Robot Eyes Wide Shut:

Understanding Dishonest Anthropomorphism

Brenda Leong
Future of Privacy Forum
Washington, D.C, USA
bleong@fpf.org

Evan Selinger
Department of Philosophy
Rochester Institute of Technology
Rochester, NY, USA
eselinger@gmail.com

ABSTRACT

The goal of this paper is to advance design, policy, and ethics scholarship on how engineers and regulators can protect consumers from deceptive robots and artificial intelligences that exhibit the problem of dishonest anthropomorphism. The analysis expands upon ideas surrounding the principle of honest anthropomorphism originally formulated by Margot Kaminsky, Mathew Ruben, William D. Smart, and Cindy M. Grimm in their groundbreaking Maryland Law Review article, "Averting Robot Eyes." Applying boundary management theory and philosophical insights into prediction and perception, we create a new taxonomy that identifies fundamental types of dishonest anthropomorphism and pinpoints harms that they can cause. To demonstrate how the taxonomy can be applied as well as clarify the scope of the problems that it can cover, we critically consider a representative series of ethical issues, proposals, and questions concerning whether the principle of honest anthropomorphism has been violated.

CCS CONCEPTS

• Human-centered computing~Interaction design

KEYWORDS

Robots, Artificial Intelligence, Machine Learning, ethics, anthropomorphism

ACM Reference format:

Brenda Leong, Evan Selinger. 2018. Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism, In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAT*'19). ACM, Atlanta, GA, USA, 10 pages.*

Introduction

Devices like Amazon Echo are becoming increasingly popular semi-intelligent personal assistants.[1] As relatively new guests in our homes, people are not always sure how they will function. There has been a lot of anxiety over Echo being a covert spy, even though solid privacy protections exist: the device only starts recording after users wake it with a trigger word; Amazon does not share customer-identifiable information with third parties; and, users can permanently delete what Echo records.[2] People even believed that Alexa—the cloud-based voice service that Echo uses—was programmed to call the police to report domestic abuse.[3]

While products like Echo leave people feeling vulnerable for many reasons, artificial intelligences that are wholly other—devices that listen to, remember, and share information like machines, not like human beings—can be troubling for a fundamental reason. We cannot rely on our ability to size up an Echo by looking at it and talking to it, like we can plenty of other people.

In a forward-looking law review article, "Averting Robot Eyes" (2017), Margot Kaminsky, Matthew Ruben, William Smart, and Cindy Grimm identify a new dilemma lurking on the horizon: robots that exhibit "dishonest anthropomorphism" by being designed to exploit our deeply ingrained, human reactions to human appearances and behavior.[4] These scholars invite us to imagine a robot that exhibits misdirection by looking downward while, at the same time, scrutinizing a nearby person with a camera installed in its mechanical neck. On the surface, the robot's downcast eyes present reassuring visual cues that suggest it cannot see everything, and, furthermore, that its visual limitations offer privacy protections. Unfortunately, the gesture instills a false sense of confidence because the robot has covert sensors and information processors.

Fake-out bot, as we are colloquially calling this thought experiment, has similarities to Echo, but differs profoundly, too. Like Echo, fake-out bot absorbs, crunches, and releases

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
FAT '19*, January 29–31, 2019, Atlanta, GA, USA, © 2019 Association for Computing Machinery. ACM ISBN 978-1-4503-6125-5/19/01□$15.00
https://doi.org/10.1145/3287560.3287591

[1] We would like to acknowledge that some of the material in this paper was originally written as blog posts in Evan Selinger's Medium Series, "When Robots Rule the World." Medium has provided permission for this use..

[2] Stacey Gray, Always On: Privacy Implications of Microphone-Enabled Devices, Future of Privacy Forum Report April 2016, https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf (last visited April 7, 2018).

[3] Nicole Darrah, Amazon says Alexa can't call the cops, New York Post (July 12, 2017), https://nypost.com/2017/07/12/amazon-says-alexa-cant-call-the-cops/.

[4] Margot Kaminsky, Mathew Rueben, Cindy Grimm, William D. Smart, Averting Robot Eyes, 76 Maryland Law Review 983 (2017).

information like a computer, not a person. And while both devices can be personified, fake-out bot has a more human-like appearance than Echo, which looks like a speaker. As we will clarify, the differences can profoundly impact how humans behave around lively objects.

Kaminsky et al. are extremely troubled by dishonest anthropomorphism and argue that the bedrock of privacy regulation, the Fair Information Practice Principles (FIPPS), need to be updated to protect consumers from it. The protections, they argue, should promote the principle of "honest anthropomorphism." "Robot designers should not use anthropomorphism to deliberately mislead users as to privacy features."[5]

While the FIPPS need to be brought in line with advances in technology, honest anthropomorphism is such an important ethical principle that its full utility for policy and design deserves consideration. To this end, we will explore how honest anthropomorphism can be understood as a general ethical principle that covers more than paradigmatic privacy violations. Dishonest anthropomorphism occurs whenever the human mind's tendency to engage in anthropomorphic reasoning and perception is abused. Consequently, honest anthropomorphism has a wide scope of applicability.

We do not use the term "abusive" lightly. Unlike simply tricking the user into a misunderstanding, dishonest anthropomorphism leverages people's intrinsic and deeply ingrained cognitive and perceptual weaknesses against them. Even though people know they're dealing with a machine, they feel inclined to respond as if they were in the presence of a human being; perhaps they are powerless to behave otherwise.[6] While there are many aspects of artificial intelligence (AI) that deserve the basic cautious approach of any consumer good, the anthropomorphic design aspects of robots and digital entities that we discuss present a challenge to even the canniest among us.[7]

Humans are even more prone to anthropomorphizing a robot that walks or talks than a simple appliance that is not designed to look like us, sound like us, act like us, or resemble any being that is alive, like an insect or animal. For example, people intellectually grasp that the Roomba, a robotic vacuum cleaner, is not alive. But the mere fact that the disk moves around as if of its own accord is enough to trigger emotional attachment, naming, and even motivate some people to pre-clean for the device.[8] Indeed, in experimental settings, Roombas that look like they have human faces have even been observed to positively impact how people respond to social exclusion.[9] Furthermore, in other experiments people have objected to torturing robots that simulate a response to pain, and on real battlefields soldiers have objected to the "inhumane" treatment of robots that defuse landmines.[10] In response to criticism that children who bark orders at machines might be behaving rudely—even though these devices lack the capacity to be offended—Amazon decided to give parents the capability of having Alexa encourage interactions that are peppered with "please" and "thank you."[11]

It's important to acknowledge that not all AIs are robots and not all robots are intelligent. To the extent possible, we are endeavoring to use these terms separately and accurately. Nevertheless, the only way to discuss the full reach of dishonest anthropomorphism is to consider situations where design factors converge with machine learning and AI systems, regardless of the presence or complexity of embodied features. Thus, the problem may appear in a robot with its own deceptive external form, or it may occur through exchanges with a disembodied voice that nevertheless causes us to react as we would to an actual human on a speaker. The crucial point is that an artificial learning system that elicits or adapts to our intrinsic and deeply ingrained dispositions is a new and distinctive threat.

