Cognitive effects of the anthropomorphization of artificial agents in human-agent interactions

Bas Vegt & Roy de Kleijn

Bas Vegt
Cognitive Psychology Unit
Leiden University
Netherlands
basvegt@gmail.com

Roy de Kleijn (corresponding author)
Cognitive Psychology Unit
Leiden University
Netherlands
kleijnrde@fsw.leidenuniv.nl

Published as: Vegt, P. S., & de Kleijn, R. (2023). Cognitive effects of the anthropomorphization of artificial agents in human–robot interactions. In S. Mukherjee, V. Dutt, & N. Srinivasan (Eds.), *Applied cognitive science and technology: Implications of interactions between human cognition and technology.* Springer.

Abstract

As our social interaction with artificial agents is expected to become more frequent, it is necessary to study the cognitive effects evoked or affected by these social interactions. Artificial agents come in all shapes and sizes, from vacuum cleaners to humanoid robots that in some cases can be difficult to distinguish from actual humans. Across this wide range of agents, different morphologies are believed to have different effects on humans in social interactions. Specifically, the extent to which an agent resembles a human has been shown to increase anthropomorphization, the tendency to attribute human characteristics to non-human agents or objects. From an evolutionary perspective this response is completely reasonable, since for most of our existence as a species, if something looked like a human, it would almost always have behaved like one. However, this is not necessarily the case for artificial agents, whose intelligence can be implemented independently of morphology. In this chapter we will review the cognitive and behavioral effects of anthropomorphization such as prosocial behavior, empathy, and altruism, as well as changes in subjective experience that can occur when interacting with human-resembling artificial agents. We will discuss results from behavioral experiments, economic games, and psychophysiological evidence. We will first give a review of the current state of the field before discussing some inconsistent findings, and shed light on areas that have been underinvestigated as of yet.

Keywords: human–robot interaction, anthropomorphism, anthropomorphization, artificial agents, robots

Words: 5290

Introduction

Of the many different types of useful robots, a substantial number of them can work for or alongside humans. Whether performing some service to customers, working together with laborers, or merely sharing a workspace with them, humans and robots will need to socially interact. Considering this, roboticists have been turning to anthropomorphism in their design to better serve these needs, and to ultimately involve robots in day-to-day life more seamlessly. Intuitively, one may think that robots are easier and more pleasant to interact with if they resemble humans, since such interactions must feel more natural. After all, people are generally more accustomed to interacting with a human than with any given complex technology. However, reality paints a slightly more complex picture, as we will discuss in this chapter.

Anthropomorphism is the extent to which an object or agent is reminiscent of a human, and robot behavior and external features can be explicitly designed to be more anthropomorphic—in other words, to represent human features more closely. Under the right conditions, human observers can attribute humanlike characteristics to non-human agents. This tendency is known as *anthropomorphization*, and can be partly induced by anthropomorphic design, although this feature is neither necessary nor sufficient for the phenomenon to occur.

Let us first examine the approach of anthropomorphic design, from its theoretical validity to the degree of success it has had so far and is projected to have in the future. We will also consider its effects on human users, such as anthropomorphization, and its effects in turn.

Anthropomorphic design

If we want humans and robots to interact, human users should be taken into account in robot design. Any well-designed tool should not only allow the user to fulfil its intended purpose, but it should do so while minimizing harm to and effort from the user, and robots are no exception. If at all possible, we want them to elicit a positive user experience.

When we call a product, program, tool or robot “easy to use”, we generally means that it is effective and intuitive (Norman, 2013). In other words, it should be easy to remember or find out the steps required to yield a desired result, and they should be easy to perform. People generally prefer using familiar tools over learning new ones because they already know how the former work, often as a result of much time spent training, whereas learning to use new tools or conventions from scratch takes more time and effort. Even if new tools promise to be more efficient in the long run, people often stick to what they know either because they deem the time investment required to switch to be too high, or because they do not feel tech-savvy enough to do so (Norman, 2013).

