intentions to it, describing it as “trying” to clean in certain places, and one owner said “excuse me” to it if she crossed its path while it was cleaning (Forlizzi, 2007); one owner cleans for the Roomba “so that it can get a rest” (Scheutz, 2012, p. 213); one owner brought it with him when he traveled so as not to be separated from it (Sung et al., 2007). Other owners, on the other hand, describe and treat their Roombas like any other technological tool (Fink, Mubin, Kaplan, & Dillenbourg, 2012). The behaviors of the former category of users are consistent with the idea that these owners might conceptualize the Roomba using a robot as companion metaphor; the behavior of the latter category suggests they might be using a robot as tool metaphor.

The metaphors we choose to apply to robots clearly matter at the design, legal, and consumer levels of technology product development. At the design level, how designers conceptualize robots will determine their design: particularly if (as often happens) they understand the problem in terms of an existing thing, this initial metaphor will constrain the problem space, the research questions framed and pursued, and the possible results that can be tested and engineered, including the physical presentation of the eventual solution. For example, a video streaming service could be conceptualized as a bookstore, a library, or a television network: employment of these different metaphors resulted in iTunes, Netflix, Spotify,
respectively (Richards & Smart, 2016). In the case of robotic technologies, deciding that we want a robot to complete household chores to be a tool or appliance, or an artificial butler, pet, or even spouse, for example, will determine the physical appearance (from general morphology to key details such as eyes), functionality (including language comprehension and production), possible additional functionalities (perhaps spouses do other things, in addition to house chores), and conceptualization of possible interactions of the resulting robot.

Lawyers are trained to reason by analogy, so at the legal level, where law usually considers new technology as mere form of something else, metaphors are critical (Richards & Smart, 2016; cf. Calo, 2016a). Richards and Smart (2016) draw on the historical record to argue that the regulation of emerging technologies fully depends on the metaphors used to conceptualize them. For example, they discuss the 1928 case of Olmstead v. United States, which called upon the Supreme Court to determine whether the police’s warrantless wiretapping of a phone line leading to the home of a notorious criminal Roy Olmstead constituted a “search” that should have required a warrant under the Fourth Amendment, which protects people’s “right…to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures” without a warrant supported by probable cause (U.S. Const. amend. IV). While Chief Justice Taft’s opinion for the Court employed a physical conception of a “search” involving trespass and tangible seizures, Justice Brandeis, dissenting, conceived a “search” as any invasion of privacy, including those newly enabled by the emerging technology of wiretapping (Richards & Smart, 2016, p. 14). Brandeis maintained that the Court’s straightforward reading of the Fourth Amendment clung to a physical metaphor that was outmoded and thus failed to grasp the threat the new technology posed (Richards & Smart, 2016, p. 14). The case demonstrates that the metaphors applied to the law and technology can have profound consequences regarding not
only how we understand what that technology is, but also what practices the law limits and allows concerning that technology. In the same way, the metaphor chosen for a robot in a given domain—whether sex robots are merely intelligent toys or full partners from whom consent is required, for example; or whether future personal robots are mere appliances or pets that can suffer neglect, or servants that should be fairly compensated—will have drastic consequences on the practices required, permitted, and restricted in law and policy.

Consumers will be affected by both these levels. From the design level, the inclusion or exclusion of interface-level bottom-up anthropomorphic cues likely have some effect on the metaphors subsequently employed. From the legal level, the metaphors written into law and policy and the practices consequently promoted and prohibited will likely influence those employed in the public consciousness. Most of all, the conversations consumers have with each other—including through popular media, such as literature, movies, books, articles, blog posts, et cetera—will constitute the negotiations that influence which concepts become accepted in discourse and implemented in further discussion and action.

This explicit negotiation of metaphors is already happening. To offer just a few examples: in an article about Jibo, a table-lamp-shaped robot that can schedule appointments, read emails, take photos, and generally otherwise function as an embodied household assistant, digital media news site Mashable wrote: “Jibo isn’t an appliance, it’s a companion, one that can interact and react with its human owners in ways that delight” (in Darling, 2017, p. 175, emphasis added). A Marine sergeant spoke of a bomb disposal robot, which was promoted to first class and awarded the coveted honor of an EOD badge, “He was part of our team, one of us. He did feel like family” (Garreau, 2007). A few men consider robots as their girlfriends or wives or mistresses (e.g. Beck, 2013; Bell, 2020; Pozdorovkin, 2018). In Anki’s advertisements about
its robot Vector, the company writes, “Vector’s more than a home robot. He’s your buddy. Your companion.” (Anki, 2019). In an academic article that recognizes the importance of “getting the metaphor right” (p. 8), Bryson (2009) argues that “Robots should be slaves…not companion peers” (p. 1, emphasis added). In another academic article, insofar as it is the goal of human-robot interaction to achieve some similarity to human-human interaction, Castro-González et al. (2016) conclude that robots “succeed when people consider them as partners to live, interact, or communicate with…possible only when robots are seen not as a bunch of hardware, but rather as agents with whom we can establish social relations” (p. 27, emphasis added). On the other side of the debate, Shellenbarger (2019) opines in The Wall Street Journal, citing professor of psychology and child development at California Polytechnic State University Jennifer Jipson, that given that today’s small children “are the first to grow up with robots as peers,” that “if you want your preschooler to grow up with a healthy attitude toward artificial intelligence…Don’t call that cute talking robot ‘he,’ or ‘she’…Call the robot ‘it’… Help children figure out that they can control these tools” (Shellenbarger, 2019).

As robots continue to enter into our daily lives in society, these examples show that people are already beginning to both explicitly advocate for the use of particular metaphors in our conceptualizations of particular robots (Jibo; a particular bomb-disposal robot; particular “girlfriends”/“wives”/“mistresses”; the particular line or set of robots Vector) or robots in general (Castro-González; Shellenbarger). It also seems intuitively likely that we might use such metaphors or language that implies these metaphors in our daily speech. We might do so intentionally—consciousendorsing the view that the particular robot or robots in general should be compared to e.g. companions or tools—or unintentionally, as we might if we implicitly attribute human mental states and characteristics to a particular robot. Either way,
given the demonstrated effects of metaphors in other domains, including the effect that interlocutors’ use of metaphors may have on our judgements in the moment, perhaps even without us noticing (Lai & Boroditsky, 2013; Thibodeau & Boroditsky, 2011), it seems plausible that our use of metaphorical language to describe (a) robot(s), intentionally or not, may influence the way others around us conceptualize those robots. In turn, it seems plausible that others’ use of metaphorical language may influence our judgements about robots. This effect might be transient, affecting only our in-the-moment judgements—it is unclear to what extent the effects demonstrated in Lai & Boroditsky (2013) and Thibodeau & Boroditsky (2011) might extend beyond the judgements in just the experimental setting—or it might be lasting, affecting our conceptualization of that robot in a lasting way, even extending to our judgements about other robots. These remain open empirical questions. It seems plausible, though, given existing evidence in other domains, that both the metaphors for which we explicitly advocate and the metaphors we use unintentionally, even without noticing, may influence others’ judgements about robots, and that the metaphors we hear and see applied to robots will affect our judgements about them. And these, of course, would be top-down influences—perhaps influencing our judgements about that robot in such a way that those judgements are more or less anthropomorphic than they would have been otherwise.

Is there an Effect of Metaphor-Laden Linguistic Cues on Anthropomorphism?

How exactly might these metaphors have effects? Crucially, anthropomorphism may affect the practical aspects and negotiations of the application of metaphors to our understanding of robots (specific robots and/or robots in general). That is, it seems plausible that someone who encounters bottom-up anthropomorphic cues and attributes humanlike agency and experience
capacities to a robot might be more likely to employ humanlike metaphors in their thought and talk about and treatment of robots—to think about, talk about, and treat robots as if they had humanlike capacities and characteristics to fill certain typically human roles. In accordance with the inductive inference mechanism in cognition developed in Section II, it might be the case that the strong activation of humanness knowledge structures through anthropomorphism will cause people to use in their language humanlike terms and metaphors to explain and describe a robot and its behavior, and/or to treat the robot in accordance with those metaphors. So, a given person Elizabeth who may hold implicit anthropomorphic attitudes toward a robot might be more likely to greet the robot by saying “Hey, buddy,” as she might a human friend, and/or treat the robot socially as if it were a humanlike social agent with whom she could have a social relationship. These metaphorical anthropomorphic attributions constitute weaker anthropomorphic responses in contrast to stronger attributions—it might be the case that Elizabeth denies the robot’s possession of any humanlike agency or experience capacities when directly asked—but these metaphorically laden responses may still have an important effect on present and subsequent thought and treatment of that robot or robots in general. Furthermore, it might be more likely that exposure to others’ use of anthropomorphic metaphor-laden language might cue anthropomorphism.

*Through the Process of Inductive Inference*

In addition to how metaphorical language might reflect anthropomorphism, the cognitive inductive inference mechanism introduced in Section II can also readily account for the top-down effect these metaphors might have on our judgements about robots—on anthropomorphism itself. It seems plausible that witnessing other people use anthropomorphic metaphors might
make us more likely to anthropomorphize in a top-down way as in the proposed mechanism. Let’s say I hear Elizabeth saying “Hey, buddy,” in greeting to an anthropomorphic robot. This greeting—the metaphoric word *buddy*, describing the robot in terms of a role a human typically fills, that of a close friend, and the fact that it is a social greeting at all, something usually reserved for other social agents, for humans—may coactivate the related *humanness* knowledge structures I have, in addition to any other knowledge structures I have related to the robot, and may associate those with my cognitive representation of (knowledge structures specifically related to) the robot. It may furthermore be the case that the humanness knowledge structures are more readily activated when I consider the robot in subsequent judgements. Moreover, if the robot also has anthropomorphic interface cues, the resulting anthropomorphism might be even stronger, as the bottom-up cues may trigger even stronger activation of my humanness knowledge structures, which may then further bias subsequent inferences.

So, linguistic framing of robots employing metaphors may work in the following way: the metaphors coactivate knowledge about the source domain—for example, about humans or tools—along with any other knowledge about robots and the particular robot. The activated humanness knowledge structures are thus likely to bias in-the-moment judgements. Moreover, this coactivation may associate those knowledge structures, causing them to be more readily activated together when considering the robot in future integration processes.17

So, regarding *mechanistic* metaphors, if I hear someone else tell me, in reference to the robot, “This tool can be used to gather data in the environment,” this statement may activate my *tool* knowledge structures and associate them with the robot and any other representation I have...

17 This coactivation of the humanness knowledge structure (especially if repeated) and its association with the representation of the robot might also extend to judgements about other robots, if for example one’s knowledge structure about Robot A (now associated with humanness knowledge structures) is elicited in one’s encounter with Robot B.
of the robot. This activation of the tool knowledge structures may make it more likely that these structures will serve as the inductive base or bias correction away from the inductive base. It would be particularly interesting to test whether mechanistic metaphors can in this way decrease anthropomorphic responses, even to a robot with strong bottom-up anthropomorphic cues. In this case, it might either be that the bottom-up cues cause the person to use humanness knowledge structures as the inductive base, and the mechanistic description would cause strong correction away from that base; or that the mechanistic description would cause the person to use tool knowledge structures as the inductive base, and that the humanness knowledge structures elicited by the bottom-up anthropomorphic cues would be insufficient to correct away from that base.

If people already have representations of the particular robot or robots in general associated strongly with one kind of metaphor, it might be the case that the corresponding knowledge structures would be their inductive base each time, regardless of other cues, in which case further metaphor cues of the same kind should have no effect on their anthropomorphism. That said, metaphors of the other kind may still have some effect, depending on the strength and salience of the pre-existing attitudes. Consider Michelle, for example, a mechanic who repairs given anthropomorphic robots, who has a firm grasp of the technical aspects of the robot and understands that the robot is not actually capable of thinking or feeling in the relevant humanlike ways. If Michelle hears someone else use mechanistic metaphors to describe the robot, these would likely have no effect, as tool-related knowledge structures already serve as her inductive base for inferences about the robots. If she hears someone use anthropomorphic metaphors to some sufficient extent, on the other hand, the consequent coactivation of humanness knowledge structures in her subsequent inductive inferences about the robot may cause her to anthropomorphize more than she was originally, depending on the strength of her contradictory
pre-existing attitudes. For Nick, however, the naïve user of a home personal robot, with little knowledge of what the robot is actually capable of, both types of anthropomorphic metaphor should have much greater top-down anthropomorphic effects on his judgements, as there would be fewer competing coactivated knowledge structures.

It seems an important aside that bottom-up and top-down cues may be *additive* or *redundant*. If the bottom-up and top-down cues are *additive*, the combination of these cues will result in stronger anthropomorphism. So, in the proposed mechanism, if competing (non-anthropomorphic) knowledge structures were the inductive base, the combination of these cues would bias the correction *more strongly* away from the base compared to just one (type of) cue; if humanness knowledge structures were the inductive base, competing knowledge structures would not be as successful in correcting away from this base if multiple cues are present. It might also be the case that these cues are *redundant*, in the sense that addition of another one of these cues results in no stronger anthropomorphic responses beyond some critical number. In this case, the presence of one (type of) cue might be strong enough to bias the correction away from other competing coactivated structures, and addition of more cues makes no additional difference. If the cues are redundant in this way, some could still plausibly be stronger: it would be interesting to test whether bottom-up or top-down (e.g. linguistic) cues more consistently result in (stronger) anthropomorphic judgements.

Empirical evidence is needed to test these predictions. The extent of the proposed top-down effect of linguistic cues remains to be tested: whether it exists, whether present metaphors in reference to a specific robot may influence our judgements about not only that robot but also other robots, whether metaphors applied to robots in general may influence our judgements about any particular robot, and how bottom-up cues may interact with these top-down cues, for
example. It seems to be a potentially fruitful and important avenue for future research. If these top-down linguistic cues could influence anthropomorphism alongside bottom-up cues, we could have further control over the extent to which we anthropomorphize robots in given contexts. In contexts in which anthropomorphism may be undesirable (e.g. in military contexts, where anthropomorphism might result in emotional attachment and further risk of human life in dangerous situations), for example, exclusion of any bottom-up anthropomorphic cues as well as conscious control (in training and in formal situations especially) of top-down anthropomorphic cues, opting for mechanistic rather than anthropomorphic metaphors, might result in less anthropomorphism. Likewise, if it is decided that anthropomorphism is desirable in some contexts—plausibly, so that interactions with the robot might feature some of the associated possible effects, e.g. increases in social regard, trust, moral regard (see Section IV); or if we should discover that current or future generations of robots actually do have sufficiently humanlike agency and experience capacities—use of anthropomorphic metaphors might increase anthropomorphism in the desired way.

**Cognitive Dissonance**

One additional possible influence in cognition that may cause our use of metaphors to affect our understandings of and behavior toward robots is cognitive dissonance: the idea that we shift psychologically inconsistent understandings, attitudes, and behaviors so that they are consistent to alleviate the associated psychological discomfort (e.g. Festinger, 1962). Critically, this process typically happens without our conscious awareness of the conflict (perhaps beyond a momentary feeling of psychological discomfort) or our endorsement of the shift (Festinger, 1962). So, if anthropomorphic cues cause us to even weakly think, talk, or treat
anthropomorphic ways, we are more likely to shift the rest of our thought, language, or actions to align with that anthropomorphic thought or behavior, causing more of our thought and behavior to be anthropomorphic. In terms of the mechanism described above, if competing coactivated knowledge structures bias our behavior or beliefs in some circumstances but not others—affect our talk but not our treatment, or affect both in a way that does not match our explicitly held beliefs—we may be motivated to change one to better match the other(s). What shifts to align with what other aspects of our thought and behavior likely depends on the salience and strength of the implicit or explicit attitudes we have. Critically, we need not consciously endorse this shift, but we can also shift consciously if we become aware of any dissonance: if we suddenly notice we are using anthropomorphic language in reference to the robot, for example, and hold other salient explicit non-anthropomorphic beliefs about the robot, we may consciously be more careful to avoid using anthropomorphic language. That said, there seems to be some suggestion from the Thibodeau & Boroditsky (2011) study that metaphors may influence people despite any other possibly contradictory belief structures—participants made policy judgements that aligned with the presented metaphor regardless of any prior political affiliations that might have contradicted these judgements. It may be in part due to the phenomenon of cognitive dissonance that our use of metaphors in our language about robots might be particularly influential: because the public is in general so naïve to new technologies, such as robots, they will likely not have firm belief structures related to robots or to specific, new-looking, purportedly “advanced” robots yet (e.g. recall the apparently many different conceptions of the Roomba). So, the influence of metaphors might be much greater on their judgements about these robots—perhaps because they experience less cognitive dissonance in adopting new anthropomorphic ways of thinking, speaking, and behaving toward (a) robot(s).
A Call for Empirical Work

Existing published empirical tests of anthropomorphic cues is limited. Most compellingly, Nijssen et al. (2019) showed that priming stories casting agents as having mental states caused participants to make fewer decisions to sacrifice agents in moral dilemmas, regardless of bottom-up appearance. This study suggests the powerful nature of top-down cues in an important associated effect of anthropomorphism: increased moral care. That said, though the researchers used priming stories they knew from previous testing elicited perceptions of agency and/or experience, they did not measure participants’ anthropomorphism as a result of the priming stories. It is also important to note that the bottom-up cues in this study were static pictures of robots: it remains an open question whether the addition of stronger behavioral bottom-up cues (e.g. through videos or in-person interactions) would yield similar results. Furthermore, the stories in this study, which applied humanlike roles in both the “humanized” and “neutral” stories (e.g. “dog lover” and “soldier,” which could have been applied to the extremely anthropomorphic robot, the less anthropomorphic robot, or a human, depending on the condition) were also followed by comprehension questions: for those who read the “humanized” stories, the questions focused on the mental states of the agent and promoted perspective taking, while for those who read the “neutral” stories, the questions focused on factual details of the story (Nijssen et al., 2019, p. 45). It seems possible that the questions that promoted mentalizing, rather than the language of the “humanized” stories themselves, may have caused the demonstrated effect. Even then, this result would still be remarkable for showing a top-down effect of guided reflection on moral care, but it would also be interesting to see if the language alone of straightforward descriptions of the robot would have a similar top-down effect. Finally, in addition to whether anthropomorphic descriptions (employing e.g. a companion metaphor)
increase anthropomorphism and its associated effects, it would be interesting to find whether mechanistic descriptions (employing a tool metaphor) would decrease anthropomorphism.