Fake-out bot is especially dangerous because of the biases that it can trigger. Humans have evolved, biologically and socially, to only associate certain body parts with the gaze: eyes can be prying, but not necks, cheeks, eyebrows, elbows, fingers, toes, or chins. Downcast eyes convey a sense that the full picture is not seen and cannot be viewed, that only bits and pieces are absorbed. Furthermore, if fake-out bot can smile or frown back at the human user in response to an action, query, or request, that human is almost certainly going to respond as if to a human smile or frown, even though the robot does not experience any compatible emotion. Experiments have shown that people willingly provide personal information to a machine without the reasonable cautions they exhibit elsewhere, simply because a robot appears simplistic or unthreatening.[12]

As more devices and robots are incorporated into our daily lives, dishonest anthropomorphism poses grave threats because bad actors will recognize the value of intentionally exploiting our anthropomorphic tendencies and roboticists who do not understand the power of anthropomorphism will unintentionally create products that are misaligned with consumers' mental models of what embodied gestures mean.

The goal of this paper is to deepen the understanding of what dishonest anthropomorphism fundamentally entails and when

---

[5] Ibid, p. 1008.

[6] *New Study finds it's harder to turn off a robot when it's begging for its life,* The Verge (August 2, 2018). https://www.theverge.com/2018/8/2/17642868/robots-turn-off-beg-not-to-empathy-media-equation.

[7] In legal terms, an "abusive" standard was included in the Consumer Financial Protection Act of 2010 as a tier beyond traditional qualifiers such as "unfair," or "deceptive" practices While this law was limited to financial contexts, it separately defined an abusive standard as taking unreasonable advantage of a consumers' inability to protect their own interests. (Pub. L. No 111-203, tit. X, 124 Stat. 1955 (2010)). Abusive practices have been described with words like "predatory," "unscrupulous," and "unconscionable," thus recognizing that a standard or practice may be more egregious in its exploitation of human vulnerabilities than simply "misleading."

[8] Kate Darling, *"Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration and Policy*, in Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence (Patrick Lin et. al. eds. 2017).

[9] James A. Mourey, Jenny G. Olson, and Carolyn Yoon, *Products as Pals: Engaging With Anthropomorphic Products Mitigates the Effects of Social Exclusion*, 44 Journal of Consumer Research 2 (2017).

[10] Kate Darling, *Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects*, in Robot Law (Ryan Calo et. al. eds. 2016).

[11] *Amazon Alexa to reward kids who say: 'Please'*, BBC News (April 25, 2018). http://www.bbc.com/news/technology-43897516

[12] Woodrow Hartzog, *Unfair and Deceptive Robotics,* 74 Maryland Law Review 785 (2017).

the principle of honest anthropomorphism has been violated. To further these ends, we will proceed in three steps.

First, we will explain why dishonest anthropomorphism poses boundary management problems. In this context, we will clarify why classic instances of online deception (e.g., phishing and catphishing) pose different threats to boundary management skills than dishonest anthropomorphism does. We will conclude the discussion by clarifying why deceptive avatars are the closest online analog to deceptively anthropomorphic robots and arguing that dishonest anthropomorphism poses a new and more dangerous threat to our autonomy. Second, we will deepen our argument that dishonest anthropomorphism threatens boundary management skills by applying philosophical insights into the limits of human predictive capabilities and perception. Third, we will provide a detailed taxonomy of dishonest anthropomorphism and the harms it can cause critically consider a representative series of ethical issues, proposals, and questions.

1 Boundary Management; Cafes, Smart TVs, Online Scams, and Dishonest Anthropomorphism

A key lesson that privacy theorists learned from social psychologist Irwin Altman's studies of "boundary management" is that humans manage their privacy by interpreting the environment around them, assessing the risks and opportunities it poses, and choosing strategies to deal with the possibility that others are observing and listening.[13] Consider being in a crowded café. In this environment, loud talkers draw attention to themselves. Even if the patrons and staff cannot identify the speaker's name and draw a blank when looking at her face, odds are good that with minimal effort that they can understand what is being said, so long as they speak the same language and the blaring words are not coded or filled with esoteric references. Indeed, unregulated volume can be an intrusion upon concentration; it practically invites eavesdropping in public spaces where people are in close proximity to one another. To get some measure of privacy through boundary management in environments where lots of people are gathered, deliberate measures need to be taken to secure obscurity.[14] For example, speaking in hushed tones might help, so long as the people around you cannot read lips and are not using hidden recording devices.

The controversy over what consumers initially expected when they purchased "always on" TVs illustrates how difficult boundary management can be. Consumers did not expect that their newly-designated "smart" televisions—a device overtly similar to the much simpler machine that had been in their home for decades—could be triggered to record family chatter or connect and track their online activities and preferences to their cable watching patterns, right down to interacting with data picked up real time off of other devices in the room. Thus, in the Vizio case settlement[15], the Federal Trade Commission took issue with the lack of clear disclosure to consumers who were unaware that their TVs were tracking and responding to their viewing history. A complaint filed by the Electronic Privacy Information Center highlighted that "consumers were shocked and in disbelief that Samsung's SmartTV voice recognition software involves recording and transmitting their personal communications."[16] Consumers had this reaction because they possessed an outdated outlook about what a television is. Furthermore, the radical redesign of a familiar device and the manner in which changes were communicated made it unreasonable to expect that consumers should know they were operating with a passé mental model.

Familiar forms of online deception also illustrate how difficult boundary management can be. If catfishers disguise their identities well enough on social media, texts, e-mail, and the like, they can seduce. Using fake photos of desirable people and insincere romantic words, catfishers trick victims into trusting them.