Whereas many older products enjoy the benefit of having defined what many users are now familiar with and thus expect, many newer products take advantage of this by fitting their design to mirror older designs, building on this existing foundation. In other words, rather than making the user familiar with the product, they make the product to be familiar to the user by design. This is one of two purposes served by robot anthropomorphism: rather than have to learn how to interact with a robot, ideally we would simply interact with one as if it were a human, which is something we all have experience with. The other purpose behind anthropomorphic design is a more pleasant user experience. There are hurdles in the way of both of these goals, but the pleasantness of interacting with human-like robots faces a noteworthy one in particular: the *uncanny valley*.

The uncanny valley

First hypothesized by Mori et al. (2012; translated from the original in 1970), the uncanny valley describes a non-linear relationship between an object's human resemblance and its elicited emotional response. While humans tend to rate an object's aesthetics as more pleasant the more it resembles a human, it seems that high but imperfect resemblance actually evokes negative affect in most people. The response can range from a mild distaste or faint eeriness to profound discomfort and genuine revulsion, explaining the phenomenon's prevalence in design philosophies for fictional horror (Tinwell, 2014).

While perhaps it makes intuitive sense that we would be wired to like smaller creatures that resemble ourselves (usually in the form of our children), the severity of the uncanny valley is more difficult to explain. Historically, the term *uncanny* has been conceptualized as familiar yet unknown—not quite so novel as to be mysterious, but all the more unsettling for appearing in a seemingly wrong, unreal context (Jentsch, 1906; Freud, 1919). Indeed, much research initially seemed to support the suggestion that the eerie feeling comes from an inability to classify an object as either human or non-human (e.g. MacDorman et al., 2009; Burleigh et al., 2013; Strait et al., 2017) which fits alongside another concept from horror and mythological monsters, called *category jamming:* humans, being categorization machines, deeply dislike being unable to classify something they encounter (Carroll, 1990). For example, most people experience no eerie sensation looking at a Roomba or other cleaning robots designed using simple geometric shapes, and are at worst mildly uncomfortable around stationary mannequins. However, put a mannequin's torso and head on a Roomba, and suddenly shoppers feel uneasy. And indeed, numerous studies attribute the uncanny valley to these same two factors: atypicality and ambiguity (Strait et al., 2017). These studies suggest it is best to keep a robot's features internally consistent, designing either something which does or does not bear resemblance to humans, rather than producing hybrids.

It seems, however, that for now the uncanny valley cannot be simply avoided by following the above guidelines. Both Burleigh and Schoenherr (2014) and MacDorman and Chattopadhyay (2016), previously among those to identify atypicality and ambiguity as the main factors contributing to the effect, have since published results that contradict their previous conclusions. Meanwhile, results confirming their original findings are still being published, indicating that there is no real consensus on this yet.

Social robotics

Not all machines need to be anthropomorphic in their design. Automatic cars, dishwashers, and smart-home systems, although they may recognize and produce human speech, function perfectly well without a humanlike morphology. However, there are some which benefit immensely from being modeled after a human, because their functionality does not entail mechanical tasks but rather providing a soft service or filling an emotional need. An android serving as a personal assistant, bartender, or home-care giver distinguishes itself from the earlier examples precisely due to their human form factor. So-called *social robots* rely on establishing a connection we would normally not extend to computers. These robots could in theory navigate spaces and perform much of their operations with *any* form factor, but they would far less effectively serve their primary function: to put humans at ease or provide company. We visit restaurants, for example, not just for tasty food, but also to be treated by friendly and socially apt staff. Similarly, in fields such as education, medical care, and therapeutic interventions, we expect well-trained, compassionate professionals to take good care of us or our children. It is worth noting that some of these roles can also be assumed by robots modeled after pets and other animals rather than humans but nevertheless these robots all employ the same power: empathy.

Empathy and attribution of mind

Empathy is an emotional affect in which a person identifies with another to a sufficient degree to mentally take their perspective, and to experience an emotional state similar to that of another, albeit in a milder form. It effectively prevents interactions from going south by reflecting part of the negative experience back on the offender (Radzvilavicius et al., 2019). It is simply difficult to do something terrible to someone while looking them in the eye. Empathy is also heavily involved in compassion, although the terms are not synonymous. Briefly, compassion includes a desire to help alleviate the negative situation the victim is in, whereas empathy merely entails feeling the victim's plight. Thus, while each can occur without the other, compassion is almost always preceded by empathy.