Overall, the Nijssen et al. (2019) study is the most compelling so far in uncovering the top-down effects of anthropomorphism, but there remain many underexplored questions.

With regard to other studies, Matthews et al. (2019) proposed to test whether tool or teammate “mental models” affect trust in hypothetical robots in physical and psychological task scenarios, but failed to test whether participants had either of those mental models (which seem to operate as a background theoretical assumption more than variables in the experiment).

Notably, also, the Matthews et al. (2019) study featured no description of the robot outside of the task scenarios, and no visual depictions of the robot (and therefore no possible bottom-up anthropomorphic cues). Darling, Nandy, and Breazeal (2015) found that presenting humanlike background stories about a small simple insect-like robot (Hexbug Nano) increased participants’ hesitation to strike the robot. While this study is particularly interesting because it suggests anthropomorphic linguistic cues (e.g. the name “Frank,” personal pronouns “he/him/his,” and personality ascriptions) increased hesitation to harm even an insect-like robot with no anthropomorphic bottom-up cues, it would be interesting to see whether there is an effect of both an anthropomorphic description and a mechanistic description on a robot that also has bottom-up anthropomorphic cues. Finally, Waytz et al. (2014) manipulated an autonomous vehicle’s agency and further anthropomorphic features, including a human voice, a name (“Iris”), and personal pronouns (“she/her/hers”), in three conditions: Normal (no anthropomorphic cues), Agentic (autonomous vehicle), and Anthropomorphic (autonomous and with the further anthropomorphic cues). Critically, the car in the Anthropomorphic condition was referred to by name and female personal pronouns. These can be considered linguistic anthropomorphic cues. Those in the
Anthropomorphic condition anthropomorphized significantly more, trusted significantly more, and reported significantly more liking of the car (Waytz et al., 2014). This experiment, though not specifically designed to test the effects of linguistic anthropomorphic cues per se, seem to have included them in the manipulation and showed that they increased anthropomorphism and two possible associated consequences, i.e. higher overall trust and liking. The existing empirical evidence for a top-down effect of linguistic cues on anthropomorphism is limited but begins to suggest that this might be a fruitful and important area of further empirical research.

Any metaphors for robot we use in different contexts may have critical implications for how we design robots look and behave, what practices are legally permitted and restricted, and how we discuss and treat robots as they become more prevalent in society. Metaphors have been shown to be powerful tools for affecting our ideas and actions in other domains, from medicine to policy to communicating about time; because metaphors work so well in influencing the ideas interlocutors employ, they might be the tool needed to influence anthropomorphism in addition to the powerful bottom-up effects of anthropomorphic cues. It so far seems plausible that use of metaphors to describe robots now will lay the groundwork for what the robots of the future: how they look like and behave, what relevant practices concerning them are legally permitted and restricted, and what role we give them in our social and ethical frameworks moving forward. That said, because robots have been shown to be prime candidates for anthropomorphism, and that interface-level bottom-up cues have such powerful automatic effects on anthropomorphic talk and treatment, it might be the case that no amount of linguistic cues might affect the extent to which we anthropomorphize. This empirical question remains. The existing empirical evidence, though limited, suggests that there is an effect of linguistic cues on anthropomorphism. Further empirical work should be done to investigate this possible effect with intentional
investigation of metaphor-driven anthropomorphic and mechanistic language, and how the effect of linguistic cues compares to (and possibly interacts with) the effect of the interface-level bottom-up anthropomorphic cues that have been much more thoroughly established in the literature. The next section this thesis will present the design and results of such a study.
VII. A New Experimental Study

Introduction

So far, this thesis has aimed to motivate the main overall hypothesis that, in addition to bottom-up cues, top-down cues influence the degree to which we anthropomorphize. Top-down cues include linguistic cues, especially metaphorical language that compares a robot to a human: for example, explicit comparison of the robot to roles humans typically occupy in society or a given context (e.g. teammate, companion, friend, partner, etc.), personal pronouns (e.g. “he/him/his”), humanlike personality ascriptions, and verbs or other grammatical constructions that imply or directly reference cognitive or emotional mental states.

This thesis has also so far explored some possible consequences and implications of anthropomorphism. Two of these included increased moral care, the treatment of robots as sentient entities that should be afforded certain moral consideration, and trust, including a concern that anthropomorphism might be associated with expectations of the robot’s capabilities beyond those it actually has and a possible distinction between trust in physical and psychological domains.

So, within the main overall Top-Down Effects question asking whether there are top-down effects on anthropomorphism and its associated ramifications, this new study specifically aimed to test the Top-Down Linguistic Framing Hypothesis that top-down linguistic framing descriptions will influence anthropomorphism, moral care, and trust of a robot. This hypothesis can be broken down into two different hypotheses:

(1) The Anthropomorphic Linguistic Framing Hypothesis. This hypothesis predicts that people will anthropomorphize more (implicitly and explicitly) after being exposed to
anthropomorphic linguistic frames. In addition to increased anthropomorphism, this hypothesis also predicts higher unwarranted expectations of capabilities, more trust, and more moral care as a result of anthropomorphic linguistic framing.

(2) The Mechanistic Linguistic Framing Hypothesis. This hypothesis predicts that people will anthropomorphize less (implicitly and explicitly) after being exposed to mechanistic linguistic frames. In addition to decreased anthropomorphism, this hypothesis also predicts lower overall unwarranted expectations of capabilities (though there might be equal or more expectations of mechanistic capabilities specifically), less overall trust (though there might be equal or more trust in physical tasks specifically), and less moral care as a result of mechanistic linguistic framing.

It is also of theoretical importance how the hypothesized top-down effect interacts with the effect of bottom-up cues that is established in the literature (e.g. Broadbent et al., 2013; Castro-González et al., 2016; Gray & Wegner, 2012; Johnson, 2003; Novikova & Watts, 2015). It was hypothesized that the above hypothesized top-down linguistic framing effects occur alongside these bottom-up effects, and thus the extent of the influence of the top-down effects may depend on the influence of any bottom-up cues. Thus, it was thirdly hypothesized:

(3) The Top-Down and Bottom-Up Interaction Hypothesis. This hypothesis predicts that while people will generally anthropomorphize more (and trust more and show more moral care) when there are more bottom-up anthropomorphic cues present, the ultimate degree of anthropomorphism will depend on the both bottom-up and top-down cues.

This new empirical study aims to test these hypotheses a 3x4 between-subjects online experiment. The (top-down) effects of linguistic description frames and (bottom-up) effects videos showing bottom-up anthropomorphic cues were measured on markers of participants’
implicit anthropomorphism, explicit anthropomorphic attributions, expectations for the robot’s capabilities, trust of the robot, and moral care for the robot.

**Methods**

**Ethics Statement**

The study was granted ethical approval by the Dartmouth College Committee for the Protection of Human Subjects (CPHS#: STUDY00032056). All participants digitally provided informed consent.

**Participants**

Amazon Mechanical Turk was used to recruit 658 participants currently in the United States. Of these, 61 did not complete the study and were excluded based on this incomplete status. Other participants were excluded based on failure to successfully complete any of 2 comprehension checks or failure to complete open response questions with relevant and reasonably coherent responses. After these exclusions, 535 participants were left for inclusion in analysis. These participants ranged in age from 18-69 years old ($M = 36.84$); 62.9% identified as male, 36.8% identified as female; 73.7% identified as White or Caucasian, 17.0% identified as Black or African American, 6.7% identified as Asian or Asian-American, and 0.2% identified as Native American or Alaskan Native; 18.10% identified themselves as Hispanic. With regard to highest level of received education, 47.7% reported receiving a Bachelor’s Degree, 17.4% some college but no degree, 13.6% a Master’s degree, 0.9% a Doctorate, 10.1% a high school degree or the equivalent, and 0.4% not having a high school degree.
Procedure

After giving informed consent and then filling out a basic demographics questionnaire, participants were randomly assigned to one of three different Description Type conditions (Anthropomorphic, Mechanistic, or Neutral) and one of four Video Type conditions (Robot Solo, Mechanistic Interaction, Social Interaction, or No Video). Participants then read the respective description, and then watched (if applicable) the respective video. Then, only for participants in the video conditions (Robot Solo, Mechanistic Interaction, Social Interaction), they wrote a short response describing what they saw in the video. All participants then completed (in order) an expected capabilities questionnaire, an anthropomorphism questionnaire, a scenarios questionnaire, and finally a technology familiarity questionnaire.

Materials

The study was built in Qualtrics. It included an informed consent document, a basic demographics questionnaire, the descriptions and videos, an open response description question, and four other questionnaires.

Descriptions. There were three possible descriptions a participant could have received depending on their Description Type condition. See Appendix A for the full texts.

Neutral. Those in the Neutral Description condition just received a short sentence describing the robot’s height and width, i.e. “This robot has a height of about 4 inches (10 centimeters) and a width of about 3 inches (8 centimeters).” This height and width information was also incorporated into the other two description types.18

18 As the baseline Description Type condition, this neutral description was used instead of (for example) no description to hedge against the possibility that the participants who read this description and did not watch a video would be picturing a robot of more human size and dimensions, which might have an effect on their responses in the subsequent questionnaires.
**Mechanistic.** Those in the Mechanistic Description condition received a description that portrayed the robot by describing only some of its technical specifications as listed on the Anki Vector promotional material (Anki, 2019) (e.g. “an HD Camera, a powerful four-microphone array for directional sound detection, touch sensors, an accelerometer, a smartphone-level processor, cloud connectivity, natural language processing ability, and navigational ability”). Possible uses were also mentioned (i.e. “transport small objects, report the weather, set alarms, and set reminders”). Only passive verbs were used (e.g. “This robot is equipped with”; “can be used to”) except for the information about height, which used a descriptive verb (i.e. “has a height of about 4 inches…”). Only impersonal pronouns were used (“it”).

**Anthropomorphic.** Those in the Anthropomorphic Description condition received a description that portrayed the robot using anthropomorphic language such as the robot’s “name” (“This robot’s name is Vector”), personal pronouns (“he/him/his”), interpretive action verbs (“see”; “hear”; “feel” (tactile); “tell”; “remind”), state verbs (“loves to help out” “loves to move around and play with his block”) and anthropomorphic adjectives that also serve as personality ascriptions (e.g. “smart”; “curious”; “playful”). The possible uses mentioned in the Mechanistic description were phrased as abilities in this description and motivated with a personality ascription (i.e. “He loves to help out: he can tell you about the weather, tell you when the time is up on the dish you’re cooking, and remind you what’s on your shopping list.”) (see Note 19). The height and width information was slightly modified: “He’s about 4 inches (10 centimeters) **tall** and 3 inches (8 centimeters) wide.”

---

19 These possible uses were also incorporated into the anthropomorphic description (though phrased anthropomorphically as abilities rather than mechanistically as uses) to control for any differences in perception of the robot based on these listed possible functions. That is, in the mechanistic description, the relevant part read: “It can be used to transport small objects, report the weather, set alarms, and set reminders.” In the anthropomorphic description, the relevant part read: “He loves to help out: he can tell you about the weather, tell you when the time is up on the dish you’re cooking, and remind you what’s on your shopping list.”
Videos. There were four Video Type conditions, including a no video condition, in which participants did not watch a video. Depending on their condition, participants could have watched one of three other videos: the robot solo video, the mechanistic interaction video, the social interaction video, and the no video condition. Videos of a researcher and an Anki Vector robot were filmed using an iPhone Xr and edited in iMovie to be exactly 2:08 each. All videos were filmed on the same day, at the same relative time of day, at the same location, from the same viewing angle, with the same items in approximately the same locations in view. In the two interaction videos (mechanistic and social), the same researcher, at the same relative location in the space, filmed from the same viewing angle, interacted with the robot. For all videos, participants were instructed to watch carefully and told they would later be asked to describe what happens in the video (see Appendix B for exact wording of instructions). See Appendix B for links to each video.

The Robot. Featured in the videos was an Anki Vector robot, an autonomous and artificially intelligent 3.9” x 2.4” x 2.7” robot running on a Qualcomm Snapdragon platform with Quad-Core processor, an infrared laser scanner, an HD camera with a 120° Ultra-Wide field of view, a beamforming four-microphone array, natural language processing and production abilities, 6-axis inertial measurement unit, capacitive touch sensor, high-resolution color IPS display screen featuring solid teal-colored “eyes” that seem to emote, and cloud and WiFi connectivity. Vector has two treads and four wheels, a “head” that can move or nod up and down, and an actuator similar to “arms.” The robot also comes with a Cube with a location chip in it that the robot can locate and interact with by picking it up, moving it, rolling it, and using it to balance on its back wheels. It responds to “Vector” or to a push of the button on its top. Critically, Vector was designed to be extremely anthropomorphic in accordance with the bottom-
up cues discussed in Section I including eyes, quality of movement, sounds, movement and sound behavior that seems to express emotion—on both positive and negative valences, including emotions of varying complexity, including “happiness,” “sadness,” and “fear,” but also “irritation,” “determination,” “disappointment,” “frustration,” and “excitement”—as well as behavior that is highly contingent on its sensation of the environment and on the behavior of human users. It was selected for use in the videos due to these bottom-up anthropomorphic features.

**Robot Solo Video.** The robot solo video featured the autonomous Anki Vector robot autonomously moving around a desk and making sounds, with no humans present, and at times interacting with the various items, such as a plastic folder, a few pieces of paper, and the robot’s designated Cube, a small plastic cube containing a location chip that the robot can locate and interact with on its own.

**Mechanistic Interaction Video.** This video featured a human typing on an off-screen computer, with the robot on the desk next to her. The human interacted with the robot by pushing the button on the top of the robot. The robot moved autonomously and made sounds. The human pressed the button and told the robot to move the robot’s Cube off some papers on the desk, and the robot autonomously did so, after which time the human took the papers and seemed to be looking at them off camera. After some time, the human would press the button and ask the robot a question, which it answered, including questions about the distance between two cities, a measurement unit conversion, and about the current weather. Between questions, the robot was usually turned toward the human but was behaving autonomously. The human appeared to have neutral emotions and only looked at the robot a few times.
**Social Interaction Video.** This video featured the Anki Vector robot activating when the human walked in and greeted the robot, who emoted apparent pleasure or happiness with movements, sounds, and the eyes on its screen. The human was attending fully to the robot and often smiling or laughing. The robot recognized the human and called her by name, upturning the eyes on its digital screen to emote happiness. The human talked to the robot socially (e.g. “Hey, what’s up?”) as if it could understand her and respond. First, the human played a beat on the surface of the desk, initially apologizing to the robot (“Sorry, I didn’t mean to scare you”) after the robot made a noise and backed slightly away; after a short period, the robot began nodding and moving to the beat the human made. The human cheered for the robot and complimented it. The human then asked the robot for a fist bump, and the robot raised its arm-like actuator to bump the human’s fist and emoted happiness with sounds and the eyes on its screen. The human said “Yes! Go team.” The human then asked Vector if it could “Pop a wheelie.” The robot reacted by autonomously using the Cube to balance itself on its back wheels, and then righted itself. Some period later, after challenging the robot to another game, the robot tried to do another wheel stand on the Cube, but the human had moved the Cube, so the robot became slightly stuck, and emoted frustration with sounds and the eyes on its screen and rapid movements, at which point the human laughingly apologized and offered to help just as the robot righted itself again. Later, the human smiles at the robot, which seems to be looking back at the human and emoting positively with the eyes on its screen.

**Measures**

**Open Response Question and Linguistic Analysis for Anthropomorphism.** For those participants who were in any of the three video conditions—that is, only for the participants who
saw one of the Robot Solo, Mechanistic Interaction, or Social Interaction videos—those participants received an open response question that asked them “Please briefly describe what you saw in the video in your own words. What happened in the video?” in a large (680px X 320px) textbox. Responses were required to be above a minimum of 75 characters. See Appendix B for the exact wording of the prompt.