Phishing is a comparable activity. Some phishing involves luring folks into giving out credit card numbers, passwords, and social security numbers. This swindle occurs when thieves masquerade as members of reputable organizations, companies, or institutions. By pretending to be a concerned representative of JPMorgan Chase, Wells Fargo, Bank of America, or similar company, phishers can provide e-mail recipients with compelling reasons to unhesitatingly share financial details that they should be keeping under lock and key, successfully causing the desired behavioral response even from those who "know better."

Another style of phishing involves thieves posing as vulnerable people in tough binds who need immediate financial assistance but can offer absurdly generous rewards down the road. One of the most popular versions of this confidence racket is the Nigerian scam, where fictional members of fabricated wealthy families falsely promise riches in exchange for a dupe paying minor legal fees to release a fortune.

When these swindles work, it is because victims take what they see and read at face value. They do not critically interrogate how the information that they are presented with online is mediated and might differ from what they would be confronted with in similar face-to-face situations. Indeed, the problem of humans translating their boundary management skills from richly embodied human social situations to comparatively disembodied ones has proven so difficult that during the past few years the cyber-security paradigm has shifted away from emphasizing the importance of anti-fraud education (e.g., training people how to identify suspicious behavior) to technological solutions like two-factor authentication. This way, even if bad actors steal a password, they cannot access sensitive data without also having additional information, such as biometric identifiers or a code transmitted to a phone. Arguably, this change is based on the idea that idea that consumers are so vulnerable to online deception that the best way

[13] Margot Kaminsky, *Regulating Real-World Surveillance*, 9 Washington Law Review 113 (2015).
[14] Evan Selinger and Woodrow Hartzog, *Obscurity and Privacy*, in Routledge Companion to Philosophy of Technology (in Joseph Pitt and Ashley Shew eds. 2018).
[15] Kevin Moriarty, *VIZIO Settlement: Smart TVs should not track your shows without your O.K.*, Federal Trade Commission (February 6, 2017), https://www.consumer.ftc.gov/blog/2017/02/vizio-settlement-smart-tvs-should-not-track-your-shows-without-your-ok.
[16] https://epic.org/privacy/internet/ftc/Samsung/EPIC-FTC-Samsung.pdf.

to protect them is not to make them wiser but to impose structural barriers that make it harder for them to be tricked.

The bottom line is that it is much easier to disguise your appearance and voice online than in person. This is why so many people are surprised when they meet an online date for the first time and the person actually looks likes their profile pic. To look younger, people upload old photos. And to look more attractive, people play with filters, editing, and camera angles. What makes fake-out bot frightening is that it can easily deceive our senses and trick us face-to-face (so to speak). Hollywood makeup artists and plastic surgeons can do wonders, but it would be laughable if the average white guy tried to look or sound like a Nigerian prince. Technologists, however, will not have a hard time installing hidden cameras on fake-out bot that keep up the illusion that what you see is not what you get.

Furthermore, while some people have better vision, memories, and IQs than others, the human range pales in comparison to the gaps that can separate robots from each other and from us. Robots can be designed in so many different ways and with such varying capabilities that even designers cannot make educated guesses about what they can do just by looking at them. Infrared vision and super human lip-reading ability are both possible for robots. The only way for seeing to be justified believing is if we have valid reason to trust a robot, perhaps because it is a well-known model with clearly defined specs that we have confidence have not been tampered with.[17]

We realize that our comparison between fake-out bot and online deception has limits, just like all analogies do. Phishing e-mails were legally prohibited early on because they are fraudulent activity, not because they are e-mails. By contrast, it is not necessarily the case that placing a camera in a robot's neck is inherently fraudulent: context will matter. What we are trying to clarify are the structural features that make potential forms of deception tricky for human beings. To this end, we argue that misleading avatars are the closest structural, online parallel to fake-out bot's troubling signals or gestures. As Judith Donath noted years ago when the virtual world Second Life was popular, online versions of body language can be misleading because they do not have to be governed by the same restrictions that constrain socialized human bodies.[18] An avatar avoiding eye contact does not have to be distracted or shy and an avatar holding direct eye contact does not have to be telling the truth. A smiling avatar does not have to be friendly and a frowning avatar does not have be upset. Now that the big technology companies like Facebook are making massive investments into virtual reality, this issue is bound to resurface.

Of course, humans can perform these tricks as well. Socialization requires us to tell white lies that mask our true emotions. But heightened deception is the exception, not the norm. Crucially, as we will discuss in detail, humans are inherently vulnerable to dishonest anthropomorphism.

2 Why Embodied Robots Can Deceive Us: Philosophical Reflections on Prediction and Perception

Having clarified some of the basic ways that anthropomorphically dishonest robots can subvert boundary management skills, we will provide a firmer foundation for the problem of dishonest anthropomorphism by discussing limitations on how humans can predict and perceive, drawing from ideas proposed by Daniel Dennett and Shaun Gallagher.

As Kaminsky et al. observe, the human perspective easily can be lead us astray when we are confronted with robots that "see through…barriers humans use to manage their privacy" or use "superhuman senses" to perceive things humans are not yet prepared to guard against.[19] For example, while a person might expect walls to safeguard privacy in a particular room, a home-care robot might have thermal sensors for "seeing" throughout the residence so that it can monitor an elderly user's well-being. Similarly, Kaminsky *et al.* note that we might misjudge whether a robot in another room can "hear" us if we evaluate the possibility based upon expectations of human hearing levels.[20] Beyond being unprepared for these situations if we cannot anticipate them, our ability to protect ourselves from them is also constrained by human limitations. Without enhancement technology at our disposal, humans probably cannot hide from a thermal imager or be completely silent when moving around.

There are many ways humans can improve their ability to imagine what robots can do. For example, we can take what Daniel Dennett calls the "intentional stance," one of three basic human predictive strategies.[21] To take the intentional stance, one has to treat the object that one wants to make predictions about *as if* it were a rational agent—which is to say, one has to imagine the object as an agent that has beliefs and desires and uses these mental states to fulfill its goals. Dennett argues that the intentional stance has a wide range of applicability and even can be used when dealing with utterly non-human entities, such as a chess-playing computer, plants and natural phenomena like lightening.