The effect of empathy is stronger when we feel closer to the offended party, and this emotional distance is strongly determined by perceived similarity and physical distance between the two parties (Bregman, 2019). This is perhaps best illustrated by how any reasonable person would readily ruin a $100 suit to save a child drowning in front of them, whereas most people would neglect to donate such sums to charity to save lives daily (Singer, 1972). By the same principle, historians describe how it is possible for citizens of opposing countries to do unspeakable things to one another at time of war (Bregman, 2019), group dynamics can explain why there is so much more hostility between rather than within homogeneous teams at a company (Forsyth, 2006), and why sociologists worry about the adverse effects of relative distance and a lack of accountability on social media (Lapidot-Lefler & Barak, 2012). The mechanisms of empathy and compassion can be seen anywhere there is human interaction. At its best, it's the most effective hostility dampener yet conceived, and it can protect human–computer interaction in the same way, if only we can extend people's empathic instincts to non-human and even non-sentient targets.

There is evidence to suggest that humans can empathize with machines that resemble humans to a sufficient degree (Misselhorn, 2009). This similarity need not be limited to the aesthetic sense that an android *looks* more like a human than does a Roomba, but also in the sense that a Roomba, being a relatively autonomous agent, already bears more resemblance to a human than for example a refrigerator. And indeed, we can be made to feel empathy for machines as simple as Roombas under the right conditions, such as when we try to judge the morality of seemingly hostile actions (Hoenen et al., 2016).

There are myriad approaches by which one can attempt to judge the morality of an action, but for now we only need to consider what is at the center of moral judgements systems: lived experiences of conscious agents. An action is morally good because elicits or attempts to elicit, on balance, positive experiences. Conversely, it is bad if it elicits or attempts to elicit, on balance, negative experiences. The way in which these intentions or actions, or the resulting experiences, are judged and weighed differ substantially between worldviews, but experiences of conscious agents are involved directly or indirectly in all moral judgements as a matter of course (see Harris (2010) and Harari (2017) for a more elaborate account and discussion on the topic).

One framework that tries to capture the process of this moral judgment in so-called *moral dyads*, suggests that in order to judge the morality of an action, that action must involve at least one agent in both the causal and receiving sides of that action (Gray, Waytz, et al., 2012). If a person breaks a tool or machine in anger, we might judge them because it speaks ill about the ability to keep their temper and not hit other people in heated situations, or because perhaps it was not their machine, or because it took time, money, materials and effort to produce it and someone else could have used those resources, all of which translate to the lived experience of others. But if one removes all outside agents from consideration, most people would not judge a person breaking a tool to be a villainous act, nor a tool falling on a person and injuring them, negligence (by another human) notwithstanding (Gray, Young, et al., 2012).

Of course, this reasoning becomes smudgy as it becomes debatable whether the machine in question should be considered a person in its own right. Humans are not rational beings at the best of times: our intuitions rarely map perfectly onto calculated logic—as illustrated by Singer (1972)'s earlier example—and we can be made to make some rather questionable decisions when they are framed in certain ways (Tversky & Kahneman, 1981). In human–robot interaction, research has suggested we can be tricked into ascribing sentience where we normally would not.

Ward et al. (2013) coined the term *harm-made mind* to describe how participants tended to attribute more "mind" to fictional humanlike inanimate objects (corpses, robots, and permanently comatose patients) when they were damaged or harmed by humans with ill intent. The reasoning goes that people, following the moral dyad intuition, judged harming a patient, humanlike robot, or even a dead body as so morally objectionable that they subconsciously deemed the victims as being more sentient so that they could ascribe a higher degree of moral fault to the offending agent. Participants not only rated these victims' capacity for pain higher than was the case for non-victimized or accidentally harmed equivalents in the control groups, but also felt they were more capable of experience and agency, indicating a higher capacity for planning, self-control and hunger. In other words, it seems that humans, under the right conditions and to a certain extent, can feel empathy towards non-human and even nonconscious entities.