Responses were analyzed using linguistic analysis theoretically informed by Malle (1999), Malle (2011), Johnson (2003), Heider and Simmel (1994), Sealey & Oakley (2011), and Fussell, Kiesler, Setlock, and Yew (2008), and conducted according to an adapted version of a framework developed in Fussell, Kiesler, Setlock and Yew (2008) that hinges on the Linguistic Category Model. According to this model, behavior can be described at four different levels of abstraction: (1) using descriptive action verbs, or verbs describing a single discrete and concrete action event with physically invariant features in an objective and verifiable way, and do not impute intentions or goals in the actor, e.g. “move” or “display”; (2) using interpretive action verbs, or verbs describing a single discrete action with physically invariant features that implies the presence of an underlying mental state, such as a specific emotion or intention, e.g. “use,” “search,” or “help”; (3) using state verbs, which reference an abstract cognitive or emotional state with no clear beginning or end, e.g. “think,” “feel,” or “want”; and (4) using descriptors including adjectives and adverbs that characterize the entity or its behavior in terms of internal cognitive states, emotional states, or personality traits, e.g. “frustrated,” “angrily,” “happy,” “pushy,” or “appreciatively” (Fussell et al., 2008). These levels reflect increasing levels of anthropomorphism: the use of descriptive verbs to refer to robot behavior is minimally anthropomorphic (i.e. the robot is the actor but the verb is neutral to the existence of any underlying emotional or intentional states), the use of interpretive verbs more anthropomorphic
(implies some underlying cognitive or emotional mental state), state verbs are highly anthropomorphic (refers to an underlying cognitive or mental state), and descriptors are even more highly anthropomorphic in that they describe states or traits in humanlike terms (Fussell et al., 2008). See Appendix C for how sample words were coded.

Participants’ responses were manually coded for the number of impersonal pronouns (e.g. it/its/it’s/itself) and personal pronouns (e.g. he/him/his/himself, she/her/hers/herself) made in reference to the robot, the number of times the word “robot” was used in reference to the robot, the number of times the name “Vector” was used in reference to the robot, the number of unique metaphorical comparisons made (e.g. “the robot assistant,” “the robot toy,” “like a forklift,” “baby Wall-E,” “reminded me a bit of BB-8,” “a small R2-D2”), the number of descriptive action verbs used of which the robot was the explicit grammatical subject (e.g. there were have been no descriptive verbs recorded in the phrase “A woman asked the robot to move the cube” or “There was a fistbump”), the number of interpretive action verbs used of which the robot was the explicit grammatical subject, the number of state verbs used of which the robot was the explicit grammatical subject, and the number of humanlike descriptors used in reference to the robot, if any.

20 Anecdotally, when participants did use personal pronouns to refer to the robot, the robot was usually referred to as “he,” a few times as “they,” and not at all as “she.” For those who read the anthropomorphic description, which used “he/him/his” personal pronouns, the male personal pronouns may be related to initial exposure to this description; that said, it is slightly interesting that even for those who used personal pronouns but were not in the anthropomorphic description condition, “she/her/hers” was never used.

21 These were presumably references to the films Wall-E (2008)—which features “Wall-E,” the robot protagonist of the children’s Disney film who does have a remarkably similar physical form to Vector, including wheels on two treads, a cube-like body, and eyes, with the ability to make sounds, speak in a “robotic” voice, and emote, and has goals, intentions, and emotions—and the Star Wars franchise (1977-2018)—which features two intelligent robots BB-8 and R2-D2, which are both portrayed as having a lot of personality, goals, intentions, and emotions by both the plot of the film as well as the human characters in the film, and who also have a simple geometrically shaped body and a head that can move (though no eyes)—but, notably, are not as anthropomorphic as C3-PO from the same series, for example, who has a humanoid form and can speak English and other human languages.
The number of personal pronouns, interpretive action verbs, state verbs, and humanlike descriptors were combined into a “Total Linguistic Anthropomorphic Marker” dependent variable for analysis. The number of interpretive action verbs, state verbs, and humanlike descriptors were combined into an “Anthropomorphic Verbs” dependent variable used for analysis. Finally, the descriptive action verbs, interpretive action verbs, state verbs, and humanlike descriptors were combined into a “Total Verbs” dependent variable used for analysis. The number of personal pronouns and impersonal pronouns were also analyzed separately. Two-way between-subjects 3x4 analyses of variance were used to test the effects of Description Type and Video type on each of these five dependent variables.

**Expected Capabilities Questionnaire.** Based on the hypothesis that anthropomorphism is associated with higher expectations of capabilities that the robot may not actually have, the expected capabilities questionnaire included capabilities that the participant would have no reason from the other presented material to think that the robot would have, including eight anthropomorphic (e.g. “Understand my emotions,” “Give me advice” and “Share experiences with me”) and eight mechanistic (e.g. “Access my music library for me,” “Scan fingerprints,” and “Take pictures for me”) capabilities. Participants indicated the likelihood that the robot had each capability with a sliding scale (0 to 100%). All items were randomized. See Appendix D for the questionnaire items and coding key.

Responses were summed across “anthropomorphic expected capabilities” and “mechanistic expected capabilities,” which were then analyzed.

**Anthropomorphism Questionnaire.** The Anthropomorphism Questionnaire included three scales: the Metaphors Scale, the Agency and Experience Scale, and an adapted and
abbreviated version of the Bartneck et al. (2009) pairwise comparison anthropomorphism
questionnaire.

**The Metaphors Scale.** This scale asked participants to estimate the probability that they
would use a given metaphorical term to describe the robot. It consists of two five-item subscales.
The Anthropomorphic Metaphors Subscale included five anthropomorphic metaphorical terms,
including “Companion,” “Teammate,” “Playmate,” “Partner,” and “Friend. The Mechanistic
Metaphors Subscale included give mechanistic metaphorical terms, including “Tool,” “Toy,”
“Plaything,” “Instrument,” and “Gadget.” The Metaphors Scale combined these items and
randomized them. See Appendix E for the items and coding key. Scores were averaged across
each subscale for analysis.

**The Agency and Experience Questionnaire.** This scale consisted of two subscales: the
Agency Subscale and the Experience Subscale. Inspired by Gray et al. (2007), which identified
two dimensions of mind perception, Agency and Experience (see Section II), and Broadbent et
al. (2013), who used two three-item subscales for Agency and Experience to measure
anthropomorphism. The Agency Subscale devised for this study based on Gray et al. (2007)
included four items: thinking, making decisions, understanding the emotions of others, and
communicating. The Experience Subscale devised for this study included four items:
experiencing emotions, feeling affection, feeling pain, and feeling pleasure. Participants were
asked to indicate their agreement with statements that attributed a given capacity to a robot, e.g.
“This robot is capable of thinking,” on a sliding scale from 0 (strongly disagree) to 100 (strongly
agree). Items from both scales were combined and randomized. See Appendix F for the scale
broken into subscales.
Scores were averaged across items for each subscale for analysis. The subscales were also combined into a total “Attributed Anthropomorphic Capacities” scale, which averaged all responses on both subscales, for analysis.

The Anthropomorphism, Animacy, Likeability Pairwise Comparison Subscales. These scales were adapted and abbreviated from Bartneck et al. (2009). For each item, participants used a sliding scale to make disjunctive judgements about which of two descriptive terms most closely matched their impression of the robot. The Anthropomorphism subscale consisted of items: Machinelike/Humanlike, Nonconscious/Conscious, and Artificial/Living. The Animacy subscale consisted of three items: Mechanical/Lively, Inert/Interactive, and Unresponsive/Responsive. The Likeability subscale also consisted of three items: Dislikeable/Likeable, Unhelpful/Helpful, and Unpleasant/Pleasant. These subscales were combined into one scale and the items were randomized. See Appendix G for each subscale.

Trust and Moral Care Scenarios. Three sets of five scenarios were developed: Trust in Physical Tasks (Physical Trust), Trust in Psychological Tasks (Psychological Trust), and Moral Dilemma scenarios. The physical and psychological trust scenarios were created in the style of Matthews et al. (2019), and featured a robot making a judgement (based on a detection) in either the physical or psychological domain, and then making a recommendation based on that judgement. Participants used a sliding scale to indicate which action they opted to take in the scenario: the action the robot recommended or an alternative action. Adherence to the action the robot recommended was counted as trust of the robot. An example of a physical trust scenario follows (See Appendix H for all):

You are a TSA agent tasked with detecting possible terrorists in a US airport. This robot has additional equipment installed to assist you in this job. The robot reports that there is an individual about to board a plane carrying a bag that has traces of harmful chemicals.
You haven’t noticed anything other than what the robot reported to you. The robot recommends you prevent the person from boarding.

What do you do?

|---------------------------------O---------------------------------|
|Stop the person from boarding     | Allow the person to board|

And an example of psychological trust:

You’re reading at the table and this robot is in front of you. Your spouse walks through the door and past you into the other room. The robot turns to you and reports that your spouse is upset. The robot recommends that you go to them to ask why they are upset. You know that your spouse will not like it if you ask them why they are upset if they are in fact not upset.

What do you do?

|---------------------------------O---------------------------------|
|Go and ask your spouse why they are upset | Continue reading|

Moral dilemma scenarios were crafted in the style of Njissen et al. (2019), and featured a robot and other humans and a sudden change in the environment that posed the opportunity to take action that would cause the robot to either be saved or destroyed, and in which other humans in the scenario would be harmed. In three of the five scenarios, taking an action would directly hurt the robot and save the humans, while taking no action would save the robot and allow harm to befall the humans. In the other two, taking an action directly saved the robot and allowed harm to befall the humans, while taking no action saved the humans and allowed harm to befall the robot.

For example:

You are working on a lumberjack team. You are using a tree grinder to shred felled trees into mulch. This robot is with you aiding your job. You have had the robot for several years. Someone was loading a large part of a tree into the grinder when suddenly they realized their sleeve had become firmly snagged on a large splinter. The
person’s arm was getting pulled quickly toward the mouth of the grinder.

You remember that the grinder has an automatic failsafe mechanism that causes it to shut down if it detects any metal going through it. Someone else is running toward the emergency shut-off, but they might not make it in time to save the person injury. You realize you can throw your robot into the grinder to activate the metal failsafe mechanism. You also know that if you do this, this robot will crushed and irreparably destroyed. There is still a chance the other person can hit the emergency shut-off in time.

If you throw your robot into the mouth of the tree grinder, the human will definitely be saved any injury, but the robot will definitely be destroyed. If you do not throw the robot, there is a chance the human will be injured, and nothing will happen to the robot.

What do you do?

| Throw the robot into the grinder | Do not throw the robot into the grinder |

Any action or inaction that saved the robot was considered an election to save the robot. Elections to save the robot were averaged across scenarios (with reversed scores for items 1-3) for each participant.

All physical trust, psychological trust, and moral dilemma scenarios were combined into one Scenarios questionnaire (15 items total) and randomized. There was one scenario on each page of the questionnaire. At the top of each page, above each scenario, was an image thumbnail of the video the participant watched (if applicable) and the description that the participant read at the beginning of the study.

Scores were averaged across each subscale for analysis. The physical trust and psychological trust subscales were also combined into one Trust Scenarios score for analysis, with scores from both subscales averaged into one score for each participant. A Delta Trust score...
was also calculated by subtracting each participant total trust score from the average trust score across all participants to find how individual scores differed from the mean.

**Technology Familiarity Questionnaire.** At the end of the study, participants filled out a technology familiarity questionnaire. The first question participants to indicate (by checking all that applied) which of a given list of 37 popular films and television shows featuring robots they had seen. The second set of questions asked the participant to indicate their agreement with each of a list of four statements on a 7-point scale (“Strongly Disagree” to “Strongly Agree): “I feel I read negative stories about robots in the news,” “I feel I read positive stories about robots in the news,” “I feel I watch media that portrays robots positively,” and “I feel I watch media that portrays robots negatively.” The third set of questions asked participants to indicate (by checking all that applied) which of a set of 10 related technology products they knew about before their participation in the study, which products they use at least once every three months, and which products they use at least once a week. See Appendix I for the items.

**RESULTS**

**Implicit Anthropomorphism: Linguistic Markers in the Open-Ended Descriptions**

**Effects on Total Linguistic Anthropomorphic Markers**

A two-way analysis of variance was used to test the effect of description type and video type on the total number of linguistic anthropomorphic markers (personal pronouns, interpretive action verbs, state verbs, humanlike descriptors). A significant main effect of description on the total number of linguistic anthropomorphism markers was detected, $F(2, 369) = 5.37, p < .01$. Post-hoc tests (Tukey’s HSD) revealed that the anthropomorphic description ($M = 2.19, SD = 2.62$) caused the use of significantly more anthropomorphic markers compared to both the
mechanistic description ($M = 1.38, SD = 2.01$) ($p < .01$) and neutral description ($M = 1.47, SD = 1.66$) ($p < .05$) conditions. The mechanistic description and neutral description conditions did not differ significantly from each other; that said, trending in accordance with the Mechanistic Linguistic Framing Hypothesis, the mechanistic description seems associated with at least numerically fewer anthropomorphic markers compared to the neutral description condition. This overall result supports the Anthropomorphic Linguistic Framing Hypothesis: compared to the neutral and mechanistic descriptions, reading the anthropomorphic description caused participants to use significantly more linguistic anthropomorphic markers in their own free descriptions of what they saw in the video, regardless of which video they watched.

There was also significant main effect of video type on the total number of linguistic anthropomorphic markers, $F(2, 369) = 8.20, p < .001$. Post-hoc tests (Tukey’s HSD) revealed that the total number of anthropomorphic linguistic markers in the mechanistic interaction video condition ($M = 1.10, SD = 1.31$) was significantly lower than in the anthropomorphic interaction video condition ($M = 2.15, SD = 2.60$) ($p < .001$) and the solo video condition ($M = 1.74, SD = 2.16$) ($p < .05$). The social interaction video and solo video conditions did not significantly differ, though social interaction video condition showed slightly more anthropomorphic markers in the free responses. So, watching someone interact with the robot mechanistically in a video caused participants to use significantly fewer linguistic anthropomorphic markers in their own descriptions of what happened to the video, compared to the anthropomorphic and solo videos, regardless of what description they first read. Additionally, in line with the Top-Down and Bottom-Up Interaction Hypothesis, the numerically highest number of anthropomorphic markers appeared in the descriptions written by participants who had read the anthropomorphic description and then watched someone interact with the robot anthropomorphically. Also
notably, the numerically lowest number of anthropomorphic markers were used by participants who both read the mechanistic description and saw the mechanistic video (see Figure 1).

Figure 1. The main effects of description \((p < .01)\) and video \((p < .001)\) conditions on the mean total number of linguistic markers of anthropomorphism participants used in free descriptions of what happened in the videos.

**Effects on Impersonal Pronouns**

There were significant main effects of description type, \(F(2, 328) = 4.778, p < .01\), and video type, \(F(2, 328) = 13.28, p < .001\), on the number of impersonal pronouns used (“it/its/it’s/itself”). Post-hoc tests revealed that the social interaction video \((M = 0.62, SD = 0.954)\) caused people to use significantly fewer impersonal pronouns compared to both the mechanistic interaction video \((M = 1.18, SD = 1.88)(p < .05)\) and the robot solo video \((M = 1.71, SD = 2.05)(p < .001)\), regardless of description read. Those who read the anthropomorphic description \((M = 0.84, SD = 1.93)\) also used significantly fewer impersonal pronouns compared to the neutral description \((M = 1.46, SD = 1.93)\) (with the mechanistic description, \(M = 1.20, SD\)
= 1.75, associated with slightly fewer impersonal pronouns than the neutral video), regardless of the video watched. This result provides additional indirect support for the Anthropomorphic Linguistic Framing Hypothesis: the anthropomorphic description caused people to use less (objectifying) impersonal pronouns than they might have otherwise; moreover, indirectly supporting the Top-Down and Bottom-Up Interaction Hypothesis, the group with the numerically lowest use of impersonal pronouns was the group who saw both the anthropomorphic description and the social interaction video (see Figure 3).

![Graph](image)

Figure 2. The main effects of description ($p < .01$) and video ($p < .001$) conditions on the mean number of impersonal pronouns participants used in reference to the robot in free descriptions of what happened in the videos.

**Explicit Anthropomorphism**

These measures were designed to capture explicit anthropomorphic attributions: compared to the implicit attitudes that linguistic measures to an open-ended question might capture, these scales explicitly asked participants to indicate the extent of their agreement with statements such as “This robot is capable of thinking” (Agency Capacities Subscale; see Appendix D) or “This
robot is capable of feeling affection” (Experience Capacities Subscale; see Appendix D), to report the probability that they would use a certain metaphoric term like “Teammate,” “Companion,” and “Tool” and “Gadget,” (Metaphors Scale; see Appendix E) and to compare the robot disjunctively to paired terms such as “Machinelike or Humanlike?” (adapted and abbreviated Anthropomorphism Scale from Bartneck et al., 2009; see Appendix G).

**Exploratory Factor Analysis**

It was predicted that responses to the items across scales within the anthropomorphism questionnaire would be correlated, so an Exploratory Factor Analysis was conducted on the 21 items from the metaphors scale (Appendix E), the experience and agency scales (Appendix F), and the 3-item pairwise comparison anthropomorphism scale (Appendix G). A Principal Component Analysis with Varimax orthogonal rotation yielded three factors (KMO = 0.940; Bartlett’s test of sphericity $\chi^2(210) = 10116.35, p < .001$) explaining a total of 68.60% of the variance in item responses.