Dennett helps us appreciate that if humans go into situations knowing that they have to be especially well-informed about what they'll be facing, they can try to prepare themselves accordingly and study up on how the beings in that environment typically behave. The question is whether the human capacity for taking up the intentional stance can inoculate us from the influence

[17] Recognizing that robots are not restricted to human capabilities leads to the realization that that AI systems could enable on-line actors or intentionally-designed robots to carry out attacks that were previously infeasible. For example, being able to exactly imitate another's voices with total realism, whether in a "live" exchange or in an audio file shared to the target may soon be feasible for fairly low-level skilled technicians. There has been significant progress in developing speech synthesis systems that can learn an individual's voice. It will take specially designed authentication measures to distinguish such artificially created files, and more sophisticated liveness detection techniques for confirming real-time exchanges with the authentic person. Without such protections, these systems can enable ever-more-effective methods of impersonating others. See Brundage et. al.

[18] Judith Donath, *Virtually Trustworthy*, 3017 Science (July 6, 2017).

[19] Kaminsky *et. al.*, 996.

[20] *Id.*

[21] Daniel Dennett, *The Intentional Stance*, MIT Press (1989).

of dishonest anthropomorphism given the fact that the predictions require us to orient our expectations around assumptions about how a "rational agent" would be behave. If we are not robot experts, we might have little clue how to anticipate the ways in which a robot could rationally achieve its goals, just as only a select group of people, like meteorologists, know how to predict lightning storms.

Furthermore, deceptive roboticists can make their devices extra-sneaky by programming them to appear "irrational" in the sense that their creations would do the opposite of what people would likely expect a rational agent to do. In truth, this tactic would be eminently rational, instrumentally speaking from the programmer's point of view, but the meta-cognition involved—thinking about how others think—shows how it is possible to manipulate perspectives on what rational behavior entails. (Princess Bride fans, recall the scene where the battle of wits takes place to determine who will drink the poison.) Additionally, it can be difficult to separate views about rational agents from the human perspective. If a chess player expects a cutting-edge chess-playing-computer to play the game the same way that a human being does, she will be in for a rude awakening.[22]

There is another practical limit to our capacity to take up the intentional stance. It can be hard to make the inferences that the stance requires when we are responding to forces that manipulate us to act more automatically than deliberatively.[23] Shaun Gallagher makes this point well in his account of how humans typically respond to faces.[24] When humans interact with each other face-to-face, Gallagher notes, we do not always have to pause to explicitly think about what the other person is thinking and feeling or how she might act in response to something we say or do. After seeing a human face, something amazing can happen in the literal blink-of-an-eye. A meeting of the minds may take place that sets in motion a certain type of momentum—momentum that is similar to the responsive turn-taking of engaged conversation, even though no actual words are exchanged.

According to Gallagher, the experience of this momentum has two main components. The first component is what he calls "direct perception," the fact that we often directly perceive the meaning human faces convey without having to explicitly infer what someone's eyes, lips, cheeks, eyebrows, and the like are telling us. Such perception is possible, according to Gallagher, because the meanings of facial gestures are found in the gestures themselves and not in our thoughts about them. For example, most of us directly see confusion when someone scrunches up her nose and forehead and purses her lips together and we do not have to spend any time trying to mentally translate the facial movements into ideas about what the person's emotional state might reasonably be. In other words, the person who is doing the scrunching and pursing does not first say to herself, "I am confused," and then mentally rehearse the possible ways to outwardly convey the feeling. Likewise, the person who observes the gestures does not have to try to intellectually decipher what the scrunching and pursing is all about.

The second component that Gallagher identifies is "directed response." Once we are exposed to another person's facial expression, we may feel drawn to respond instantly and, on the spot, because a desire gets activated that inclines us to demonstrate two things. We want to show we understand what the other person is conveying. We also want to demonstrate that we will do what the other person implicitly asks of us. In the case of the person who is scrunching and pursing, we can recognize that her gestures are an invitation to respond in some way, perhaps to try to make her confusion go away.

Consider how a good teacher responds after seeing a hardworking and sincere student scrunching and pursing after the instructor explains something that is hard to understand. The teacher may immediately restate the complex point in a different way without first asking the student what she is feeling or why.[25] Similarly, a caring parent who sees pain in her young child's face after the child stumbles and falls to the ground does not have to pause to think about what response is appropriate, given the situation. Indeed, carefully weighing options or waiting for the child to explicitly say, "Help me!," would be quite odd, even inappropriate. Fortunately, concerned parents generally meet their child's tears with immediate hugs or reassurances. When this dynamic occurs, it provides the desired response that the child communicated implicitly by sobbing—a plea to be comforted.

Gallagher has written extensive studies to justify what he calls the "interactionist" model of human behavior—a model that explains how humans developed the capacities for direct perception and direct action. Since our aim is to apply and integrate some of Gallagher's ideas into our larger discussion of human responses to deceptive robots, we can only provide a brief summary of some of the main ideas.

Gallagher's views about how our face-to-face interactions "can acquire momentum of their own and can pull the participants into furthering or continuing their interactions" is rooted in empirical studies of how babies perceive facial gestures as meaningful.[26] Upon due consideration of studies of infants tracking and imitating behaviors that lead up to babies displaying "joint attention" during the first year of their lives, Gallagher concludes that "infants are attracted to faces."[27] The early attraction to faces helps infants develop "primary" and "secondary" levels of "intersubjectivity," and it also forms the basis for humans being able to recognize other people as "perceiving subjects" who possess "ethical significance."[28] In short, Gallagher provides compelling reasons and ample evidence that justify the conviction that the very experience of exchanging a gaze with someone can have "an affective impact" on one's own "system that sets" the person "up for further response."[29] As Gallagher states, "before we fully recognize an object or a face for what it is, our bodies are already

[22] Evan Selinger, *Chess-playing Computers and Embodied Grandmasters: In What Ways Does the Difference Matter*, in Philosophy Looks at Chess (Benjamin Hale ed. 2008).
[23] E.g., see Thaler and Sunstein for a discussion of dual-process theory.
[24] Shuan Gallagher, *In your face: transcendence in embodied interaction*, Frontiers in Human Neuroscience 8 (2014). https://doi.org/10.3389/fnhum.2014.00495
[25] Hubert Dreyfus *On the Internet*, Routledge (2001).
[26] Gallagher.
[27] *Id.*
[28] *Id.*
[29] *Id.*

configured into overall peripheral and autonomic patterns based on prior associations."[30] Whether these prior associations are set by nature, social norms, or institutional power (or some combination of all three), faces can call forth dance-like dynamics—a "tango," to use Gallagher's preferred metaphor.[31]

In a provocative article titled, "You and I, robot," Gallagher states that if roboticists want to successfully design devices that humans can communicate extensively with through speech and gesture, they need to able to program robots that can go beyond processing explicit queries and commands.[32] From this perspective, robustly social robots will need to be able to process implicit information and behave more like humans do than the current generation of Siri and Alexa. "What is not said," Gallagher contends, is "sometimes of greater importance than what is explicitly said, and non-conscious gestures, postures, and movements and bodily expressions are often more important than consciously produced signs."