However, Ward et al.'s (2013) findings should be considered carefully. The robot that participants read about in the experimental condition's version of the story was described as being "regularly abused" with a scalpel. This language was not present in the control condition due to the nature of the experiment, and the use of suggestive language to describe the act of damaging a robot or its sensors could have confounded the findings. Since damaging a non-sentient object is not typically cause to accuse someone of "abusing" it, and since it is a word normally reserved only for describing harmful actions against a feeling agent[1], the mere use of this word potentially suggested sentience on the part of the robot, considering subtle differences in language can influence people's accounts of events in substantial ways (Loftus & Palmer, 1974). More recent studies (e.g. Küster & Swiderska, 2020) used visual vignettes instead of text-based ones, largely eliminating language as a potential confound. Although results were largely similar, differences were found between human avatars and humanlike robotic avatars, pertaining to their perceived levels of mind and applied moral standards.

With that caveat in mind, it is worth noting that research has since demonstrated the principle of the harm-made mind in a more concrete manner as well (Hoenen et al., 2016). Participants report marginally more compassion for robots which they felt were treated more aggressively. This could also be seen on a neurophysiological level as measured by mirror neuron activity, showing that there is potential for us to sympathize with these mechanical entities.

It is one thing to ask someone about the idea that machines *could* possess a mind, it is quite another to try to determine where we stand with current robots. Findings such as the above suggest that people are generally even open to the possibility of machine consciousness, but only as a function of how humanlike the machine is. However, a common intuition by which people determine whether an agent is conscious or not is through a combination of its complexity and similarity to a human (Morewedge et al., 2007). When an

[1] Reminding one of the same category mistake famously illustrated by computer scientist Edsger W. Dijkstra: "The question of whether machines can think is about as relevant as the question of whether submarines can swim."

agent's actual complexity is hard to determine, all that remains to go on are their aesthetic and behavioral similarities to us. The field might benefit from research investigating the upper and lower limits of this theory, searching to identify people's thresholds for mind-attribution and empathy. An extended replication of the harm-made mind study could, for example, including a broader range of objects, some intentionally pushing on the edge of the uncanny valley, and others in the other extreme, resembling humans less and less, such as mannequins, dolls, toys, or even planks of wood, to fully test the limits of the suggested effects as a function of complexity and human-likeness.

Physical human–robot interaction

We also seem to extend certain social norms to robots subconsciously, as seen when participants show physiological signs of emotional arousal when touching a robot's intimate (low-accessible) body parts (Li et al., 2017). Here as well, the authors pose as an open question to what extent such signs would occur when touching dolls, mannequins, etc. However, the only known attempted replication of this study thus far (Zhou et al., 2021) found no difference in arousal levels between body parts, although this could be due to methodological limitations of that study, including a low number of participants and the use of a robot which lacks some of the most intimate (i.e. least accessible) body parts: this robot does not have legs as such, and thus no inner things nor clear buttocks or part which would correspond to its genitals. As this is a new, yet-to-be replicated finding, we must employ caution in drawing conclusions on its basis, but at the very least it suggests that people naturally see humanoid robots as something else than mere objects.

Goal-directed action and mirroring

Human beings, as well as some other animals, learn by both observing and doing. We observe and mimic the behavior of others to expand our own skillset. It is believed that at the heart of this phenomenon is a subconscious "mirroring" mechanism, which has been identified on a neurological level (Gallese et al., 2004). When somebody observes another perform an action, brain activity shows patterns similar to when they perform that action themselves (Gallese et al., 1996). We can also infer the intention behind an action as we observe it, allowing us to understand each other's behavior as well as replicate it (Alaerts et al., 2010). The neurological signature of mirroring has been recognized when observing robot actions as well (Oberman et al., 2007). Mirror neurons seem to play a similar role in imitating robot actions as they do in imitating human actions, as long as they are not too repetitious (Gazzola et al., 2007). In one study, participants were instructed to mimic a robot's hand movement as soon as it was finished, but the timing of their response was instead determined by when the robot looked at them, mimicking a common social cue for turn-taking (Bao & Cuijpers, 2017). The fact that participants responded to this cue despite not being told about it suggests targets of mimicry are readily perceived as social agents with intentionality (Wykowska et al., 2014). However, participants seemed to use an action's movement rather than its goal for mimicry, a phenomenon not observed for mimicking humans unless the goal is unclear (Bao & Cuijpers, 2017). This could either be because a robot's more clunky movements require more conscious effort to imitate or because a robot's actions are represented on a different abstraction level than those of a human.