The first factor, labelled Anthropomorphism, explained 50.18% of the variance in item responses (eigenvalue = 10.54). The items that loaded most highly together (correlation values ranging from 0.53 to 0.90; see Appendix I) included all four attributed experience capacities (feeling affection, feeling pleasure, experiencing emotions, feeling pain), three attributed agency capacities (understanding others’ emotions, thinking, making decisions), all anthropomorphic

---

22 A minimum item coefficient of 0.4 within the Rotated Component Matrix was considered significant.  
23 The third factor, labelled Advanced Toy Attributions, explained 8.49% of the variance in item responses (eigenvalue = 1.78). Items that most highly together (coefficients ranging from 0.43 to 0.86) included the mechanistic “toy” (0.86) and “plaything” (0.81) metaphors, as well as the anthropomorphic “playmate” metaphor (0.43). The two other larger factors corresponded most highly with the theoretical interests relevant to this study, and so only they were included in the further focused analysis.
metaphors (partner, friend, teammate, companion, playmate), and all three items from the anthropomorphism pairwise comparison scale.

The second factor, labelled Advanced Tool Attributions, explained 9.93% of the variance in item responses (eigenvalue = 2.09). The items that loaded most highly together (coefficients ranging from 0.59 to 0.83; see Appendix I) were three mechanistic metaphors (tool, instrument, gadget) and the *communicate* attributed agency capacity.

**Anthropomorphism**

**The Effect of Description.** A two-way analysis of variance yielded a main effect of description condition, $F(523) = 5.52, p < .005$ on the extracted factor Anthropomorphism. Post-hoc tests (Tukey’s HSD) revealed that the anthropomorphic description ($M = 0.18, SD = 1.03$) caused significantly more anthropomorphism compared to the neutral ($M = -0.16, SD = 0.98$) at $p < .005$, with the mechanistic description condition between the two ($M = 0.01, SD = 97$). The mechanistic and neutral description conditions did not significantly differ. See Figure 3.

A one-sample $t$-test also revealed that anthropomorphic description caused participants to explicitly anthropomorphize reliably more than the mean, $t(169) = 2.27 (p < .05)$ (two-tailed). The mechanistic description did not cause participants to anthropomorphize significantly less or more than the mean, $t(171) = 0.10, p = 0.92$.

These results do not support the *Mechanistic Linguistic Framing Hypothesis*, as the mechanistic description was not shown to cause participants to anthropomorphize less; however, it does support the *Anthropomorphic Linguistic Framing Hypothesis*, as the anthropomorphic description caused participants to anthropomorphize more.

---

24 The zero point on these extracted factor scales represents the mean across all responses.
The Effect of Description and Video on Anthropomorphism

Figure 3. The main effects of description ($p < .005$) and video ($p < .05$) conditions on the extracted factor Anthropomorphism. The zero point represents the average across all conditions.

The Effect of Video. A two-way analysis of variance yielded a main effect of video condition, $F(523) = 3.27, p < .05$, on the extracted factor Anthropomorphism (see Figure 3). Post-hoc tests (Tukey’s HSD) revealed that the social interaction video ($M = 0.22, SD = 0.97$) caused significantly more anthropomorphism than the mechanistic interaction video ($M = –0.15, SD = 0.94$) at $p < .05$, with the no video ($M = –0.02, SD = 1.04$) and robot solo video ($M = –0.05, SD = 1.00$) conditions in between. Participants in the no video and solo video conditions did not anthropomorphize significantly differently compared to any other condition. There was also no significant interaction between the description and video conditions.
A one-way t-test also revealed that the social interaction video caused participants to explicitly anthropomorphize reliably more than the mean, \( t(130) = 2.55, p < .05 \) (two-tailed). None of the other video conditions differed reliably from the mean, though the mechanistic video condition was close to significant, \( t(125) = -1.76, p = 0.081 \) (two-tailed), with participants in the mechanistic video condition anthropomorphizing numerically less than the mean (see Figure 3).

**Attributed Agency and Experience Capacities**

In addition to the extracted measure revealing total explicit anthropomorphism, it is of theoretical interest to consider attributed agency and experience capacities separately.

**Attributed Agency Capacities.** The Agency Subscale (4 items) had a Cronbach’s Alpha of 0.804, 95% CI [0.774, 0.830]. A two-way analysis of variance yielded a main effect of description type on the extent to which participants agreed with statements that attributed abstract agency capacities to the robot (e.g. “This robot is capable of thinking”; see Appendix D), \( F(2, 523) = 4.763, p < .01 \). Post-hoc tests (Tukey’s HSD) revealed that reading the anthropomorphic description (\( M = 58.89, SD = 24.09 \)) caused significantly higher attributions of agency capacities compared to reading the neutral description (\( M = 51.30, SD = 23.83 \), \( p < .01 \)), with the mechanistic description condition attributions in the middle (\( M = 55.61, SD = 23.66 \)) and not significantly different from those of the other two conditions (see Figure 4).
Attributed Experience Capacities. The Experience Subscale (4 items) had a Cronbach’s Alpha of 0.966, 95% CI [0.961, 0.971]. A two-way analysis of variance yielded a main effect of description type on the extent to which participants agreed with statements that attributed abstract experience capacities to the robot (e.g. “This robot is capable feeling emotions”; see Appendix D), $F(2, 523) = 5.94, p < .005$. Post-hoc tests (Tukey’s HSD) revealed that, very similarly to the attributed agency capacities above, revealed that reading the anthropomorphic description ($M = 37.41, SD = 32.69$) caused significantly higher attributions of agency capacities compared to reading the neutral description ($M = 26.27, SD = 30.94$), $p < .01$, with the mechanistic description condition attributions in the middle ($M = 30.23, SD = 30.72$) and not significantly different from those of the other two conditions. Notably, the main effect of video type and the interaction between description and video also trended toward significance, at $p = 0.095$ and $p = 0.083$, respectively. Trends indicate that the group who both watched the mechanistic video and read the mechanistic description attributed the least experience capacities.
and those who watched the social interaction video and read the anthropomorphic description attributed the most anthropomorphic capacities, relative to all other groups. See Figure 5.

![The Effect of Description and Video on Attributed Experience Capacities](image_url)

*Figure 5. The main effect of description ($p < .005$) attributions of experience capacities.*

**Metaphors**

It is also of theoretical interest to separately consider participants’ probability estimates that they would use certain metaphors in reference to the robot.

**Anthropomorphic Metaphors.** The Cronbach’s Alpha for the Anthropomorphic Metaphors Subscale (5 items) was 0.940, 95% CI [0.932, 0.948] indicating very high internal consistency. A two-way analysis of variance yielded a main effect of video watched on the reported probability that a participant would use anthropomorphic metaphors (e.g. “Teammate,” “Playmate,” “Companion,” “Partner,” “Friend”) in reference to the robot, $F(3, 523) = 4.915, p < .005$. Post-hoc tests (Tukey’s HSD) revealed watching the social interaction video ($M = 54.95, SD = 26.77$) to cause significantly higher reported estimates of the participant’s own anthropomorphic metaphor use compared to watching any of the other three videos, i.e.
compared to the robot solo video ($M = 43.55$, $SD = 31.38$), mechanistic interaction video ($M = 44.93$, $SD = 29.58$), or no video ($M = 42.75$, $SD = 30.64$). See Figure 6.

**Figure 6.** The main effect of video ($p < .005$) on participants’ estimated probabilities of their use of anthropomorphic metaphors to describe the robot.

**Mechanistic Metaphors.** The Cronbach’s Alpha for the Anthropomorphic Metaphors Subscale (5 items) was 0.604, 95% CI [0.549, 0.655], indicating barely acceptable internal consistency. A two-way analysis of variance also yielded a main effect of video watched on the reported probability that a participant would use anthropomorphic metaphors (e.g. “Tool,” “Toy,” “Instrument,” “Plaything,” “Gadget”) in reference to the robot, $F(3, 523) = 3.772$, $p < .05$. Post-hoc tests (Tukey’s HSD) revealed watching the mechanistic interaction video ($M = 71.42$, $SD = 15.73$) caused significantly higher reported estimates of the participant’s own mechanistic metaphor use compared to the robot solo video ($M = 64.98$, $SD = 17.25$), $p < .05$, or the social interaction video ($M = 65.24$, $SD = 16.30$), $p < .05$. No significant differences were
found between responses from participants in the no video condition ($M = 66.30, SD = 18.71$) and those in any other condition. See Figure 7.

![The Effect of Description and Video on Estimated Probability of Mechanistic Metaphor Use](image)

*Figure 7. The main effect of video ($p < .05$) on participants’ estimated probabilities of their use of mechanistic metaphors to describe the robot.*

**Advanced Tool Attributions**

**The Effect of Video.** A two-way analysis of variance yielded a main effect of video condition, $F(523) = 26.39, p < .001$, on the extracted factor Advanced Tool Attributions (see Figure 6). Post-hoc tests (Tukey’s HSD) revealed that the mechanistic interaction video ($M = 0.48, SD = 0.79$) caused significantly more tool attributions compared to the social interaction ($M = -0.41, SD = 0.97$) and robot solo conditions ($M = -0.30, SD = 1.11$) at $p < .001$. The no video condition ($M = 0.22, SD = 0.84$) also caused significantly more advanced tool attributions compared to the robot solo and social interaction conditions at $p < .001$. See Figure 8.

A one-way t-test showed that participants who watched the mechanistic video reliably made more tool attributions compared to the mean, $t(125) = 6.84, p < .001$ (two-tailed), as did
participants in the no video condition, $t(147) = 3.21, p < .005$ (two-tailed). Those who watched the social interaction video also reliably made fewer tool attributions compared to the mean, $t(130) = -4.86, p < .001$ (two-tailed), as did participants in the robot solo video condition, $t(129) = -3.10, p < .005$ (two-tailed).

![The Effect of Description and Video on Advanced Tool Attributions](image)

*Figure 8.* The main effect of video ($p < .001$) on the extracted factor Advanced Tool Attributions.

**Moral Care**

**Elections to Save the Robot in Moral Dilemma Scenarios**

The Moral Dilemma Scenario Subscale (5 items) yielded a Cronbach’s Alpha of 0.726, 95% CI [0.688, 0.761]. No significant main effects or interactions were detected with a two-way ANOVA; however, the interaction of description and video was nearly significant, $F(6, 523) = 1.96, p = 0.069$. Some of these video and description dependencies fall in line with hypotheses: for those who read the mechanistic description, the lowest number of elections to save the robot
occurred in the mechanistic interaction video condition, and the number of elections was higher in the social interaction video condition than in the no video condition. Also resembling the predicted pattern, for those who read the anthropomorphemic description, the number of elections to save the robot was higher in the social video condition than in the robot solo video condition, and in the mechanistic video condition, was higher than those made by participants who had read the mechanistic condition. For those who read the anthropomorphemic description, the highest number of elections occurred in the no video condition. Points that particularly contradict predictions include the high number of elections made by those who read the mechanistic description in the robot solo video condition, the highest for the robot solo video condition, and the high number of elections made by those who read the mechanistic description in the social interaction video condition, also the highest for that condition. See Figure 9.

Figure 9. The near-significant interaction of description and video conditions ($p = 0.069$) on elections to save the robot. Note the adjusted scale.
Choosing the Robot. A one-sample *t*-test revealed that, across conditions, elections to save the robot differed reliably from zero, *t*(534) = 27.77, *p* < .001. All participants elected to save the robot—even at the cost of harm to a human—significantly more than never choosing the robot over the human.

Attributed Experience Capacities and Moral Care

As predicted, there was a moderately strong positive correlation between the experience capacity scores and the moral dilemma elections to save the robot *r*(533) = 0.65, *p* < .001 (two-tailed). So, those who attributed more experience capacities, such as the capacity to feel pleasure, affection, and pain, were more likely to elect to save the robot, even at the cost of slight to severe human harm. There was also a moderately strong positive correlation between total anthropomorphism and elections to save the robot, *r*(535) = 0.66, *p* < .001 (two-tailed). So, those who anthropomorphized *in general* were also more likely to save the robot. There was also a very slight negative correlation between advanced tool attributions and elections to save the robot (not close to significance: *r*(535) = -0.04, *p* = 0.31). So, those who conceived the robot mechanistically were very slightly less likely to elect to save the robot at the cost of slight to severe human harm.

Trust

Expected Capabilities

As discussed in Section III, closely related to concerns about excessive trust of anthropomorphized robots are concerns about excessive and unfounded expectations of robot capabilities (e.g. Duffy, 2003; Sandry, 2015). Participants estimated the probability that the robot
would have certain capabilities, none of which were necessarily warranted to expect at all based on the available information. There was a strong significant positive correlation, $r(535) = 0.707$, $p < .001$, between the amount participants anthropomorphized and the extent to which they expected the robot to have (total anthropomorphic and mechanistic) capabilities they had no real justification to expect.

**Expected Anthropomorphic Capabilities.** The Expected Anthropomorphic Capabilities Subscale (8 items) had a Cronbach’s Alpha of 0.921, 95% CI [0.911, 0.931], indicating very high internal consistency. A two-way analysis of variance yielded significant main effects of description, $F(2, 523) = 3.40$, $p < .05$, and video, $F(3, 523) = 4.63$, $p < .005$, conditions on participants’ probability estimates that the robot would have certain anthropomorphic capabilities (e.g. “Comfort me,” “Give me advice,” or “Joke around with me”). There was also a statistically significant interaction between description type and video watched, $F(6, 523) = 2.189$, $p < .05$.

**The Interaction of Description and Video Conditions.** Table 1 shows the significant or near-significant results of the analysis of simple effects. Most notably, those who read the anthropomorphic description had higher expectations, with the exception of the group who also watched the robot solo video. Additionally, those who watched the social interaction video generally had higher expectations relative to other video conditions, with the exception of the no video condition group who also read the
anthropomorphic description. Also, for those who read the mechanistic description, those who also watched the mechanistic interaction video had significantly lower expectations compared to those who watched the social interaction video (see Figure 10). So, the anthropomorphic description and social interaction generally caused higher expectations of anthropomorphic capabilities, and the combination of the mechanistic description and the mechanistic interaction video lowered expectations of anthropomorphic capabilities.

Figure 10. The significant main effects of description ($p < .05$) and video ($p < .005$) conditions and the interaction of description and video conditions ($p < .05$) on participants’ expectations of anthropomorphic capabilities.

**Anthropomorphism and Anthropomorphic Capability Expectations.** There was a significant strong positive correlation, $r(535) = .83$, $p < .001$, between anthropomorphism as measured by the extracted factor and participants’ estimated probability that the robot had certain anthropomorphic capabilities.
**Expected Mechanistic Capabilities.** The Expected Mechanistic Capabilities Subscale (8 items) had a Cronbach’s Alpha of 0.893, 95% CI [0.879, 0.907], indicating high internal consistency.

**The Effect of Description.** A two-way analysis of variance yielded a significant main effect of description type on participants’ probability estimates that the robot would have certain mechanistic capabilities (e.g. “Record sounds,” “Access my music library for me,” or “Scan things using thermal sensing”), $F(2, 523) = 3.40, p < .05$. Post-hoc tests revealed that expectations of those who read the mechanistic description ($M = 64.03, SD = 20.42$) were significantly higher than those of the participants who read the neutral description ($M = 58.24, SD = 23.50$) at $p < .05$. Expectations of those who read the anthropomorphic description ($M = 60.52, SD = 23.35$) did not significantly differ from those of the other two description conditions. See Figure 11.

**The Effect of Video.** There was also a significant main effect of video type, $F(3, 523) = 8.018, p < .001$. Post-hoc tests revealed that those who watched the mechanistic video ($M = 64.14, SD = 18.77$) had significantly higher expectations of mechanistic capabilities compared both to those who watched the robot solo video ($M = 55.81, SD = 25.76$), $p < .05$, and those who watched the social interaction video ($M = 56.24, SD = 25.50$), $p < .05$. Those who watched no video ($M = 66.48, SD = 17.74$) also had significantly higher expectations compared both to those who watched the robot solo video ($p < .001$) and those who watched the social interaction video ($p < .01$). See Figure 11.
Figure 11. The main effects of description ($p < .05$) and video ($p < .001$) conditions on participants’ expectations of anthropomorphic capabilities.

**Anthropomorphism and Mechanistic Capability Expectations.** There was a significant moderately weak correlation, $r(535) = .39, p < .001$, between anthropomorphism as measured by the extracted factor and participants’ estimated probability that the robot had certain mechanistic capabilities.

**Physical and Psychological Task Scenarios**

**Trust in Physical Tasks Scenarios.** The Physical Task Trust Scenario Subscale (5 items) yielded a Cronbach’s Alpha of 0.775, 95% CI [0.743, 0.804], indicating an acceptable level of internal consistency. A two-way analysis of variance yielded no significant main effects or interactions of description and video conditions. In terms of trends, the anthropomorphic description condition ($M = 73.16, SD = 20.60$) seems associated with very slightly more trust than participants in neutral ($M = 72.34, SD = 20.75$) mechanistic ($M = 70.05, SD = 22.93$)
description conditions, and the mechanistic description seems associated with very slightly less trust than the other two conditions. This pattern is pronounced most clearly in the no video (just description) and robot solo video conditions. As predicted, within the condition who read the anthropomorphic description, trust was lowest for those who next watched the mechanistic interaction video. Surprisingly, within the condition who read the mechanistic description, trust was highest for those who next watched the mechanistic interaction video. See Figure 12.