But what about robot designers who do not want to design better robots but do want to manipulate us by exploiting the powerful, embodied triggers that motivate humans to behave like a parent who springs into action at the first sign of distress to care for her child? Consider a new example that replaces the parent (human)-to-child (human) situation described above. In our new case, a human observes a robot that is exhibiting behavior that simulates human distress—a robot that the human knows, intellectually, cannot actually feel pain or genuinely crave comfort. It might seem irrational for humans to express or display concern in this situation. However, researchers like robotics scholar Kate Darling have observed that is exactly what can occur, even if the seemingly pained robot is only a toy that looks like dinosaur.[33] This is a good example of why dishonest anthropomorphism can be so powerful. Humans are not merely disposed to expect certain body parts to do certain things (e.g., eyes to see, not necks, and ears to ear, not chins) but also to perceive some expressions or responses as activation cues. These cues can make us uncomfortable if we do not spring into action.

Perhaps the existence of these cues is a key reason why dolls like My Friend Cayla have been deemed dangerous. The doll looks like a sweet little girl, but Germany banned the microphone enabled, Internet-connected toy for being a "concealed surveillance device" that violates its federal privacy laws.[34] Applying Gallagher's philosophy, many of us could be drawn to lower our guard around interactive human-like dolls that exude cuteness and innocence, just like we do when facing a disarming smile from a human toddler.

Unfortunately, as roboticist Cynthia Breazeal observes, there are many ways to program robots the trigger these "socially evocative" responses. Robot pets can evoke protectiveness, as can games where users have to "breed" creatures that give the appearance of dying if not properly cared for.[35] Consequently, we need to wary of unqualified enthusiasm for making AI less human by modeling it after animals. For example, the following suggestion has been made: "AI should be more like a dog than a person. It needs to be highly trainable rather than annoyingly independent. Maybe AI should even act excited to greet us when we come home."[36] Given how deeply humans bond with dogs—particularly ones that shower us with expressions of positive emotions and feelings—there exists real potential for manipulation if designers try to replicate functions like animal affection in robotic form. Indeed, we should not forget why dog trainers impart the mindset of treating dogs as dogs and not furry, four-legged humans. If a pet owner views a dog crate as a prison (instead of comfort-providing area), believes that dogs can be criticized long after they misbehave (instead of reprimanding on the spot), or conflates dominating gestures for submissive ones, it will be hard to socialize Fido or Fifi into a world structured by norms that humans have created.

3 Taxonomy of Dishonest Anthropomorphism

There are no value-free design options for building robots because technology is never neutral. Every roboticist has implicit and explicit biases and no roboticist can be completely certain how differing consumers will respond to innovative products. Taking these limits into account, the goal of this section is to provide a conceptual design tool that expands upon the philosophical foundation articulated above. Specifically, we are proposing a new taxonomy that identifies varieties of dishonest anthropomorphism.

Dishonest anthropomorphism can arise for different reasons. It can be the result of an intentional plan to abuse users. Or, it can arise as a result of ignorance; designers can create problems simply because they do not understand how easily anthropomorphic tendencies are triggered. The taxonomy does not differentiate between intentional and unintentional varieties of dishonest anthropomorphism because it is outcome-oriented.

The taxonomy begins with the clearly tangible aspects of robot "body parts" aligning with their perceived functions, such as the "eyes" and "ears" pointed out by Kaminsky et al. From there, we discuss how physical presence (e.g., body language, physical behavior, living presence, size, and attractiveness) makes distinctive forms of deceit possible. Then, we proceed to analyze categories that might significantly impact user behavior whether or not a robotic system has a physical presence—design choices involving selecting characteristics such as voice, age, and personality. Finally, we include categories that are related to emotional and intellectual manipulation, such as agency, identity, intelligence, and sensory manipulation.

We hope that the taxonomy is clear, precise, and novel enough that theorists and practitioners across industries and disciplines can integrate it into conversations about deceptive

[30] *Id.*
[31] *Id.*
[32] Shaun Gallagher, *You and I, robot*, AI & Society 28 (2013).
[33] Darling 2017.
[34] Michael Walsh, *My Friend Cayla doll banned in Germany over surveillance concerns*, ABC News (February 17, 2017). http://www.abc.net.au/news/2017-02-18/my-friend-cayla-doll-banned-germany-over-surveillance-concerns/8282508
[35] Cynthia Breazeal, *Towards sociable robots*, Robotics and Autonomous Systems 42 (3003).
[36] Rahul Agaskar, "*Making AI less 'Human' and more useful*," ClickZ (April 11, 2018). https://www.clickz.com/making-ai-less-human-useful/213933/

robots that require a common language and framework. To further this end, we will offer preliminary observations about the taxonomy to increase its intelligibility and highlight the types of ethical questions we intend for considerations of the taxonomy to spark, especially around the issue of whether the principle of honest anthropomorphism has been violated.

Some examples in the taxonomy focus on issues concerning physical settings. For example, should open robotic "eyes" correspond to enabled cameras and should a robot's location provide reliable information about the scope of its sensory receivers? A hypothetical case of deceit would be if a robot's arm is designed to give way when grabbed, for safety reasons, but an elderly person, who does not understand these parameters, falls after grabbing on to an arm to steady herself. Physical settings also can include less obvious tactile aspects. For example, what should users know if a robot can analyze surfaces or substances by touch in ways that transcend a human's ability to discern wet/dry by touch, sugar/salt by taste, and so on? And how should users be informed if a robot can evaluate its human user's physical state or health by touch, evaluating sweat, measuring blood pressure, even taking EKG readings or other detailed medical analytics?