Altruistic and strategic behavior toward robots

Economic games can be used to simulate complex real-world situations using relatively simple rulesets in a controlled environment. These experiments can reveal behavior not easily predicted by pure theory, as illustrated when they were used to lay bare the irrational decisions people make (Harsanyia, 1961). For instance, in the *dictator game*, played with two parties, one party is given a certain amount of money, and asked to divide it over the two parties. Whichever distribution the party dictates is then realized. The *ultimatum game* is similar, but with the addition that the other party may at this point exercise a veto on the offer, in which case neither party gets anything. Strictly rational actors in a single-round ultimatum game would never reject a proposal, as a little bit of money is always better than no money. However, many people are willing to receive nothing if it means preventing the other party from "unfairly" receiving much more (Güth et al., 1982), even though such retribution yields no material benefit for them.[2] Such behavior seems to be driven by an emotional response towards a transgression of fairness norms (Moretti & di Pellegrino, 2010). As such, it only appears when playing against opponents that are held to such norms. This is not the case for computers, which are generally perceived to lack intentionality (Sanfey et al., 2003; Moretti & di Pellegrino, 2010). As for humanoid robots, there is evidence to suggest they are treated more like humans than like computers in ultimatum games (Torta et al., 2013).

[2] However, it should be noted that this effect is moderated by absolute amount, e.g. Anderson et al. (2011) found that rejection rates approach zero as stake size increases.

However, behavior in economic games is itself influenced by expectations and social norms, since participants know they are being watched (Fehr & Schmidt, 2006). This objection has important implications for investigating human–robot social interaction, as behavior towards people is guided by different norms than towards non-human animals and inanimate entities. Therefore, we cannot be sure that only the participants' tendency to anthropomorphize was measured. In addition, the results of these studies tend to be rather fragile. For instance, the effect in which participants more readily accept unfair offers from computers than from humans, can disappear or even reverse under certain circumstances (Torta et al., 2013). In an experiment with non-economic games, children's urge to win was smaller when playing against a humanoid robot rather than a computer (Barakova et al., 2018). The robot was also perceived to be smarter, despite both opponent types using the same game strategy and purposely performing suboptimal moves to give the children an advantage.

De Kleijn et al. (2019) continued the research on the topic of robot anthropomorphism and anthropomorphization tendencies in economic games, flipping the earlier paradigm of Torta et al. (2013) on its head somewhat: rather than use behavior as a measuring tool for anthropomorphization, they include the latter as an independent variable to investigate its effects on strategic and altruistic behavior. They found that sharing behavior in the dictator game was influenced by the physical appearance of robots, but not the participants' anthropomorphization levels, while the reverse was true for the ultimatum game. Based on the results, they posit that playing against entities which they anthropomorphized led people to exhibit more fairness and strategy in their responses, although the anthropomorphization measures were taken after the game, so it is possible participants' scores were also influenced by the game rather than just vice versa. For example, one might rationalize selfish decisions by retroactively minimizing their opponents' human-likeness, or one might attribute more or less human-likeness to the opponent based on their reaction upon receiving their offer.

This study merits careful consideration for the purposes of this chapter. Like in other economic game studies, the researchers used the same text-based interaction between participant and opponent, regardless of opponent type, to control for possible confounds, as is often done in similar studies. As such, the human player could not draw on their charisma and cunning to plead, threaten and shame the participant during the bargaining. Restraining the responses of the human and robotic opponents diminishes the difference between them, which lies partly in the fact that they do not have access to the same toolset. Thus, the decision to facilitate comparison also necessarily undermines it somewhat. Furthermore, participants in this study anthropomorphized all non-human opponents equally, even though they were meant to be ordinal in this aspect. It is possible that vastly different outcomes could be observed with different opponents. Lastly, the fact that participants shared any money at all in a one-off dictator game versus a laptop is hard to account for based on just anthropomorphization, since people are unlikely to feel much empathy for a laptop. This might indicate that other factors, such as a strategic assumption that their choice would influence a later part of the game, or the fear of being judged as greedy, or simply familiarity, had a larger hand in determining participants' choices.