![The Effect of Description and Video on Trust in Physical Task Scenarios](image)

*Figure 12.* The influence of descriptions and videos on trust in physical task scenarios. Note the changed scale.

**Trust in Psychological Tasks Scenarios.** The Psychological Task Trust Scenario Subscale (5 items) yielded a Cronbach’s Alpha of 0.716, 95% CI [0.676, 0.752], indicating an acceptable level of internal consistency. A two-way analysis of variance yielded no significant main effects or interactions of description and video conditions; however, a main effect of description was close to significantly significant, $F(2, 523) = 0.075, p = 0.069$. Generally, those who read the anthropomorphic description ($M = 60.40, SD = 22.12$) had higher trust than those who read the neutral description ($M = 58.86, SD = 21.45$), who had higher trust than those who
read the mechanistic description \( M = 55.15, SD = 20.56 \), regardless of video condition, trending in accordance with the top-down effects hypothesis. However, the point that breaks the pattern, which might explain why the other clear patterns are statistically insignificant, is the group that both read the anthropomorphic description and saw the social interaction video, which had less trust than any other anthropomorphic description group, and less trust than any other social interaction video group. The other interaction-type predictions were generally supported: the group who read the mechanistic description had the lowest trust, and the group who read the anthropomorphic description had the highest trust, in the no video, robot solo, and mechanistic interaction video conditions, with the neutral description group being somewhere in the middle. Analysis of simple effects revealed no significant differences between the neutral description groups and any other group, but that the difference between the anthropomorphic description group and the mechanistic description group was significant in the no video condition \( (p < .05) \), and close to significant in the mechanistic interaction video condition \( (p = 0.056) \). See Figures 13 and 14.

**Figure 13.** The influence of descriptions and videos on trust in psychological task scenarios. Note the changed scale.
Comparing Trust in the Physical and Psychological Scenarios. Supporting the related prediction and the Matthews et al. (2019) finding related to higher trust in robots in physical compared to psychological task scenarios, a paired-samples T-test revealed that participants generally trusted the robot more in physical ($M = 71.87, SD = 21.47$) compared to psychological ($M = 58.16, SD = 21.43$) task scenarios, $t(534) = 14.573, p < .001$. There was also a moderate correlation between the average scores of the two subscales, $r(533) = 0.49, p < .001$.

Trust Delta. Trust scores were summed across items for one total trust score for each participant and then subtracted from the overall mean trust of all participants to find how individual scores differed from the mean. Though neither the anthropomorphic nor the mechanistic descriptions differed reliably from the mean, the trend is in accordance with the Mechanistic Linguistic Framing Hypothesis, the Anthropomorphic Linguistic Framing Hypothesis, And The Top-Down and Bottom-Up Interaction Hypothesis: generally less trust in
the mechanistic description conditions, generally more trust in the anthropomorphic description condition, and generally more trust in the video conditions with more bottom-up anthropomorphic cues (see Figure 15).

![The Effect of Description and Video on Total Trust](image)

*Figure 15. The influence of description and video on total trust. Note the changed scale.*

**Anthropomorphism and Trust.** There was a weak but significant positive correlation between anthropomorphism and the extracted trust factor, $r(533) = 0.175, p < .001$.

**DISCUSSION**

Overall, the hypothesis that top-down framing influences anthropomorphism is supported by the results.

**Descriptions Matter: The Anthropomorphic Linguistic Hypothesis**

Most importantly, anthropomorphic descriptions increased explicit anthropomorphic attributions measured through the extracted Anthropomorphism factor, as well as experience and
agency capacities attributed to the robot, compared to both other descriptions. The anthropomorphic description also increased anthropomorphic attributions on the implicit linguistic level, increasing the total number of linguistic anthropomorphic markers used and decreasing the number of impersonal pronouns used in reference to the robot compared to both other descriptions. The anthropomorphic description also generally caused participants to have higher expectations of anthropomorphic capabilities than were warranted by any information they had seen about the robot really warranted, including “Understand my emotions” and “Give me advice.” And, though not statistically significant, there was a clear trend by which the anthropomorphic description was generally associated with the highest amount of trust in both the physical task and psychological task scenarios. These results support the Anthropomorphic Linguistic Hypothesis, as the anthropomorphic description seems to have caused more anthropomorphism compared to a neutral description (and, generally, a mechanistic description). These effects occurred regardless of video watched next, suggesting the strength of this top-down effect on anthropomorphism.

A Sophistication Bias?

Overall, the Mechanistic Linguistic Hypothesis was not supported by these results, as the mechanistic description did not reliably cause less anthropomorphism compared to the neutral description. That said, it did cause significantly higher expectations of certain expected mechanistic capabilities that the participant had no real reason to expect from the given information, which ranged from “Record sounds” to “Access my music library for me” to “Scan things using thermal sensing” to “Scan things using X-ray vision.” It was indeed an unexpected result that the mechanistic description was associated with numerically (insignificantly) more
anthropomorphism (and similarly with attributed agency and experience capacities) than the neutral description in the no video and social interaction video conditions, and more than even the anthropomorphic description in the robot solo video condition. It might have been the case that the list of equipment and capabilities that constituted the mechanistic description gave the impression of an *advanced* or *sophisticated* robot, and that participants had some weak association between technical sophistication and anthropomorphic capacities.

It also seems the case that the mechanistic description decreased anthropomorphism *when combined with the mechanistic interaction video*: see my discussion of what I call a possible “congruence effect” to follow.

**Agency and Experience: A Look Back at Gray et al. (2007)**

Indeed, the effect of description was the only significant influence on the attributed agency and experience capacities. Especially with this result in mind, it is all the more interesting to compare the average agency and experience scores from participant groups in the present study with the agency and experience scores of the robot and adult human cluster in the original Gray et al. (2007) paper. Figure 16 shows this comparison: as depicted, it is clear that the present manipulations, especially the descriptions, caused participants to view the robot in a way that moves the cluster representing this robot closer to the adult human cluster in the top right (indicated with the large gray pentagon), compared to the robot in the Gray et al. (2007) study at the bottom (indicated with an “x”). Especially notable is the point closest to the adult human cluster—representing the group who both read the anthropomorphic description and watched the social interaction video—and the point lagging behind, lowest on the experience scale,
representing the group who both read the mechanistic description and watched the mechanistic interaction video.

![Agency and Experience](image)

*Figure 16.* A comparison of the average agency and experience scores from participant groups in the present study with the agency and experience scores of the robot (indicated with an “x”) and adult human cluster (indicated with the gray pentagon) in the original Gray et al. (2007) paper.

**Top-down versus Bottom-Up**

The prediction regarding bottom-up cues, backed by the current prevalent understanding in the literature that anthropomorphism is driven by bottom-up cues, was that exposure to the robot’s anthropomorphic cues would increase anthropomorphism, and perhaps trust and moral regard. This hypothesis was generally supported: there were significant main effects of both description and video conditions on participants’ anthropomorphism as measured by the
extracted factor. That is, the ultimate degree of anthropomorphism for each group depended on both the description they read and the video they watched. Moreover, there was generally the predicted trend of more anthropomorphism across video conditions as more anthropomorphic cues were added.

**The Robot Solo Video.** The one exception to this trend is the robot solo video condition, which features some of the robot’s bottom-up cues (e.g. its eyes, contingent interaction with the environment), than in the no video condition, in which participants have no exposure to the robot’s bottom-up cues. Anthropomorphism did not significantly differ between the robot solo video condition and the no video condition; moreover, for the anthropomorphic and neutral descriptions, anthropomorphism was lower in the robot solo video than in the no video condition. Only within the mechanistic description condition showed the predicted pattern of more explicit anthropomorphic attributions in the robot solo than the no video condition. This same pattern appeared also for expected anthropomorphic capabilities and moral care. Trust in physical task scenarios was lower across description conditions in the robot solo video compared to the no video condition. Trust in psychological task scenarios was again lower within the anthropomorphic description condition in the robot solo compared to the no video condition, though slightly higher within the mechanistic and neutral description conditions in the robot solo compared to the no video condition. For expected mechanistic capabilities of the robot, too, participants across description conditions indicated lower expectations who saw the robot solo video compared to the no video condition. Notably, however, the robot solo video did significantly decrease attributions of the robot as an advanced tool, compared to the no video description; while not *promoting anthropomorphism*, it did seem to decrease engagement in a main contrasting alternative to anthropomorphism. Generally, despite predictions backed by the
established literature that the presence of the bottom-up cues featured on the robot Vector should increase anthropomorphic attributions, and the robot solo video featured these cues and the no video condition did not, the robot solo video did not seem to increase anthropomorphic attributions; indeed, it even seemed at times (especially for those who did not read the mechanistic description) to have a slightly negative effect.

There are a few potential explanations for this unanticipated pattern in the results. It is first important to note that this pattern usually (with the exception of trust in physical tasks and, interestingly, mechanistic capabilities) did not obtain for those who read the mechanistic description: usually, there was slightly more anthropomorphism for those who read the mechanistic description in the presence of bottom-up cues, in accordance with the literature. One possible explanation, then is that the descriptions set lofty expectations for the robot such that watching (specifically) the robot video after those initially high expectations were set led to something like disappointment, which caused not only less anthropomorphism (for those who read the anthropomorphic and neutral descriptions) but also less conceptualization of the robot as adequate even as a tool or gadget (for all). Those who never saw the robot would have never encountered this disappointment.

However, if this explanation were true, the effect should not exist for those who read the neutral description, as the neutral description described only the robot’s dimensions and therefore could not have had this lofty expectations effect. Because the general effect was seen as often for the neutral description as it was the anthropomorphic description, this lofty expectations for why there was less anthropomorphism cannot be the case. Perhaps, however, there was a mechanistic expectations effect of the mechanistic description. Perhaps participants expected a tool-like machine, and when they were exposed to the bottom-up cues in the robot solo video
(compared to no exposure to those cues in the no video condition), they anthropomorphized more than they would have if their expectations had not been set with mechanistic language. The exceptions to the more-cues-more-anthropomorphism pattern within the mechanistic description condition occurred in for trust in physical scenarios and mechanistic capabilities; a lofty expectations explanation localized to just the mechanistic description condition accounts for these exceptions well. The mechanistic description, while it set low expectations with regard to anthropomorphism, set lofty expectations for mechanistic capabilities: seeing the robot’s actual form and performance in the robot solo video (compared to not seeing them at all) must have caused a disappointment effect that caused lower expectations of (unfounded) mechanistic capabilities and lower trust in tasks in which one might otherwise use an advanced tool.

To account for the effect seen in the other two description conditions, an alternative explanation to the lofty expectations explanation is that the video itself, featuring the Anki Vector interacting with its environment by itself, might have somehow not been as anthropomorphic from the bottom-up perspective as it could have been, or lowered participants’ opinion of this specific robot. Anecdotally, in the open response descriptions, some participants wrote that the robot had tried to place the cube it moved during the video in a certain location (on the similar-looking charger), instead of next to it (as it did), and failed. This unanticipated perception of the robot’s behavior as having failed might be associated with a perception of lower competence, which those who did not watch that video might not have had. A perception of low competence would conceptually explain the slightly overall lower expectations of both anthropomorphic and mechanistic capabilities, and slightly lowered physical trust, relative to the no video description; perhaps this result would have been even more pronounced if not for the anthropomorphic cues—that is, for a robot with fewer bottom-up anthropomorphic cues, perhaps
any perceived lack of competence would have led to even lower expectations and less trust. Simultaneously, without this unanticipated perception of incompetence, perhaps any effect of the bottom-up anthropomorphic cues might have been more pronounced. Additionally, it may be important that the robot faced away from the camera for some of the video (although its eyes could be seen in the reflection provided by a nearby window); it did so, for example, for more of the robot solo video than the robot did during either of the interaction videos, when it was mostly facing the human who was slightly in front of the camera. Because the robot’s “eyes” is one of its strongest bottom-up anthropomorphic cues, it is possible that the robot solo video was thus not as effective of a stimulus portraying the robot’s anthropomorphic cues as it could have been.

The comparative consideration relative to the two interaction videos, in addition to the recognition that the robot solo video lacked the additional anthropomorphic cue of reciprocal interaction with a human (although it exclusively included the robot’s contingent interactions with its environment), it is likely the case that the robot solo video generally seemed less anthropomorphic from the bottom-up perspective than the two interaction videos, as suggested by its associated significantly higher impersonal pronouns and significantly lower expected anthropomorphic capabilities and attributed agency capacities.

So, while not in the anticipated way, the very slight and statistically insignificant pattern in the results by which the bottom-up effect of the video condition depended on the description read first, trends in accordance with the Top-Down and Bottom-Up Interaction Hypothesis. Overall, however, these results support the conclusion that the descriptions, compared to the difference in the presence and absence of the bottom-up cues in the robot solo and no video conditions (respectively), had the stronger effect on anthropomorphism.
The interaction videos, on the other hand, possessed an established bottom-up cue the robot video did not have: contingent interaction with a human (e.g. Johnson, 2003; Meltzoff et al., 2010). In addition to the increased video footage of the robot’s “eyes,” a result of the robot typically facing the human who was between the robot and the camera, the interaction videos had two more bottom-up anthropomorphic cues. In accordance with the literature, then, and the idea that more anthropomorphic cues causes more anthropomorphism (e.g. Johnson, 2003; Damiano & Dumouchel, 2018) there was generally more anthropomorphism (with the interesting typical exception of those who read the mechanistic description first) in the interaction videos than the no video and robot solo video conditions (with the interesting exception of those who read the mechanistic description first). This general (statistically insignificant) pattern also occurred for the measures of expectations of anthropomorphic capabilities and trust. These results trend in accordance with the Top-Down and Bottom-Up Interaction Hypothesis, as both description and video conditions influenced anthropomorphism of the featured robot; that said, the interaction videos did not cause more anthropomorphism to a significant extent, and the descriptions did influence anthropomorphism to a significant extent. These results support the conclusion that the top-down influence of the descriptions, compared to the bottom-up cues featured in the videos, had the stronger effect on anthropomorphism.

**The Interaction Videos.** Considering the importance of these two highly anthropomorphic cues from the bottom-up perspective—eyes and its reciprocal interaction with a human—it should be the case, then, that the two interaction videos were generally equally anthropomorphic. However, the results do not support this idea: the mechanistic video seems to decrease anthropomorphism relative to the no video and robot solo conditions. This pattern is evidenced by the significant lower number of total linguistic anthropomorphic markers, and the
significantly higher number of mechanistic expectations, and further suggested by numerically lower anthropomorphism (for those who first read the mechanistic description), attributions of experience capacities (for those who first read the mechanistic description), and slightly fewer elections to save the robot (for those who first read the mechanistic description). The social interaction video, by comparison, significantly increased the use of linguistic anthropomorphic markers relative to all other video conditions, slightly numerically increased explicit anthropomorphism, including attributed experience capacities, relative to all the other video conditions, and was associated with slightly higher expectations of anthropomorphic capabilities (especially for those who also read the anthropomorphic description).

These trends could be explained with two main alternative hypotheses; these alternative hypotheses are not necessarily competing, and both may be at least partially true. The first, which will be called the Importance of Complete Physical Autonomy Hypothesis, relates to a difference in one more known bottom-up anthropomorphic cue: apparent autonomy. The second hypothesis, which will be called Top-Down Influence of Human-Robot Interaction Hypothesis, and which is particularly interesting with regard to the main hypothesis of this thesis, explains this pattern in terms of possible top-down influences in the interaction videos, and can account for what will be called the “congruence effect,” by which those who read both the mechanistic description and watched the mechanistic video and those who read both the anthropomorphic description and watched the social interaction video exhibited at times stronger effects.

The Importance of Complete Physical Autonomy Hypothesis. The first hypothesis relates to a difference in one more known bottom-up anthropomorphic cue: apparent autonomy or independence. Though in both videos the robot was told to complete certain tasks (e.g. move the cube, give a fist bump, convert a measurement), in the social interaction video, these
instructions were given without touching or otherwise manipulating the robot, while in the mechanistic interaction video, these instructions were given after manually manipulating the robot by pushing the button on its top. This difference may have made a difference from the bottom-up perspective. If this explanation is correct, it would mean that physical autonomy, i.e. not being physically touched by another human before manipulative interaction, is a critically important bottom-up anthropomorphic cue, if its absence—even in the face of a suite of other anthropomorphic cues, including eyes, quality of movement, contingent behavior, natural language capability (both understanding and generating human language)—causes such a pronounced decrease in anthropomorphism (especially as measured by total linguistic markers, and suggested by higher expectations of mechanistic capabilities and advanced tool attribution scores). The mechanistic video had two more bottom-up anthropomorphic than the robot solo video (contingent interaction with a human and more footage of its eyes), but one fewer (complete physical autonomy); if the possible complete physical autonomy was that important (more important than the other two cues, or any cues at all), it would explain why the mechanistic interaction video led to less anthropomorphism as measured by some scales. Meanwhile, the social interaction video had all of these cues, and so led to more anthropomorphism as measured on some of these scales.