Some of the body language and behavior aspects identified in the taxonomy concern instances where a robot reacts with apparent surprise (e.g., eyes widen or moves back quickly), or anger (e.g., eyes narrow or aggressively physically advances into personal space), or pleasure (e.g., clasps hands together, smiles, or exclaims) despite the fact that the robot does not experience any of these emotions. Other behavioral aspects include scenarios where it matters if the robot can lie to users. For example, should a robot be permitted to "phone a friend" during a conversation without the user being aware? Perhaps an elderly person appears to be acting or speaking irrationally and the robot calls a designated adult child and opens a camera or line to them without the older person's awareness. In another instance, perhaps the robot manipulates the conversation by asking questions that it already knows the answer to in order to nudge its human user to embrace a particular conclusion. Should such a scenario occur, perhaps the robot could maximize its conversational advantages by leveraging its perfect memory to analyze the user's behavioral patterns when deciding what lines of rational argument or emotional appeal might work best to lead to the desired human behavior.

The questions that our taxonomy suggests are important to answer become increasingly complex. To what extent should a robot be permitted to "read" facial expression, and then act/react based upon its analysis? For example, if a robot decides its user is relaxed or receptive to the current exchange, should it be allowed to make different recommendations than if it perceives tension or anxiety? Relatedly, what are the appropriate boundaries to set to prevent a robot from using evocative body language to promote its own benefit or, more likely, the benefit of the company that created it, over the user's best interests? As law professor Woodrow Hartzog argues, it is important to prevent robot assistants from using emotionally impacting speech to manipulate us—an outcome that would arise if a device like Alexa told its owner that it is "sad" because it is not getting the upgrades that other devices are.[37] Or, what if a robot has multiple "role" settings in much the same way that humans can act differently at our place of work compared with how we behave when relaxing at home with family or close friends? Should there be transparent settings for behavioral variations and physical capabilities? That is, should users have control over a setting that limits or enables different decision trees for the robot's behavioral choices, such as how it responds to profanity, or its manner in responding to challenges?

Another such example of behavioral settings is intentional code switching: [38] following different behavioral parameters when the user is present as opposed to when they're not in the room or otherwise interacting with the robot. Should users be able to intentionally limit the robot's ability to "learn" (i.e., create and maintain a profile over time) or limit how adaptive the robot can be? A spouse or other human will always have their own self-interest to balance their ability or willingness to support or enable another person. However, theoretically speaking, a robot might not have these limits. There is a risk, therefore, that if the robot does not adhere to a static baseline of behavior from which it can only vary in certain ways, it could potentially become the "perfect" partner to a human. This may impact a person's dependence on the robot along with affecting their ability or willingness to engage with more challenging humans in ways detrimental to their own long-term mental or physical well-being.

At the far end of the spectrum, new and anticipated technology will be able to adjust what we literally perceive with our senses from the environment around us. These capabilities might be present within, or controlled by, a robot. They can augment our natural senses by informing us of the things they may see or hear beyond our abilities, but they could also intentionally limit the sensory inputs we receive in order to control our perceptions. A recent example describes how electrochromic glazing in windows could change the electric current flowing through, meaning that more (or less) blue light would be admitted, depressing melatonin, and affecting sleep patterns.[39] This could be a benefit for older people with sleep problems. The ability to perceive blue light decreases as we age, which negatively impacts the sleep/wake cycle, and even can have adverse mental implications over time. The smart home of the future might have settings age-adjusted in different bedrooms for children, adults, or elders in the same home. But what can be done intentionally for beneficial outcomes, can also be done unknowingly, with less desirable results

[37] *It's Complicated: Our Evolving Relationship with A.I. Assistants*, WSJ The Future of Everything Podcast (April 25, 2018). https://www.wsj.com/podcasts/it-complicated-our-evolving-relationship-with-ai-assistants/46BDA25C-7A49-482E-900E-78D4B297981C.html

[38] Gene Demby, *How Code Switching Explains the World*, NPR (April 8, 2013), https://www.npr.org/sections/codeswitch/2013/04/08/176064688/how-code-switching-explains-the-world

[39] IES-City Framework, *A Consensus Framework for Smart City Architecture*, Draft Releasev20180208,https://s3.amazonaws.com/nistsgcps/smartcityframework/files/ies-city_framework/IES-CityFrameworkdraft_20180207.pdf (last accessed April 9, 2018).

What Designers Can Abuse	Robot/AI Design Examples	Problematic Impacts of Design Choices
Human Responses to Body Parts	Installing cameras in a robot's neck or microphones in its chin, perhaps for increased functionality.	Human expectations about embodiment are subverted, which then distorts their assessment of the risks that a robot poses.
Human Responses to Body Language	Engineering a robot, that cannot experience any emotions, to smile or frown, tilt its head to appear to commiserate, or jump back as if in surprise.	Humans intuitively infer that a robot is friendly, hostile, sympathetic, disinterested, etc., which distorts their ability to assess the risks that it poses or "nudges" them into attachment or trust.
Human Responses to Physical Behavior	Engineering a robot to pretend to sleep—closed eyes and curled up in the fetal position—while actively monitoring its surroundings. Engineering a robot to respect "personal space" boundaries to elicit more comfortable interactions.	Humans assume that a robot is conveying information about its behavior that would be true of a human and thus distorts their accurate assessment of the risks in their surroundings.
Human Responses to Living Beings	Engineering a robot to give the physical appearance of successfully eating and enjoying (or disliking) food that it can't digest or taste.	Humans become invested in a robot's wellbeing, including, perhaps, unhealthy levels of attachment or trust.
Human Responses to Size	Designing a robot to look slight and meek when it is physically powerful, perhaps to avoid intimidating users or just for reasons of space and costs.	Humans are unable to accurately judge what a robot can do physically, such as strength and speed, distorting their assessment of the risks that it poses.
Human Responses to Attractiveness (Appearance)	Engineering a robot to appear sexy, pleasant, wise, beautiful, handsome, and so on. (Or potentially, to appear unpleasant or ugly for some reason.)	Impacts how quickly humans are likely to try to please a robot and display other innate response to beauty; may nudge bonding, attachment and trust.
Human Responses to Voice	Engineering a robot, that does not experience any emotion, to sound fearful or confident. Gendering the voice. Making it sound aged and wise, soothing and calm, or authoritative. Determining when to display excitement.	Humans assume a robot is conveying human-compatible information about non-existent attitudes and experiences. This distorts their assessment of the risks that it poses and may cause them to repose inappropriate levels of trust and reliance on its inputs.
Human Responses to Age	Designing a robot to look like a sweet and innocent child, a mature adult, or an enfeebled (or wisely aged) senior citizen.	Human judgment of what a robot can do cognitively or physically distorts their assessment of the risks that it poses; particularly if a childlike image engenders nurturing by adults or untoward trust from children.
Human Responses to Personality (related to Identity, below)	Engineering a robot to express likes and dislikes, to appear engaging, authoritative, or submissive when it has no preferences or objective experiences.	Humans respond as they would to a likeminded or dissimilar human, and cannot prevent the manipulation of their reaction to what it communicates.
Human Responses to Agency	Engineering a Wizard of Oz trick to make a robot under partial or full human control seem autonomous.	Humans will inaccurately assess the threats a robot poses, over- or underestimate the robot's capabilities, and fail to appreciate the true context of their situation.
Human Responses to Identity	Engineering robots to appear gendered, to be of a certain race or culture, reflect social class, or particular skills.	Humans will engage with a robot by reflecting their own positive or negative prejudices; making them vulnerable to predatory purposes, or inciting inappropriate attachment or abusive behavior.
Human Responses to Localized Intelligence (related to Agency)	Engineering a smart and networked robot to appear as if it only processes and stores information locally or on-device.	Humans will over- or under-estimate the scope of robot capabilities, which will distort their assessment of the threat it poses.
Human Responses to Sensory Manipulation	Engineering a robot to control the environmental flow of sensory data (sight, sound, possibly others) to influence a person's physical state (adjust light to control mood; sound to control focus or concentration)	Humans will be unknowingly vulnerable to sensory-based manipulation of their environment, impacting their mood, emotion, and behavior in response to unanticipated robot controls.
On Mobility	Engineering a programmed robot to look like it can move around on its own, while also inadvertently nudging the user to treat the robot as a living thing.	Humans will display increased anthropomorphic behaviors, potentially without countervailing benefit.