Participants tend to display more cooperation and desirable behavior when they anthropomorphize (Waytz et al., 2010; de Kleijn et al., 2019). However, clear, consistent data contrasting interactions between human and non-human partners is hard to find, and studies often contradict each other or yield ambiguous results themselves. While people tend to behave differently when they believe to be playing with a human than with a computer, this difference is not uniform in direction or magnitude. This is hardly surprising, given that people do not consistently share money the same way even when playing only with humans. There is large variance in the nature of a given person's interactions with different people. What is more, their strategy (and fairness thereof) is influenced by many situational factors, their impression of their opponent's character being only one.

Ethical considerations

In what could well be either the most adorable or frightening of the studies discussed so far, Bartneck et al. (2005) replicated Milgram's (1963) infamous obedience study, but with a small robot constructed from Lego bricks. The robot was programmed to tremble, scream, and beg for participants to stop administering shocks, but since the whole thing was rather quaint, not a single one of the participants headed its pleas. This would likely turn out very different if the robot to be shocked was indistinguishable from a human. Even if someone in a lab coat urges the participant on and ensures them that "it is only a robot, it cannot feel actual pain", it seems reasonable to assume that this would create considerable stress in participants. Leaving aside for now the fact that institutional review boards would likely not approve of any experiment which could cause emotional distress, take a moment to consider the implications of this suggestion.

Consider that currently, artificial agents exist which have a very convincing ability to exhibit emotions, but in fact do not possess any real consciousness—there is nothing "what-it-is-like" to be them, as they are simply computer programs or characters in a videogame. But suppose that it is in fact possible for non-biological life to exist, fully artificial but satisfying any possible demands you could set for its sentience, intelligence, qualia, etc. and undergoing lived experiences. Given this premise in one hand and our earlier observation in the other, we can conclude that the capacity for any such an entity to *feel*, should it exist, does not necessarily stand in a 1:1 relationship with its capacity to *emote*. In other words, agents are imaginable that experience genuine consciousness, but not able to convey it, like a patient with locked-in syndrome. We believe examples such as this highlight the necessity of reasoning about this, so as to not inadvertently harm a certain kind of life, the existence of which we cannot yet with certainty prove or disprove.

On the other side of the same coin, we must be careful blurring the line between human and mechanical agents. As it has been observed that customers tend to be much more mean and rude to automated customer support services than to human employees in the same service (Pozharliev et al., 2021), we must be careful in concluding that we should make robots more convincingly humanlike, so as to improve customer appraisal. For as long as customers continue to keep their temper when chatting with fellow people but not when they believe to be talking to a robot, they might well end up lashing out against a human being who has been given a script and is simply trying to do their job.

Present challenges

Assessing the cognitive and behavioral effects of anthropomorphization requires measuring the process of anthropomorphizing in participants. Two components play a role here: how humanlike a robot looks to a participant, and the cognitive and behavioral effects this perceived human-likeness produces. The interaction between these two components makes it difficult to measure either of them independently.

There are considerable challenges involved in assessing the validity of findings discussed in this chapter as this is a complex field with many unknowns, due in part to a fundamental ambiguity of the myriads of parameters involved. Any given factor A might be used to predict variable B in one study, while it is B that seems to influence A in another, both interpretations sounding equally reasonable. This frequently leaves us with little idea of the actual causality, not to mention the possibility that the effects are bidirectional. Even in cases where the cause is identified, the directionality of the effect is contested. Reasoning on the basis of theory can yield many alternate, conflicting interpretations and predictions. As a result, almost any possible hypothesis could enjoy favorable outcomes and vice versa, making it neigh-impossible to prove or disprove much definitively.

For instance, in the economic games discussed in an earlier section, we assumed that participants would display altruistic behavior by sharing at least some amount of their money with people, but would not extend this behavior to entities which are clearly non-human and with whom one does not need to share money, such as a rock or a teddy bear. We could then measure how much money gets shared, on average, with several entities and place them on a scale from 0 to human. This hypothetical scale is flawed in principle, as neither limit can be consistently defined. Different scenarios will cause some participants to share nothing, even without using robots. Likewise, there is no robot that is "so human" that participants give it all of their money, because merely being human is not sufficient to guarantee that outcome. The same problems and more are on display in the ultimatum game: if participants share more money in this game with a human than with a robot, one might conclude either that the robot was not human enough to sufficiently elicit empathy, or that it was so convincing that the participant forgot it was a robot did not trust it for this reason. As one gets closer to the uncanny valley, these considerations only get more complex and multi-layered.