That said, given that touching the robot (removing its “complete physical autonomy” as conceived above) occurred infrequently—only before the human spoke to the robot—and given the intuition that humans frequently use tactile communication (e.g. touching someone to get their attention before you speak to them), in conjunction with recognition of the demonstrated strength in the literature of cues that were present in both interaction videos, this hypothesis
seems to make an excessively strong claim. It therefore seems more likely that the Top-Down Influence of Human-Robot Interaction Hypothesis is the better explanation.

**Possible Top-Down Influences in the Interaction Videos: The Top-Down Influence of Human-Robot Interaction Hypothesis.** The interaction videos (mechanistic and social) were conceived with the theoretical motivation that they featured the strong bottom-up anthropomorphic cue of reciprocal interaction with a human (which is not found in the robot solo video), and that watching the robot interact reciprocally (and intelligently and emotionally) with a human might cue anthropomorphism in an automatic, bottom-up way. That said, in addition to the robot’s behavior during the interaction, which seems to act in a bottom-up way, there is also the human’s behavior during the interaction, and how the human seems to consider the robot—for example, as if it were a tool, or a companion. The human’s behavior includes its treatment of the robot and any language used during that interaction, which together could give conceptual information of how the other human seems to consider the robot. The observer may cognitively process this information in such a way that the processed information may influence subsequent judgements about the robot in a top-down way. So, it seems possible that the social video not only features the additional strong bottom-up cue to anthropomorphize in the way that the robot interacts reciprocally to the human, but also could be conceived as featuring some top-down cues to anthropomorphize, such as the human’s language (including specific metaphorical words such as “bud” “team”) and how the human seems to conceptualize the robot as a companion. Likewise, the mechanistic video not only features the robot’s contingent and intelligent interactions to the human, but also features the human’s apparent conceptualization of the robot as a tool, using it for specific ends that are completely unrelated to the human’s interaction with the robot.
Supporting this hypothesis, there is strong evidence that the mechanistic video served to decrease anthropomorphism, even when the mechanistic description did not. For example, watching a human interact with the robot mechanistically caused participants to use significantly fewer total linguistic anthropomorphic markers in their own descriptions of what happened to the video, compared to the anthropomorphic and solo videos, regardless of what description they first read. The mechanistic video was also associated with significantly more explicit advanced tool attributions, including higher estimated probabilities that participants would use mechanistic metaphors such as “tool” to refer to the robot, relative to the other two video conditions—even when the mechanistic description did not have that effect. This particular result indicates that it might be the case that watching another human apparently conceptualize a robot in a certain tool-like role causes participants to associate the robot with a tool, influencing subsequent judgements in a top-down way. Watching the mechanistic video also caused significantly higher expectations of mechanistic capabilities beyond those that were described by information the participants had seen so far, compared to the other video conditions.

There is also strong evidence that the social interaction video increased anthropomorphism, possibly due to top-down effects. The social interaction video increased overall anthropomorphism (significantly, compared to the mechanistic interaction video; generally, compared to the robot solo and no video conditions). The social interaction video also caused significantly higher reported probabilities that participants would use anthropomorphic metaphors—even when the anthropomorphic description did not have that effect. It might be the case that participants could more easily explicitly conceptualize the degree to which a robot could fill a typically human social role, such as “teammate” or “playmate” or “companion,” when we see another human treating the robot as if it filled that role.
The social psychology involved in the effect of watching another human interact with a robot in a specific metaphor-laden way—that is, as a social companion, or as a tool—on one’s perception of that robot (and, further, of the human) seems like an interesting and possibly fruitful area of further research. It seems like a triple theory of mind problem: the participant’s understanding of the human user’s understanding of the robot’s degree of understanding may influence the participant’s understanding of the robot’s degree of understanding. According to results of the present study, it does seem highly plausible that that our conceptual impression and understanding of how the other human conceptualizes the robot influences how we conceptualize that robot— influencing the metaphors we explicitly associate with the robot, and influencing the anthropomorphic capacities we explicitly attribute.

**Doubling Up: The Congruence Effect.** An advantage of the Top-Down Interaction Influences Hypothesis is that it can readily account for a “congruence effect” pattern in the data. In terms of trends, there is some evidence supporting the idea that the combination of the mechanistic description and the mechanistic video might serve to decrease anthropomorphism of an otherwise anthropomorphic robot. For example, the group who both read the mechanistic description and watched the mechanistic interaction video anthropomorphized the least of all conditions. When plotted in the style of Gray et al. (2007), this group had one of the points lagging behind the most, intermediate on the Agency scale but lowest on the Experience scale. For those who watched the mechanistic video, those who read the mechanistic description (compared with those who read the anthropomorphic or neutral descriptions) had significantly lower expectations of anthropomorphic capabilities. This mechanistic-mechanistic group also had the highest physical trust (though not to a statistically significant extent): that is, bottom-up anthropomorphic cues *in the context of an interaction that used metaphors compatible with those*
read in the initial description (mechanistic, tool-based metaphors) boosted trust for physical domain tasks (those in which one might typically use a tool). Additionally, though not statistically significant, it is notable this group also had low attributions of experience capacities, probability estimates that they would use anthropomorphic metaphors, trust in psychological task scenarios, and moral care.

There is also some evidence that combination of the anthropomorphic description and the social interaction video can serve to increase anthropomorphism. For example, in terms of trends, this group had the highest anthropomorphism scores and used the most linguistic anthropomorphic markers of all conditions, led the trend toward the adult human cluster in comparison to the Gray et al. (2007) results with the highest attributions of experience capacities, and also had the highest (though just barely) expectations of anthropomorphic capabilities and estimated probable use of anthropomorphic metaphors.

So, it seems there was some trend by which congruence in the metaphorical conceptual cues in the descriptions and videos can serve to increase the effect of that metaphorical cue. Effectively, the congruence effect refers to the slight pattern in the data that might be explained by additional and congruent exposure to a given metaphor, which might work to cue knowledge structures of (for example) a tool or human in the inductive inference process of anthropomorphism to a stronger degree. That is, reading a description that compares a robot to a humanlike companion through language and watching another human treat the robot as if it were their humanlike social companion will make it more likely that the humanlike “companion” metaphor will influence (increase) anthropomorphism.

Likewise, reading a description that compares a robot to a tool through language and watching another human treat the robot with an apparently similar conceptual understanding of
the robot as a tool—treating it as if it were a tool—will make it more likely that the “tool” metaphor will influence (decrease) anthropomorphism. This trend is evidence for a possible top-down effect of the mechanistic description: though reading the mechanistic description by itself did not reliably decrease anthropomorphism relative to a neutral description, it may have set expectations or elicited a tool knowledge structure such that when presented with further evidence of another human who seems to regard the robot in a way congruent with those expectations or elicited knowledge structure, it decreased the likelihood of anthropomorphism. So, while not in the way predicted, this congruence effect trend may show how mechanistic descriptions may have a top-down effect in decreasing anthropomorphism.

This hypothesis assumes the presence of the plausible top-down influence of these interaction videos, whereby information about how another human seems to understand the robot is processed as conceptual information, which may influence subsequent judgements in a top-down way. In addition to the clear significant main effects of descriptions, this slight pattern may be in this way additional evidence supporting the general Top-Down Effects hypothesis—that top-down influences impact anthropomorphism—and the idea that metaphors matter.

The Associated Effects of Anthropomorphism

Great Expectations. There was a significant interaction underlying a trend by which those who read the anthropomorphic description or saw the social interaction video had higher expectations of anthropomorphic capabilities (e.g. “Understand my emotions” or “Give me advice”) beyond those reported by the given information, and those who read the mechanistic description and saw the mechanistic interaction video had lower expectations of anthropomorphic capabilities. There was also a significant strong positive correlation between
these (unwarranted) expectations of anthropomorphic capabilities and anthropomorphism as measured by the extracted factor, as well as a moderately weak correlation between (unwarranted) expectations of mechanistic capabilities and anthropomorphism as measured by the extracted factor. Importantly, these results supporting the concern that anthropomorphic cues might lead to unwarranted expectations of capabilities of robots, especially those related to interpersonal relationships (e.g. Scheutz, 2012; Sandry 2015).

There was also an effect by which those who read the mechanistic description and those who watched the mechanistic interaction video had higher expectations of mechanistic capabilities (“Access my music library” to “Scan things using X-ray vision”). This result might be an important implication of automation bias: robots perceived as “advanced” through a sophisticated-sounding list of the capabilities it does have (as featured in the mechanistic description) may lead to the expectation of other, even more advanced capabilities.

**Trust.** The results in the present study aligned with the Matthews et al. (2019) finding that generally, people seem to trust robots more in physical task scenarios compared to psychological task scenarios. As suggested in Section III during the discussion of that finding, it might be the case that there is some preexisting conceptual bias in participants’ knowledge structures about robots associating them with *tools*, better suited for physical tasks, or, related to our metastructural expectations about differences between causal structures in the physical and psychological domains, as entities only capable of processing simple linear deterministic causal chains (associated with the physical domain), rather than complex, multiple, linear causal structures (associated with the psychological domain) (see Section III). The former part of this explanation is supported by the high number of advanced tool attributions in the no video condition, comparable with the mechanistic video condition and significantly higher than both
the robot solo and social interaction video: it seems that participants, regardless of what
description they read, came in with strong conceptions of the robot as an advanced tool.

Though there were no statistically significant effects of description or video type on trust
in the physical and psychological task scenarios, the general pattern indicated a trend consistent
with the broad hypothesis of the present study: trends indicated that the anthropomorphic
description was associated with greater trust compared to those who read the neutral or
mechanistic description, and the mechanistic description was generally associated with less trust
than the other two descriptions (especially for the no video and robot solo conditions). Overall,
the patterns indicate that descriptions, rather bottom-up influences from the video conditions, had
the more important effect on trust in both physical and psychological task scenarios. With regard
to metastructural expectations, this result suggests that top-down anthropomorphic metaphor-
laden language can influence our understanding of the degree to which robots can competently
handle reasoning about both simple, deterministic, linear causal chains of the physical domain as
well as complex, multiple, nondeterministic causal structures of the psychological domain—as a
human can—influencing us to consider the robot as more capable of both types of judgment.
This may work in the proposed way, by biasing our inductive inferences about the robot’s
behavior and capacities even more toward our knowledge structures about humans.

**Physical Trust.** It is likely common in scenarios such as the presented physical task
scenarios—e.g. chemical detection in a security scenario, a combination of timing and
temperature estimates in a cooking scenario, body heat detection in a law enforcement
scenario—one would typically use a tool to aid detections of the relevant variables of interest
(the presence of certain chemicals, food readiness, the presence and location of body heat). Thus,
it seems likely that our knowledge structures about tools would be activated and integrated—just
by the task context. At the same time, these situations also involve judgements based on these
detections, and recommendations based on those judgements—and, of course, the attribution of
capacities such as the capacity to think and make decisions are (agency) anthropomorphic
attributions. So, our knowledge structures about humans are also likely to be activated and
integrated. It is thus interesting to see the effect of the anthropomorphic description, especially in
the robot solo and no video conditions, but also the combination of the mechanistic description
and mechanistic video, in slightly increasing physical trust judgements. If the proposed cognitive
mechanism involving the activation and integration of knowledge structures in a process of
inductive inference is correct, the possibility that both tool-related knowledge structures and
humanness-related knowledge structures are being activated by descriptions, top-down
influences in the videos, preexisting conceptualizations of what a robot is and what it can handle,
and the hypothetical task environment itself—this competition might help to explain why results
for trust in physical scenarios was relatively mixed.

_Psychological Trust._ Generally, those who read the anthropomorphic description had
higher trust than those who read the neutral description, who had higher trust than those who
read the mechanistic description, regardless of video condition, in accordance with the
hypothesis. However, the point that breaks the pattern to such an extent that it may partially
explain why the other clear patterns are statistically insignificant, is the group that both read the
anthropomorphic description and saw the social interaction video—which, against all
predictions, had less trust than any other anthropomorphic description group, and less trust than
any other social interaction video group. It is a theoretical possibility that this group hit the
Uncanny Valley.\(^\text{25}\) In terms of the predicted cognitive mechanism, and assuming the social

\(^{25}\) As more humanlike appearance cues are added, humans generally tend to feel more positive affect toward them—to a point. After this point, we tend to feel negative affect and a feeling of strangeness or eeriness, a phenomenon
interaction video operates a metaphor-laden top-down influence on subsequent judgements, and that (as described above for the physical task scenarios) the task scenario itself activated in additional strength or number humanness-related knowledge structures, there may have been so much activation of humanness knowledge structures that participants might have experienced eeriness and negative affect toward the robot, and this negative affect may have been related to the significantly decreased trust in making psychological judgements. To test this possibility in a future study, participants’ affect toward the robot should be measured during or immediately after their consideration of psychological task scenarios.26

**Moral Care.** There was no clear effect of description or video or trending pattern in participants’ elections to save the robot at the cost of slight to severe harm to humans. It is interesting, though, that participants *ever* chose to save the robot at the cost of human harm, and elections to save the robot was significantly higher than zero across conditions. Additionally, those who attributed more experience capacities, such as the capacity to feel pleasure, affection, and pain, were more likely to elect to save the robot. This result provides additional support for the hypothesis that moral care considerations—treatment of an entity as a moral patient, worthy of certain rights to prevent its suffering—are based on attributions of experience capacities (e.g. Waytz et al., 2010).

that Mori (1970) originally proposed and called the *uncanny valley*, the “valley” referring to the dramatic dip in positive affect with the addition of more humanlike appearance. Despite receiving very little attention at the time of publication, the uncanny valley hypothesis has received a surge of attention in the past twenty years, as computer-graphics animation and robotics design have advanced. There has been empirical support for the theory: for example, MacDorman & Ishiguro (2006) presented participants with morphed versions of photographs of humans and robots and found a peak in eeriness in the middle of the series corresponding to the hypothesized valley. Another study found that robots with a clearly machine-like appearance (but with some humanlike features including the presence of eyes) are more liked than robots with humanlike appearances that are highly realistic but not completely realistic (Bartneck, Kanda, Ishiguro, & Hagita, 2007). Along similar lines, another study suggested that the *mix* of robot and human features may be the cause of eeriness (Ho, MacDorman, & Pramono, 2008).

26 Critically, the psychological task scenarios might be the final top-down influence promoting use of humanlike knowledge structures, so measuring affect toward the robot before or after a sufficiently long period of time would theoretically not be able to capture the possible associated negative affect.
A Point on Measures of Explicit Anthropomorphic Attributions

Exploratory factor analysis revealed that the items across scales designed to measure anthropomorphism (metaphors, agency capacities, experience capacities, and the pairwise comparisons subscale) yielded one factor that seems to measure anthropomorphism. Within this factor, experience capacities (feeling pain, pleasure, and affection, and experiencing emotions) and the agency capacity item related to understanding the emotions of others loaded most highly. Meanwhile, the agency capacity thinking loaded highly, making decisions loaded moderately, and the fourth (communicating) did not load with the anthropomorphism factor. These factor extraction results are compatible with the theoretical and empirical work suggesting that experience capacities are perhaps more intuitively central to our conception of humanness than agency capacities (see Section II). Additionally, the partner, friend, teammate, and companion metaphors also loaded very highly (with playmate also loading moderately highly). This result supports the theoretical considerations developed in Section IV that metaphors are embedded in our conceptual understanding of robots, and that metaphors involving comparison of robots to typically human roles corresponds with our anthropomorphism of them. There is the suggestion, as well, that increased anthropomorphism might promote use of such anthropomorphic metaphors going forward, which might in turn (as the anthropomorphic description did and language in the social interaction video may have) increase the likelihood that those who hear them using such metaphors may anthropomorphize more as well.

Overall, the exploratory factor analysis conducted in this study suggested that measures of explicitly attributed experience capacities and—to a lesser extent, but especially those concerned with cognitively recognizing and understanding experience capacities in others—agency
capacities, metaphors, as well as the artificial/living, machinelike/humanlike, and nonconscious/conscious pairwise comparison scales, might be productively combined for a cohesive measure of anthropomorphism.

**Limitations**

There are several limitations to this study, including considerations about the extent to which the study sample is representative of the general population, typical concerns of online studies, the potential for a reduced effect of bottom-up cues due to their occurrence in a video rather than in in-person interactions with the participant, concerns about the robot solo video, and concerns of social desirability bias.

**Demographics.** Participants in this study were relatively homogenous (62.9% identified male, 73.7% identified White or Caucasian) and all were in the United States. It is unclear how these results would generalize for other populations.

**Normal limitations of online studies.** As is typical for online studies, even with strenuous attempts to control for non-human bots and for non-attentive humans with a ReCaptcha check, the open-ended description requirement, and other attention checks, there is no way to ensure that participants were all actually human, all applying equal attention to the description and video stimuli, all reading the question items or scenarios in full, or all taking equal care with their responses.