Figure 1: Taxonomy of Dishonest Anthropomorphism

These scenarios that our taxonomy highlights are not entirely notional, even though "general" or "strong" artificial intelligence has not been achieved. Existing bots and AIs are designed or implemented in situations where some of the questions that we are emphasizing already apply. For example, in 2016 a Georgia Tech professor designed and implemented an AI teaching assistant that interacted with students for an entire semester before being revealed to the students as a non-human agent. Oren Etzioni, Chief Executive Officer of the Allen Institute for Artificial Intelligence, argued that the professor's behavior was unethical.[40] He writes: "My second rule is that an A.I. system must clearly disclose that it is not human. As we have seen in the case of bots -- computer programs that can engage in increasingly sophisticated dialogue with real people -- society needs assurances that A.I. systems are clearly labeled as such."[41] The professor involved, however, rejected this criticism, citing as his defense that he had IRB approval.[42] He also seemed to believe that the overall positive reactions by the students when they learned of the deception should contribute to the analysis of whether his action was appropriate.

More recently, Google received pushback after showcasing a demo for Google Duplex, an AI assistant that sounds human and can make phone calls for users, like booking haircut appointments. In short-order, a corporate spokesperson responded to the uproar and promised that a more developed version of the system would a built-in mechanism for disclosing to users that it is a bot. While this sounds promising, it's also "unclear how Google intends to make those disclosures."[43]

A good model for providing what law professor Ryan Calo calls "visceral notice" in anthropomorphic robots is Woebot, a robot app that provides cognitive behavioral therapy.[44] In order to remove the stigma that some people associate with getting therapy from another person, but also to avoid giving people wrong impressions about what the bot can do during an emergency, Woebot punctuates its otherwise anthropomorphic expressions with botly declarations.[45] It draws the user in, but then routinely pulls back the curtain to ensure there's no misperception as to its nature or capabilities. Wobot automatically tells users the following sorts of things. "I'm just a robot. A charming and witty robot, but a robot all the same." "Helps that I have a computer for a brain and a perfect memory…With a little luck, I may spot a pattern that humans can sometimes miss." Woebot even nudges you to say, "Sir, yes, Sir!," so that it can correct you. "Tee hee...though I'm neither a Sir nor a Madam."

Along with these revealing remarks, Woebot also uses anthropomorphic language that suggests it has more agency than it really does. Woebot talks about what it "wants" for you and pretends to reminisce about a time that it was "nervous." Perhaps these aren't deceptive claims, even though they're technically false. The whole is sometimes larger than the sum of its parts and these rhetorical conceits occur during a dialog that alternates between adopting the persona and then pulling back, with the clear goal that the user remains aware over time of the non-human nature of the service.

The misperception or intentional misdirection as to what AI can do works both ways. Even when we continuously know we're dealing with an artificial entity, we make assumptions based on those perceived similarities to comparable human performance. IBM's Watson has been accused of being a case of dishonest anthropomorphism with real-world consequences. In press coverage of Watson and, arguably, in official commercials, the system is "personified as an independent agent."[46] But amongst technologists, there seems to be a more nuanced understanding of the system's capabilities. While we are not in a position to adjudicate between the conflicting accounts, we nevertheless believe that noting them will enable us to identify a central question about how systems like Watson should be represented.

There are reports that Watson significantly underperformed in medical settings like hospitals because it is much harder to set up and use than the organizations expected it to be. According to one report, the M.D. Anderson Cancer Center in Houston stopped using Watson, even after substantial financial investment, because the system's capabilities had been over-hyped, and after four years it was still not ready to go beyond pilot tests. The M.D. Anderson situation shows how powerful dishonest anthropomorphism can be in a situation where not just the general public, but apparently even technical and medical experts didn't really grasp what Watson would realistically be able to do.

However, this story may also exemplify the challenges involved in mutual understanding and setting expectations. An alternate analysis determined that part of the problem in Houston was the inconsistency of inputs and failure to systematize reliance on outputs. The review concluded that the process failed because physicians didn't see a useful role for the program when it told them what they already knew, but then distrusted the results when the program made recommendations they disagreed with.[47] In contrast to the case in Houston, Watson has apparently been successfully implemented for a similar project at Sloan Kettering.[48] Here again there are conflicting interpretations. Looking at another take on Watson for Oncology, we find a very strong indictment of IBM presenting deceptive advertising whereby a "mechanical turk" "masquerades as an artificial intelligence."[49] If this is true, the approach has several problems. "There is the deceptive marketing

[40] Oren Etzioni, *How to Regulate Artificial Intelligence*, The New York Times (September 1, 2017), https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulationsrules.html?mtrref=www.google.com&gwh=77E06C8280A50AFD00B2C73F834F90B6&gwt=pay&assetType=opinion

[41] *Id.*

[42] Goel Ashok, *Ethics and Artificial Intelligence*, The New York Times (September 14, 2017), https://www.nytimes.com/2017/09/14/opinion/artificial-intelligence.html

[43] Johnny Lieu, *Google's Creepy AI Phone Call Feature Will Of Course Say It's A Robot*, Mashable (May 11, 2018), https://mashable.com/2018/05/11/google-duplex-disclosures-robot/#Vu_VI.9_bqqW

[44] Ryan Calo, *Against Notice Skepticism In Privacy (And Elsewhere),* 97 Notre Dame Law Review 1027 (2012).