State of the field

Science relies on the aggregation of data, each study building on the work of previous research. We believe that the field requires a stronger foundation before it can progress further. As our technology and knowledge continues to advance and develop, we must also consider critical implications before we overcome these challenges. With the amount of studies performed and published continuously, it is unavoidable that some of them contradict each other given any significance level, leaving it up to scholars of any field to examine this data and filter out the noise. Foundational fields, whose theories and paradigms have stood the test of time, do not often get torn down and rebuilt from the ground up. For that reason, when something foundational is called into question, a lot can be at stake because decades of consecutive work rests upon it, but it is smooth sailing otherwise. In contrast, newly evolving fields tend to be quite turbulent as they have not yet existed long enough to rely on such central pillars, but rather are rapidly testing new hypotheses and developing methodologies, which can be hurriedly iterated upon, but can just as easily collapse.

In the field of human–robot interaction, the weather is turbulent indeed. The field of psychology is still coming to terms with new physiological and psychometric measures which are still being improved. At the same time, the stakes are high because modern computer science and engineering are advancing at a staggering rate, and their efforts are making robots and the programs that drive them increasingly important in our lives. Any field that attempts to tackle these challenges in tandem faces a Herculean task indeed.

References

Alaerts, K., Swinnen, S. P., & Wenderoth, N. (2010). Observing how others lift light or heavy objects: Which visual cues mediate the encoding of muscular force in the primary motor cortex? *Neuropsychologia*, *48*, 2082–2090. https://doi.org/10.1016/j.neuropsychologia.2010.03.029

Anderson, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review*, *101*, 3427–3439.

Bao, Y., & Cuijpers, R. H. (2017). On the imitation of goal directed movements of a humanoid robot. *International Journal of Social Robotics*, *9*, 691–703. https://doi.org/10.1007/s12369-017-0417-8

Barakova, E. I., De Haas, M., Kuijpers, W., Irigoyen, N., & Betancourt, A. (2018). Socially grounded game strategy enhances bonding and perceived smartness of a humanoid robot. *Connection Science*, *30*, 81–98. https://doi.org/10.1080/09540091.2017.1350938

Bartneck, C., Rosalia, C., Menges, R., & Deckers, I. (2005). Robot abuse – A limitation of the media equation. *Proceedings of the Interact 2005 Workshop on Agent Abuse*, 54–58.

Bregman, R. (2019). *De meeste mensen deugen*. De Correspondent.

Burleigh, T. J., & Schoenherr, J. R. (2014). A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Frontiers in Psychology*, *5*, 1–19. https://doi.org/10.3389/fpsyg.2014.01488

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*, 759–771. https://doi.org/10.1016/j.chb.2012.11.021

Carroll, N. (1990). *The Philosophy of horror: or, paradoxes of the heart*. Routledge.

de Kleijn, R., van Es, L., Kachergis, G., & Hommel, B. (2019). Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human Computer Studies*, *122*, 168–173. https://doi.org/10.1016/j.ijhcs.2018.09.008

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism – Experimental evidence and new theories. In *Handbook of the Economics of Giving, Altruism and Reciprocity* (Vol. 1, pp. 615–691). Elsevier. https://doi.org/10.1016/S1574-0714(06)01008-6

Forsyth, D. R. (2006). Intergroup relations. In *Group Dynamics* (Fourth Ed., pp. 447–484).

Thomson Wadsworth.

Freud, S. (1919). *The Uncanny [2011 archive.org version]*. The Uncanny. https://web.archive.org/web/20110714192553/http://www-rohan.sdsu.edu/~amtower/uncanny.html

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609. https://doi.org/10.1093/brain/119.2.593

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*, 396–403. https://doi.org/10.1016/j.tics.2004.07.002

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

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*, 206–215. https://doi.org/10.1080/1047840X.2012.686247

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

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Harari, Y. N. (2017). The odd couple. In *Homo Deus: A brief history of tomorrow* (pp. 179–199). Harper Collins.

Harris, S. (2010). *The moral landscape: how science can determine human values*. Free Press.