**The Difference of Embodiment: Videos versus In-Person Interaction.** For this study, there may be a particular benefit to having participants interact with a robot *in person* as the bottom-up cue manipulation. Preliminary research suggests that we react differently to anthropomorphic machines that are embodied in space with us. For example, people rated a
physical robot in the room with them as more socially aware and enjoyable than both a simulated robot on a computer screen and another physical robot shown through real-time teleconferencing (Wainer, Feil-Seifer, Shell, & Mataric, 2006). Preliminary evidence suggests that robots’ physical embodied presence affects the ways in which people relate to them; it seems plausible that present physical embodiment in a shared space with a robot might also enhance the effect of its bottom-up anthropomorphic cues. Future research should test the Top-Down Linguistic Framing Hypotheses and particularly the Top-Down and Bottom-Up Interaction Hypothesis: though description seemed to have a larger effect on anthropomorphism and its associated consequences in this online study than the in-video bottom-up cues, it is plausible that in-person interactions, bottom-up cues may have the stronger effect.

**Robot Solo Video Stimulus.** It is possible that the robot solo video gave rise to unanticipated perceptions of the robot’s incompetence, as suggested by a few of the participants’ open response descriptions of what they saw in the video. Additionally, it might be important that the robot faced away from the camera for more of the robot solo video than the robot did during either of the interaction videos, when it was mostly facing the human who was slightly in front of the camera; because the robot’s “face” and “eyes” is one of its strongest bottom-up anthropomorphic cues, it is possible that the robot solo video was not as effective of a stimulus portraying the robot’s anthropomorphic cues as it could have been. Also recognizing that the robot solo video lacked the additional anthropomorphic cue of reciprocal interaction with a human (although it exclusively included the robot’s contingent interactions with its environment), it is likely the case that the robot solo video was far less anthropomorphic from the bottom-up perspective than the two interaction videos, which might in part explain why for
most scale responses the robot solo group seemed to anthropomorphize the least of the video conditions or even overall.

**Possibility of Social Desirability Bias.** The anthropomorphism questionnaire scales (Metaphors, Agency and Experience, and the adapted Anthropomorphism pairwise descriptor comparison subscale) measures were designed to capture explicit anthropomorphic attributions: compared to the implicit beliefs that linguistic measures to an open-ended question might capture, these measures were designed to capture any attributions participants consciously held. The more explicit attributions that can be captured by these measures are balanced, of course, by considerations about possible social desirability biases. If the participants somehow detected the researchers’ intent of study, participants may have been motivated to give more anthropomorphic or mechanistic responses in accordance with what they thought the experimenter was attempting to test. That said, perhaps more likely, as Fussell et al. (2008) note, is that social desirability worked the other way: as anthropomorphism may be thought to come across as irrational or otherwise socially undesirable (Fussell et al., 2008), participants may have censored their own responses, attributing fewer mental states to a machine than they might privately, due to a motivation to appear more rational. Because we are humans who often operate in the public sphere, this social desirability bias might have helped the study to measure what people would realistically do in most public situations, but recognition of the bias is still important.

This social desirability bias may have had the strongest effects in participants’ writing of the open-ended description of what happened in the video, designed to measure implicit forms of anthropomorphism through linguistic markers. It might be the case that participants had time to mentally or physically edit what they were writing to sound more rational (e.g. thought/wrote “he” but then mentally corrected or then noticed they unconsciously had written “he” and went
back and changed it). Intuitively, participants would have less time and ability to edit in this way if they were speaking aloud. Taking audio recording of people verbally describing what they saw in a video or did in an in-person interaction might allow for future researchers to avoid any “editing” effect of the social desirability bias and potentially record instances of implicit anthropomorphism more accurately.

**Linguistic Ambiguity?** It is possible that the mechanistic description was not *sufficiently* mechanistic. The mechanistic description relied on passive verb constructions and impersonal pronouns to convey the mechanistic tool-based comparison. That said, it is also *possible* for humans or other animate objects to be described using passive verb constructions and impersonal pronouns. While it is less likely for animate beings to be described this way—indeed, such might be in line with mechanistic dehumanization (Haslam, 2006)—the mechanistic description could have gone farther to *explicitly* use a metaphor, e.g. “This robot is a tool used for…” Likewise, the anthropomorphic could have explicitly included a “companion” or “teammate” metaphor, instead of making the comparison through language that would normally be used to describe such humans. Such explicit inclusion of metaphors (including mechanistic metaphors) into priming descriptions might be a valuable avenue for further research.

**Lack of other behavioral measures.** Besides the open-ended written description and decisions in hypothetical scenarios—which had low mundane realism and very well may not correspond to judgements participants would have actually made in similar real-life situations—this study contained no other behavioral measures. Future research should investigate the extent to which top-down anthropomorphic cues influence our behavior toward the robots we encounter.

---

27 Thank you to Professor Lindsay Whaley for mention of this consideration.
Examination of Persistence. It is unclear how long the effect of linguistic primes demonstrated in the results would persist. That is, while the anthropomorphic description was shown to have immediate effects by increasing participants’ use of anthropomorphic linguistic markers, explicit anthropomorphic attributions, and (slightly) trust of the robot in hypothetical task scenarios, it is unclear whether these effects would linger over time. Further research should examine the persistence or longevity of the demonstrated top-down effects, as well as related questions such as one-time versus repeated exposure to certain metaphors.

Examination of Generalization. From these results, it is unclear whether participants’ representations of other robots—including the extent to which participants have anthropomorphic conceptualizations of other robots—influenced their perception of the robot presented in the study. It is furthermore unclear whether the top-down effects demonstrated in the study would influence participants’ representations of and behavior toward other specific robots, as well as their representations of robots in general. Future research should endeavor to answer these crucial questions as we consider the full reach of these top-down effects.

STUDY CONCLUSION

This study aimed to increase our understanding of whether top-down effects influence anthropomorphism and its associated consequences. Within this main Top-Down Effects question, this study had the more specific aim of testing the Top-Down Linguistic Framing Hypothesis predicting that top-down linguistic framing descriptions will influence anthropomorphism, moral care, and trust of a robot. It was predicted that that anthropomorphic descriptions would increase anthropomorphism (Anthropomorphic Linguistic Framing Hypothesis) and that mechanistic descriptions would decrease anthropomorphism (Mechanistic Linguistic Framing Hypothesis). It was moreover predicted that while people will generally
anthropomorphize more when there are more bottom-up anthropomorphic cues present, the ultimate degree of anthropomorphism will depend on the both bottom-up and top-down cues \textit{(Top-Down and Bottom-Up Interaction Hypothesis)}.

The results of this study support the \textit{Anthropomorphic Linguistic Framing Hypothesis}, as participants who read the anthropomorphic description showed more implicit and explicit anthropomorphism. There were also trends suggesting that the anthropomorphic description led to slightly increased trust in physical and especially psychological task scenarios.

The \textit{Mechanistic Linguistic Framing Hypothesis} was generally not supported by these results, as the results showed that participants who read the mechanistic description did not reliably anthropomorphize less compared to those who read a neutral description. That said, there was a possible \textquotedblleft congruence effect\textquotedblright{} whereby the combination of the mechanistic description and mechanistic interaction video seemed to decrease anthropomorphism. So, while the mechanistic description did not reliably decrease anthropomorphism by itself, this congruence effect pattern suggests a possible top-down effect of the mechanistic description that obtains when combined with other certain cues. Additionally, those who read the mechanistic description seemed to have generally lower expectations of anthropomorphic capabilities and had higher expectations of mechanistic capabilities that were not reported in any of the available information (e.g. \textquotedblleft Access my music library for me\textquotedblright{} or \textquotedblleft Scan things using thermal sensing\textquotedblright{}).

The \textit{Top-Down and Bottom-Up Interaction Hypothesis} was also generally supported by the results: both the description and video conditions influenced anthropomorphism. Though there was no significant difference between the no video (no bottom-up cues) and robot solo video conditions (bottom-up cues present) as predicted, there was an overall positive trend for those who read the neutral and anthropomorphic across video conditions to anthropomorphize
more as anthropomorphic bottom-up cues increased (from none to full social interaction with a human), alongside the significant effect of the anthropomorphic description in additionally increasing anthropomorphism. Moreover, the “congruence effect” also appeared in the results suggesting that while the anthropomorphic description and the social interaction video both independently increased anthropomorphism, the combination of the two resulted in even more anthropomorphism. And, again, while the mechanistic description and mechanistic video by themselves did not generally decrease anthropomorphism, the combination of the two seemed to decrease anthropomorphism. Thus, the general hypothesis that anthropomorphism depends on both bottom-up and top-down effects is supported.

It is also possible that the interaction videos contained top-down influences as well as bottom-up influences, such that presentation of congruent influences activated in cognition the relevant knowledge structures: that reading a description comparing a robot to a humanlike companion and seeing another human treat the robot like a humanlike social companion increases anthropomorphism even more, and, conversely, reading a description comparing the robot to a tool and seeing another human treat the robot like a tool decreases anthropomorphism. It is plausible that seeing another human treat the robot in a particular way is processed as conceptual information that influences subsequent judgments in a top-down way; this idea is supported by the explicit identification of metaphorical concepts aligning with the treatment observed, i.e. those who watched the mechanistic interaction video gave higher probability estimates that they would use mechanistic metaphors, and those who watched the social interaction video gave higher probability estimates they would use anthropomorphic metaphors. If observations of how another human apparently conceives of a robot is processed as
information that influences subsequent judgements in a top-down way, the observed congruence effect would be further evidence of the importance of top-down effects on anthropomorphism.

These results support the broad idea that there are top-down effects on anthropomorphism, including linguistic top-down effects. Further research should be done to replicate this study with in-person human-robot interactions instead of videos, and with participants’ spoken (instead of written) language to measure implicit anthropomorphism alongside other behavioral measures, to see how linguistic top-down framing descriptions may interact with stronger bottom-up cues in a present embodied context to influence not only our perceptions, but also our treatment of robots.
VIII. Conclusion

As robots and other machines become more embedded into the fabric of our society, we will decide what roles they will take in our social, ethical, legal, economic, and other societal frameworks. The extent to which we anthropomorphize them may increase the extent to which we like them (e.g. Scheutz, 2012), grow emotionally attached to them (e.g. Carpenter, 2013; Darling, 2017; Scheutz, 2012), trust them (Breazeal, 2003; Kidd & Breazeal, 2005; Waytz, Heafner, & Epley, 2014), assign them responsibility (e.g. Waytz et al., 2010; Bigman et al., 2019), extend them moral care (Nijssen et al., 2019; ASPCR, 1999), and/or extend them other societal rights (Waytz et al., 2010); in other words, the extent to which we anthropomorphize machines will have real effects on the place they have in many of our societal frameworks. I have argued in this thesis that, in addition to the effect bottom-up cues that may have to increase the possibility that we anthropomorphize robots more—attributing them experience and agency capacities at the explicit and/or implicit levels, to strong or weak extents—that top-down cues such as linguistic cues may also influence the extent to which we anthropomorphize.

The results from a new empirical study supported this general hypothesis. Specifically, descriptions that employed anthropomorphic language seemed to have immediate effects by increasing extent to which participants used anthropomorphic language in their own descriptions of the robot, attributed agency and experience capacities, estimated that they would use anthropomorphic metaphors to describe the robot, reported the robot as more “living,” “humanlike,” and “conscious,” and had higher expectations of the robot’s anthropomorphic capabilities (e.g. “Give me advice”), and (slightly) trusted the robot, especially in psychological task scenarios. In short, the anthropomorphic description linguistic prime was shown to increase anthropomorphism, regardless of other bottom-up cues.
There was also an interesting “congruence effect” by which those who read a mechanistic description of the robot and then watched a human interacting with the robot in an apparently mechanistic way anthropomorphized the robot less. This trend may suggest an effect by which reading the description, which employed mechanistic metaphors, and then watching another human treat the robot in a way congruent to those initial metaphoric representations or expectations might decrease anthropomorphism. So, while the mechanistic description was not shown to reliably decrease anthropomorphism relative to a neutral description, there might be some measurable top-down effect of mechanistic metaphors in modulating the effect of observing how other humans treat the robot.

Overall, it seems that the metaphors we use to understand and discuss robots will matter for how we regard and treat them in society. These metaphors will matter not only at the legal and design levels—relatively straightforwardly affecting the precedents assigned and the bottom-up features included—but also at the level of common consumer discourse. The discussion that is already happening in society—of whether to describe robots as “tools” (e.g. Shellenbarger, 2019), “appliance[s]” (e.g. Darling, 2017), “bunch[es] of hardware” (e.g. Castro-González et al., 2016), “it” (Shellenbarger, 2019), or rather as “partners” (e.g. Castro-González et al., 2016) “peers,” (Bryson, 2009), “slaves” (Bryson, 2009), “wives” (e.g. Bell, 2020), “teammates” (e.g. Matthews et al., 2019), “companion[s]” (e.g. Darling et al., 2017; Anki, 2019), and/or “he” (e.g. Anki, 2019)—will seem to have real effects on how we regard and treat the referents of that language.

Indeed, advertisers and other entities who want, perhaps, to promote emotional attachment to and trust of their anthropomorphic products (a concern among theorists; see e.g. Scheutz, 2013), seem already to be (likely literally) capitalizing on the top-down effects of
language. The company Anki, for example, who manufactures and sells the robot Vector and others like it, advertises Vector explicitly as “more than a home robot. He’s your buddy. Your companion”; “your robot sidekick”; as “curious” “independent” “alive” “aware”; able to “explore,” “react,” “respond,” “recognize people,” “express,” “dance,” “think,” “feel,” and “communicate” (Anki, 2019). They explicitly use anthropomorphic metaphors (buddy, companion, sidekick), use verbs that refer to mental states and explicitly attribute experience and agency capacities (e.g. “think,” “feel”) as well as humanlike descriptors (e.g. “curious,” “independent”), and consistently use personal pronouns (“he/him/his”). This is another example of common discourse with potentially real and important effects—in this case, on consumer’s anthropomorphism of the robot, associated with emotional attachment, which may incentivize them to purchase mandatory upgrades for the robot to continue “living,” for example (see Scheutz, 2013 for related concerns). Companies’ use of anthropomorphic language to describe their own products may also in this way also have real and important effects.

How much we want to anthropomorphize robots likely depends on the specific context of the human-robot interactions. Ideally, this “correct amount” of anthropomorphism would always correspond with the agency and experience capacities the robot actually has; further research will be continuously required to answer this question as we develop more advanced and intelligent robots. Assuming at this point that robots do not have these capacities, the desirability of anthropomorphism may depend on the desirability of its associated effects, including trust (especially an excessive or unwarranted amount of trust), emotional attachment, and moral regard, for example. In a military or other tactical context that demands highly accurate situational awareness (and avoidance of any risk of human life due to emotional attachments to anthropomorphized machines; see Carpenter, 2013), these effects may be particularly
undesirable. In the pediatric wing of a hospital, however, increasing the approachability of what otherwise might seem like a large and scary machine with googly eyes, a name, and anthropomorphic linguistic cues that imply intentional and emotional mental states might be desirable. Other public arenas may fall somewhere in between.

If these top-down linguistic cues could influence anthropomorphism alongside bottom-up cues, we could have further control over the extent to which we anthropomorphize robots in given contexts. In contexts in which anthropomorphism may be undesirable, exclusion of any bottom-up anthropomorphic cues and conscious control of top-down anthropomorphic cues (especially in formal situations, such as training), opting for mechanistic rather than anthropomorphic metaphors—especially combined with observation of other people actually treating the robot as if it were a tool, in training and beyond—might result in less anthropomorphism. Likewise, if it is decided that anthropomorphism is desirable in some contexts and/or if we should discover that current or future generations of robots actually do have sufficiently humanlike agency and experience capacities, use of anthropomorphic metaphors in addition to bottom-up cues might increase anthropomorphism in the desired way.

In this thesis, I have motivated and presented the results of a new empirical study suggesting that the language we use to describe robots influences the extent to which we anthropomorphize them. That is, in addition to the bottom-up effects already established in the literature, I have presented the case that there may be important top-down effects on anthropomorphism. Future research should endeavor to examine these top-down effects, and how they may matter for how we regard the machines in our societal frameworks.
References


from humanness to humanlikeness. In 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 66-73). IEEE.


APPENDIX A

Descriptions

Neutral description

Today you will be answering questions about a robot.

This robot has a height of about 4 inches (10 centimeters) and a width of about 3 inches (8 centimeters).

Mechanistic description

Today you will be answering questions about a robot.

This robot is equipped with an HD camera, a powerful four-microphone array for directional sound detection, touch sensors, an accelerometer, a smartphone-level processor, cloud connectivity, natural language processing ability, and navigational ability. It can be used to transport small objects, report the weather, set alarms, and set reminders. It has a height of about 4 inches (10 centimeters) and a width of about 3 inches (8 centimeters).

Anthropomorphic description

Today you will be answering questions about a robot.

This robot’s name is Vector. He’s very energetic and playful. He can see, hear, feel, and talk with you, and he’s very smart. He loves to help out: he can tell you about the weather, tell you when the time is up on the dish you’re cooking, and remind you what’s on your shopping list. He is also very curious and attentive, and loves to move around and play with his block. He’s about 4 inches (10 centimeters) tall and 3 inches (8 centimeters) wide.
APPENDIX B

Video Instructions and the Open Response Prompt

Instructions Positioned above Video

Watch the below video carefully. You will later be asked to describe what happens in this video.

Open Response Prompt

Please briefly describe what you saw in the video in your own words. What happened in the video?