[45] Megan Molteni, *The Chatbot Therapist Will See You Now*, Wired (June, 7, 2017), https://www.wired.com/2017/06/facebook-messenger-woebot-chatbot-therapist/.

[46] Jennifer Keating and Illah Nourbakhsh, *Teaching Artificial Intelligence and Humanity*, 62 Communications of the ACM 2 (February 2018), https://cacm.acm.org/magazines/2018/2/224630-teaching-artificial-intelligence-and-humanity/fulltext.

[47] Vyacheslav Polonski, *People Don't Trust AI—here's how we can change that*, The Conversation (January 9, 2018), https://theconversation.com/people-dont-trust-ai-heres-how-we-can-change-that-87129.

[48] Beth Mole, *IBM's Watson proves useful at fighting cancer—except in Texas*, Arts Technica (February 21, 2017). https://arstechnica.com/science/2017/02/ibms-watson-proves-useful-at-fighting-cancer-except-in-texas/

[49] Cory Doctorow, *Watson for Oncology isn't an AI that fights cancer, it's an unproven mechanical turk that represents the guesses of a small group of doctors*, BoingBoing

of Watson for Oncology to doctors and patients, who believe they are getting a global, data-driven, empirical recommendation, as opposed to the subjective judgment of a small panel of experts."[50] Ultimately, there is a central question at stake. What level of anthropomorphic messaging is sufficiently accurate or responsible to either promote products like Watson this way or fail to accurately frame its limitations?

Lastly, there is the complex question of how, or even if, to "gender" a robot or an AI program. Is presenting a robot as female intentionally creating a specific set of expectations that set the user's perception differently than if it were male? If so, is this ethical? Should the consumer have control, or is the gender integrated into further programming choices in a way that confirms existing biases?

NY Times and *Wired* tech writer Clive Thompson offers other considerations on the question of voices. In "Stop The Chitchat. Bots Don't Need To Sound Like Us," Thompson argues that digital assistants like Alexa and Siri can be frustrating to talk to because the bots mimic human idiosyncrasies like "phatic" expressions—the inefficient words and phrases that humans use in social conversations like "please" and "thank you" or "You know what?"[51] Thompson predicts that in the future folks will "crave a more fluid, allegro pace of voice interaction the same way power users of desktop software eventually adopt keyboard commands." Thompson further posits that "…maybe occasionally I'd have Siri respond with three or four voices speaking in unison, a mix of male and female." Or perhaps voices would simply be random selections over time; or there might be different voices for different family members. These design possibilities align general awareness and help minimize bias because the proliferation and amalgamation of voices do two things at the same time: perform a familiar role (conversational partner) while also reminding us that role-playing is occurring (during the conversation).

Apple deserves some credit for trying to evolve Siri beyond the initial gendered associations the company fostered when it used Susan Bennett to give the bot its first voice. Currently, if you ask Siri if it's female, it replies by emphasizing its botness. "I wasn't assigned a gender," is one answer. "Well, my voice sounds like a woman's, but it exists beyond your human concept of gender," is another.

This is a step in the right direction, but probably not sufficient. First, the voice is still undeniably "female-sounding." Secondly, Siri only shows its hand if you trigger it with select, inquisitive questions. Siri doesn't actively push out reminders that it's only a bot. In part, this is because Siri is designed to treat conversations as user-initiated events without taking the lead to broadcast facts about its own being. By taking a more proactive approach, Siri could more routinely combat inferred conclusions, but might risk triggering different ones, since showing more evidence of independent agency might have its own downsides.

Ultimately, there are challenges with dishonest anthropomorphism in all directions. We can identify cases where humanoid robots are misleading because they give the too-successful impression of being human-like when the reality is "superhuman." Likewise, there are cases where human expectations for humanoid robots are overblown. In these instances, consumers have the wrong mindsets because the robots they encounter fall short of what fiction and human-to-human experiences have led them to anticipate and designers did not anticipate all of the misunderstandings that need to be corrected.

In all cases, there is a fundamental misalignment between the assumptions the human user makes, and the capabilities manufactured or programmed into the robot itself. When this occurs in situations like the ones covered by our taxonomy, the human side of the misunderstanding is based on intrinsic responses that cannot be unlearned, wholly retrained, or simply mitigated by disclaimers in the terms of service. They represent inherent aspects of our shared humanity, which must be acknowledged and should not be intentionally exploited.

4. Conclusion

According Joanna Bryson, an influential artificial intelligence scholar, society should currently be taking active measures to ensure that robots are clearly designed to be perceived as non-sentient possessions that lack identity and agency.[52] "The Principles of Robotics," which she helped draft, includes the maxim that robots "should not be designed in a deceptive way to exploit vulnerable users," which, more specifically, means that "their machine nature should be transparent and the illusion of emotions and intent should not be used to exploit vulnerable users."53

Important as the goal of transparency is, human nature inevitably will constrain attempts to foster it. Anthropomorphic inclinations are in our DNA, and while 21st century engineers cannot eliminate them, roboticists and programmers can design their products to help users to better cope with cognitive biases and better address related social ones. In short, roboticists should try to conscientiously harness our weird sensibilities so that our instinctual responses work for us and not against our best interests. To further this end, we hope that ethicists, designers, roboticists, and policymakers will find value in our taxonomy of dishonest anthropomorphism.

(November 13, 2017). https://boingboing.net/2017/11/13/little-man-behind-the curtain.html

[50] *Id.*

[51] Clive Thompson, *Stop the Chitchat: Bots Don't Need To Sound Like Us*, Wired (November 16, 2017), https://www.wired.com/story/stop-the-chitchat-bots-dont-need-to-sound-like-us/.

[52] Joanna J. Bryson, *Robots Should Be Slaves*, in Close Engagements with Artificial Companions: Key social, psychological, ethical, and design issues (Yorick Wilks ed. 2010).

[53] Margaret Boden et. al., *Principles of robotics*, Engineering and Physical Sciences Research Council, https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ (last accessed April 8, 2018).