Harsanyia, J. C. (1961). On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution*, *5*, 179–196. https://doi.org/10.1177/002200276100500205

Hoenen, M., Lübke, K. T., & Pause, B. M. (2016). Non-anthropomorphic robots as social entities on a neurophysiological level. *Computers in Human Behavior*, *57*, 182–186. https://doi.org/10.1016/j.chb.2015.12.034

Jentsch, E. (1906). On the psychology of the Uncanny. *Angelaki*, *2*, 7–16. https://doi.org/10.1080/09697259708571910

JoyForAll Home Page. (n.d.). Retrieved September 15, 2021, from https://joyforall.com/

Küster, D., & Swiderska, A. (2020). Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *International Journal of Psychology*, *56*, 454–465. https://doi.org/10.1002/

ijop.12715

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*, 434–443. https://doi.org/10.1016/J.CHB.2011.10.014

Li, J. J., Ju, W., & Reeves, B. (2017). Touching a mechanical body: Tactile contact with body parts of a humanoid robot is physiologically arousing. *Journal of Human–Robot Interaction*, *6*, 118. https://doi.org/10.5898/jhri.6.3.li

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, *146*, 190–205. https://doi.org/10.1016/j.cognition.2015.09.019

MacDorman, K. F., Vasudevan, S. K., & Ho, C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & Society*, *23*, 485–510. https://doi.org/10.1007/s00146-008-0181-2

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378. https://doi.org/10.1037/H0040525

Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines*, *19*, 345–359. https://doi.org/10.1007/s11023-009-9158-2

Moretti, L., & di Pellegrino, G. (2010). Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion*, *10*, 169–180. https://doi.org/10.1037/a0017826

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, *93*, 1–11. https://doi.org/10.1037/0022-3514.93.1.1

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics and Automation Magazine*, *19*, 98–100. https://doi.org/10.1109/MRA.2012.2192811

Norman, D. (2013). *The design of everyday things*. Basic Books. https://doi.org/10.1145/1340961.1340979

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, *70*, 2194–2203. https://doi.org/10.1016/J.NEUCOM.2006.02.024

Paro Robots Home Page. (n.d.). Retrieved September 15, 2021, from http://www.parorobots.com/

Pozharliev, R., De Angelis, M., Rossi, D., Romani, S., Verbeke, W., & Cherubino, P. (2021). Attachment styles moderate customer responses to frontline service robots: Evidence from affective, attitudinal, and behavioral measures. *Psychology and Marketing*, *38*, 881–895. https://doi.org/10.1002/mar.21475

Radzvilavicius, A. L., Stewart, A. J., & Plotkin, J. B. (2019). Evolution of empathetic moral evaluation. *ELife*, *8*. https://doi.org/10.7554/ELIFE.44269

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–1758. https://doi.org/10.1126/science.1082976

Singer, P. (1972). Famine, affluence and morality. *Philosophy & Public Affairs*, *1*, 2229–2243. http://www.jstor.org/stable/2265052

Strait, M. K., Floerke, V. A., Ju, W., Maddox, K., Remedios, J. D., Jung, M. F., & Urry, H. L. (2017). Understanding the uncanny: Both atypical features and category ambiguity provoke aversion toward humanlike robots. *Frontiers in Psychology*, *8,* 1–17. https://doi.org/10.3389/fpsyg.2017.01366

Tinwell, A. (2014). *The Uncanny Valley in games and animation*. CRC.

Torta, E., Van Dijk, E., Ruijten, P. A. M., & Cuijpers, R. H. (2013). The ultimatum game as measurement tool for anthropomorphism in human–robot interaction. *Lecture Notes in Computer Science*, *8239*, 209–217. https://doi.org/10.1007/978-3-319-02675-6_21

Tversky, A., & Kahneman, D. (1981). The framing of decision and the psychology of choice. *Science*, *211*, 453–458. https://doi.org/10.1126/science.7455683

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, *24*, 1437–1445. https://doi.org/10.1177/0956797612472343

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

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLOS ONE*, *9*, e94339. https://doi.org/10.1371/JOURNAL.PONE.0094339

Zhou, Y., Kornher, T., Mohnke, J., & Fischer, M. H. (2021). Tactile interaction with a humanoid robot: Effects on physiology and subjective impressions. *International Journal of Social Robotics*, *13,* 1657–1677. https://doi.org/10.1007/s12369-021-00749-x