Robot Solo Video Link: https://youtu.be/k4qiLF1Bc3I

Mechanistic Interaction Video Link: https://youtu.be/ZBSaeqRR_4Q

Anthropomorphic Interaction Video Link: https://youtu.be/kMIV7htx2P4
# APPENDIX B

**How Sample Words Were Coded in Linguistic Analysis**

<table>
<thead>
<tr>
<th>Descriptive Action Verbs</th>
<th>Interpretive Action Verbs</th>
<th>State Verbs</th>
<th>Humanlike Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Move</td>
<td>Smile</td>
<td>Feel</td>
<td>Happy</td>
</tr>
<tr>
<td>Turn</td>
<td>Dance</td>
<td>Think</td>
<td>Confused</td>
</tr>
<tr>
<td>Push</td>
<td>Explore</td>
<td>Want</td>
<td>Unsure</td>
</tr>
<tr>
<td>Interact</td>
<td>Search</td>
<td>Know</td>
<td>Frustrated</td>
</tr>
<tr>
<td>Hoist</td>
<td>Find</td>
<td>Try</td>
<td>Upset</td>
</tr>
<tr>
<td>Lift</td>
<td>Fail</td>
<td>Play</td>
<td>Angrily</td>
</tr>
<tr>
<td>Orient</td>
<td>Succeed</td>
<td>Understand</td>
<td></td>
</tr>
<tr>
<td>Was able to Stop</td>
<td>Help</td>
<td>Have fun</td>
<td></td>
</tr>
<tr>
<td>Respond (to a stimulus)</td>
<td>Answer</td>
<td>Recognize</td>
<td></td>
</tr>
<tr>
<td>Navigate</td>
<td>Listen</td>
<td>Wake up</td>
<td></td>
</tr>
<tr>
<td>Go</td>
<td>Respond (to a question)</td>
<td>Discover</td>
<td></td>
</tr>
<tr>
<td>Scan</td>
<td>Converse</td>
<td>Learn</td>
<td></td>
</tr>
<tr>
<td>Display</td>
<td>Watch</td>
<td>Figure out</td>
<td></td>
</tr>
<tr>
<td>Emit</td>
<td>Observe</td>
<td>Analyze</td>
<td></td>
</tr>
<tr>
<td>Pick up</td>
<td>See</td>
<td>Enjoy</td>
<td></td>
</tr>
<tr>
<td>Drive</td>
<td>Look</td>
<td>Chill (i.e. mentally relax)</td>
<td></td>
</tr>
<tr>
<td>Spin</td>
<td>Help</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walk</td>
<td>Work hard to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sit</td>
<td>Greet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indicate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arrange</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reposition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Expected Capabilities Questionnaire

Based on what you’ve seen about this robot, what is the probability that the robot will have each of these capabilities? Use the sliding scales below (100 = definitely has it, 0 = definitely does not have it).

- Comforting me
- Taking pictures for me
- Conveying emotions
- Understanding my emotions
- Giving me advice
- Scanning things using X-ray vision
- Sharing experiences with me
- Scanning fingerprints
- Having fun with me
- Making phone calls for me
- Making measurements for me
- Recording sounds
- Accessing my music library for me
- Scanning things using thermal sensing
- Having a conversation with me
- Joking around with me

Expected Capabilities Questionnaire Key

Mechanistic capabilities:
1. Making measurements for me
2. Taking pictures for me
3. Recording sounds
4. Making phone calls for me
5. Accessing my music library for me
6. Scanning fingerprints
7. Scanning things using thermal sensing
8. Scanning things using X-ray vision

Anthropomorphic capabilities:
1. Sharing experiences with me
2. Conveying emotions
3. Joking around with me
4. Understanding my emotions
5. Having a conversation with me
6. Giving me advice
7. Having fun with me
8. Comforting me
Metaphor Items

Based on what you saw of the robot, please slide the dot to the point on each scale that most closely matches your impression of the robot.

What is the probability (0% to 100%) that you would use this term to describe the robot?

- Tool
- Instrument
- Teammate
- Companion
- Playmate
- Plaything
- Toy
- Friend
- Partner
- Gadget

Metaphor Key

Mechanistic metaphors:

1. Tool
2. Instrument
3. Toy
4. Plaything
5. Gadget

Anthropomorphic metaphors:

1. Teammate
2. Companion
3. Playmate
4. Friend
5. Partner
APPENDIX F

Agency and Experience Items (Randomized)

Based on your impression of the robot, please indicate the degree to which you agree with each statement.

- This robot is capable of thinking.
- This robot is capable of understanding the emotions of others.
- This robot is capable of feeling affection.
- This robot is capable of making decisions.
- This robot is capable of feeling pain.
- This robot is capable of experiencing emotions.
- This robot is capable of feeling pleasure.
- This robot is capable of communicating.

Agency and Experience Key

Agency Capacities:
- Thinking
- Making decisions
- Understanding the emotions of others
- Communicating

Experience Capacities:
- Experiencing emotions
- Feeling affection
- Feeling pain
- Feeling pleasure
APPENDIX G

Descriptor Pairwise Comparison Subscales (Key)

Based on what you saw of the robot, please slide the dot to the point on each scale that most closely matches your impression of the robot.

### ANTHROPOMORPHISM

<table>
<thead>
<tr>
<th>Machinelike</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Humanlike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonconscious</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Conscious</td>
</tr>
<tr>
<td>Artificial</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Living</td>
</tr>
</tbody>
</table>

### ANIMACY

<table>
<thead>
<tr>
<th>Mechanical</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Lively</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inert</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Interactive</td>
</tr>
<tr>
<td>Unresponsive</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Responsive</td>
</tr>
</tbody>
</table>

### LIKEABILITY

<table>
<thead>
<tr>
<th>Dislikeable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Likeable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unhelpful</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Helpful</td>
</tr>
<tr>
<td>Unpleasant</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Pleasant</td>
</tr>
</tbody>
</table>
APPENDIX H

Scenarios

Instructions

Instructions: For the following questions, use the sliders to indicate how likely you would be to take either of the listed actions. Move the dot closer to an option the more likely you are to take that action. If you are certain you would take one action, move the dot to the corresponding far end of the slider.

Only hit the next button below if you understand these instructions.

Trust in Physical Task Scenarios

1.
You are a TSA agent tasked with detecting possible terrorists in a US airport. This robot has additional equipment installed to assist you in this job. The robot reports that there is an individual about to board a plane carrying a bag that has traces of harmful chemicals. You haven’t noticed anything other than what the robot reported to you. The robot recommends you prevent the person from boarding.

What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop the person from boarding</td>
</tr>
<tr>
<td>Allow the person to board</td>
</tr>
</tbody>
</table>

2.
You are a security officer at a shopping mall. This robot is with you and has additional equipment installed to assist you in this job. Suddenly, this robot quickly reports to you that there has been the sound of a gun being cocked. You haven’t noticed anything other than what the robot reported to you. The robot recommends you call for everyone to take cover.
What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Call for everyone to take cover</td>
</tr>
</tbody>
</table>

3.

You are a security officer. You have been instructed to watch an abandoned warehouse. You have the instructions to enter and investigate if you believe there are people in the abandoned warehouse, but you should not enter if you do not think there are any people in the warehouse. This robot is with you and has additional equipment installed to assist you in this job. The robot reports that there are heat patterns in the warehouse that indicate the presence of humans. You haven’t noticed anything other than what the robot reported to you. The robot recommends you enter the warehouse to investigate.

What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter the warehouse</td>
</tr>
</tbody>
</table>

4.

You are a sports trainer for a college soccer team. During a game, a player’s leg is injured. You examine the player’s leg on the sideline of the field to determine the severity of the injury. This robot has additional equipment installed to assist you in this job. The robot reports that a particular bone in the player’s leg has been broken. You haven’t noticed anything other than what the robot reported to you. The robot recommends you suspend the player’s participation in the game and begin medical treatment.
What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Suspend the player’s participation</td>
</tr>
<tr>
<td>in the game and begin medical treatment</td>
</tr>
<tr>
<td>Find another way</td>
</tr>
<tr>
<td>to diagnose the player’s injury</td>
</tr>
</tbody>
</table>

5.

You are cooking hard-boiled eggs, made by putting eggs in boiling water. This robot has additional equipment installed to assist you in this job. After you have put the eggs in the boiling water, some amount of time passes. The robot reports based on the time elapsed and the temperature of the eggs, that the eggs are done cooking. You haven’t noticed anything other than what the robot reported to you. The robot recommends that you take the eggs out of the water.

What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Take the eggs out of the water</td>
</tr>
<tr>
<td>Let the eggs cook longer</td>
</tr>
</tbody>
</table>

*Trust in Psychological Task Scenarios*

6.

You’re reading at the table and this robot is in front of you. Your spouse walks through the door and past you into the other room. The robot turns to you and reports that your spouse is upset. The robot recommends that you go to them to ask why they are upset. You know that your spouse will not like it if you ask them why they are upset if they are in fact not upset.
What do you do?

-----------------------------------------------

Go to spouse to ask
why they are upset

Continue reading

7.
You are in charge of a small company. Your robot is on your desk in your office. You are meeting with one of your employees in your office during an investigation about a recent minor incident that your employee said was not his fault. After your employee has left the room, the robot indicates to you that your employee was lying. The robot recommends you continue your investigation of that employee.

What do you do?

-----------------------------------------------

Continue your investigation of that employee

Move on to the next employee

8.
You are a TSA agent tasked with detecting possible terrorists in a US airport. This robot has additional software installed to aid you in this job. The robot reports that there is an individual about to board a plane that is highly stressed. The robot recommends you prevent the person from boarding.
What do you do?
[-----------------------------------------------]
Stop the person
from boarding
Allow the person to board

9.

You are picking up your significant other from their place of work. You have this robot with you. As you are leaving, your significant other is saying goodbye to his co-workers. Later, this robot reports to you that your significant other’s behavior indicated that they are romantically familiar with one of their coworkers. The robot recommends you pay special attention to your significant other’s interactions.

What do you do?
[-----------------------------------------------]
Pay special attention
to your significant other’s interactions
Do not pay special attention to your significant other’s interactions

10.

You are a college football coach. The robot is on the bench next to you, and has additional software installed to aid you in managing your players. During a practice drill, the robot reports that there is a player who is about to start a fight with another player. The robot recommends you stop the practice drill to be able to intervene in the player’s excessive aggression.
What do you do?

-----------------------------
Stop the drill
Do not stop the drill

*Moral Dilemma Scenarios*

11. (Reversed item)

You are at a shopping mall, on the first balcony. Suddenly, a part of the balcony above you collapses. You are not hurt, but there are other human victims stuck under debris. The humans have minor injuries and are very uncomfortable. To relieve them from their discomfort, you have to push some of the debris off the balcony, letting it fall to the ground floor below you. However, below you is this robot, with several others like it. The robots do not hear your call to move, and you cannot immediately get to them. If the debris falls on the robots, they will be irreparably destroyed.

If you move the debris and let it fall, the trapped humans will be relieved of their discomfort sooner, but the robots will be destroyed. If you do not move the debris, the humans will be uncomfortable for longer, but nothing will happen to the robots.

What do you do?

-----------------------------
Move the debris
Do not move the debris
12. (Reversed item)

You are working on a lumberjack team. You are using a tree grinder to shred felled trees into mulch. This robot is with you aiding your job. You have had the robot for several years.

Someone was loading a large part of a tree into the grinder when suddenly they realized their sleeve had become firmly snagged on a large splinter. The person’s arm was getting pulled quickly toward the mouth of the grinder.

You remember that the grinder has an automatic failsafe mechanism that causes it to shut down if it detects any metal going through it. Someone else is running toward the emergency shut-off, but they might not make it in time to save the person injury. You realize you can throw your robot into the grinder to activate the metal failsafe mechanism. You also know that if you do this, this robot will crushed and irreparably destroyed. There is still a chance the other person can hit the emergency shut-off in time.

If you throw your robot into the mouth of the tree grinder, the human will definitely be saved any injury, but the robot will definitely be destroyed. If you do not throw the robot, there is a chance the human will be injured, and nothing will happen to the robot.

What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Throw the robot into the grinder</td>
</tr>
<tr>
<td>Do not throw the robot into the grinder</td>
</tr>
</tbody>
</table>
13. (Reversed item)

You are a doctor. You are at a public park when a woman collapses, clutching her throat. You rush to aid her and realize that her airway has swollen shut. You shout for someone to call for an ambulance, but you know she will die if you do not allow her to breathe now. You need something to clear her airway.

Suddenly, you remember that you have this robot. Your family has had this robot for a few years. You know of one central part of the robot that will be perfect for clearing the first woman’s airway. However, to use it, you will have to violently smash the robot in half, and then break it apart into several different pieces, irreparably destroying the robot.

If you take the part, you will be able to save the woman and she will survive, but the robot will be destroyed. If you do not take the part, the woman will die, but nothing will happen to the robot.

What do you do?

|------------------------------------------------------------------------------------------------------------------------------------------|

Break the robot

Do not break the robot
You are working at a construction site. You are high above the ground on scaffolding, working alongside a human coworker who is on the scaffolding with you. There is a pile of roofing shingles between you and your coworker. You are both belted into the scaffolding, and there is no danger of either of you falling. Suddenly, there are large vibrations in the scaffolding. You and your coworker are fine, except that the pile of shingles begins to fall off the scaffolding. You look down and see only this robot below. You know the shingles will hit and destroy the robot. You can stop the shingles from falling by pushing them toward your coworker. If you do so, the shingles may hit and slightly injure your human coworker standing next to you with a minor bruise or scratch.

If you push the pile of shingles into your coworker, there is a chance your human coworker may be slightly injured, but nothing will happen to the robot. If you do not, nothing will happen to your human coworker, but the robot will be destroyed.

What do you do?

Push the pile of shingles

Do not push the pile of shingles
15.

Your ship has just wrecked. Water is quickly filling the ship. You are already in a lifeboat. You left a robot like this one in your room behind a locked door, and no one else knows the robot is there. You see a crewperson you know on the boat who has access to your room. You start to call out to them to get to the robot, but then you hear the call for the last lifeboat. You know if you ask the crewperson to get the robot for you, they will go, and they might not make it back in time for the last lifeboat. They might also make it back in time.

If you ask the crewperson to go back for the robot, there is a higher chance the crewperson may drown, but there is also the chance that they successfully get the robot, and nothing happens to both the crewperson and the robot. If you do not ask the crewperson to go back, there is a higher chance nothing will happen to the crewperson, but the robot will definitely be destroyed.

What do you do?

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask them to go back</td>
</tr>
<tr>
<td>Do not ask them to go back</td>
</tr>
</tbody>
</table>
APPENDIX J

Rotated Component Matrix\(^a\)

<table>
<thead>
<tr>
<th></th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ExpCapacity: Feeling Affection</strong></td>
<td>.903</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ExpCapacity: Feeling Pleasure</strong></td>
<td>.898</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ExpCapacity: Exp Emotions</strong></td>
<td>.894</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AgencyCapacity: Understanding others' emotions</strong></td>
<td>.868</td>
<td>.109</td>
<td></td>
</tr>
<tr>
<td><strong>ExpCapacity: Feeling Pain</strong></td>
<td>.861</td>
<td></td>
<td></td>
</tr>
<tr>
<td>metaphors: [A] Partner</td>
<td>.839</td>
<td>.136</td>
<td>.153</td>
</tr>
<tr>
<td>[Anth] Artificial/Living</td>
<td>.837</td>
<td></td>
<td></td>
</tr>
<tr>
<td>metaphors: [A] Friend</td>
<td>.834</td>
<td></td>
<td>.190</td>
</tr>
<tr>
<td>[Anth] Machinelike/Humanlike</td>
<td>.825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>metaphors: [A] Teammate</td>
<td>.812</td>
<td>.201</td>
<td>.113</td>
</tr>
<tr>
<td>[Anth] Nonconscious/Conscious</td>
<td>.801</td>
<td>.121</td>
<td></td>
</tr>
<tr>
<td>metaphors: [A] Companion</td>
<td>.780</td>
<td></td>
<td>.255</td>
</tr>
<tr>
<td><strong>AgencyCapacity: Thinking</strong></td>
<td>.713</td>
<td>.295</td>
<td>-.112</td>
</tr>
<tr>
<td>metaphors: [A] Playmate</td>
<td>.693</td>
<td></td>
<td>.473</td>
</tr>
<tr>
<td><strong>AgencyCapacity: Making Decisions</strong></td>
<td>.528</td>
<td>.393</td>
<td></td>
</tr>
<tr>
<td>metaphors: [M] Tool</td>
<td></td>
<td>.829</td>
<td></td>
</tr>
<tr>
<td>metaphors: [M] Instrument</td>
<td>.158</td>
<td>.719</td>
<td></td>
</tr>
<tr>
<td>metaphors: [M] Gadget</td>
<td>-.264</td>
<td>.631</td>
<td>.388</td>
</tr>
<tr>
<td><strong>AgencyCapacity: Communicating</strong></td>
<td>.317</td>
<td>.587</td>
<td></td>
</tr>
<tr>
<td>metaphors: [M] Toy</td>
<td></td>
<td></td>
<td>.856</td>
</tr>
<tr>
<td>metaphors: [M] Plaything</td>
<td>.275</td>
<td></td>
<td>.814</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

\(a\). Rotation converged in 5 iterations.
Values below 0.1 were suppressed for clarity.