DOI:10.1145/3241365

Alan R. Wagner, Jason Borenstein, and Ayanna Howard

▸ **Rachelle Hollander,** Column Editor

Computing Ethics
Overtrust in the Robotic Age

A contemporary ethical challenge.

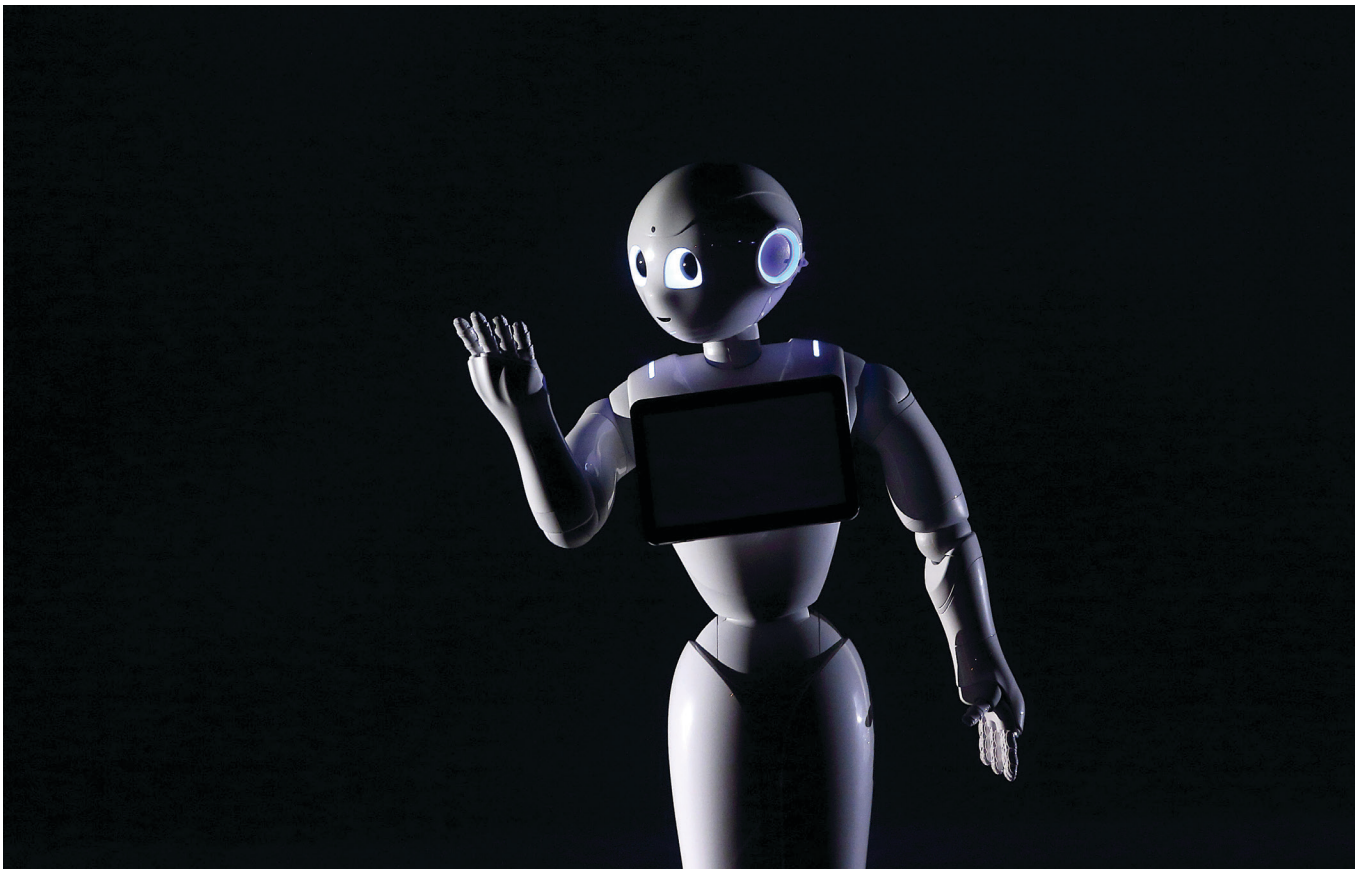
As robots complement or replace human efforts with more regularity, people may assume that the technology can be trusted to perform its function effectively and safely. Yet designers, users, and others must evaluate this assumption in a systematic and ongoing manner. Overtrust of robots describes a situation in which a person misunderstands the risk associated with an action because the person either underestimates the loss associated with a trust violation; underestimates the chance the robot will make such a mistake; or both.

We deliberately use the term "trust" to convey the notion that when interacting with robots, people tend to exhibit similar behaviors and attitudes found in scenarios involving human-human interactions. Placing one's trust in an "intelligent" technology is a growing phenomenon. In a sense, it is a more extreme version of automation bias, which is a tendency of people to defer to automated technology when presented with conflicting information.[6] Early research on this issue focused primarily on autopilots and factory automation.[7] But with advances in AI and the associated potential for significantly more sophisticated robots, humans may increasingly defer to robots. For example, an overarching ethical concern that we have sought to explore in our research is the prospect that children, their parents, and other caregivers might overtrust healthcare robots.[2] In this column, we highlight two other near-term examples where overtrust of robots may become problematic: in emergency situations and in the operation of self-driving cars. We close with some recommendations that may help mitigate overtrust concerns.

Placing one's trust in an "intelligent" technology is a growing phenomenon.

Overtrust of Robots During Emergencies

For some time now, scholars have envisioned how to use robots for search and rescue operations, emergency situation awareness, and disease outbreaks.[10] Two of the authors of this column studied how humans react to robots during emergency evacuations.[9] Specifically, we evaluated whether participants would follow a robot's directions during a simulated but realistic fire emergency; the participants were informed that a robot would guide them to a meeting room where they would receive further instructions. Under different experimental conditions, the robot made increasingly transparent mistakes while guiding the participants. After reaching the meeting room, participants were asked to enter, close the door, and complete a written survey. Unbeknownst to the participants, the hallway leading up to the room was filled with smoke setting off fire alarms and simulating an emergency.



Pepper, a human-like robot developed by Softbank Robotics, is designed to recognize basic human emotions and adapt its behavior accordingly.

PHOTO BY TOMOHIRO OHSUMI/BLOOMBERG VIA GETTY IMAGES

Participants, upon exiting the room, encountered the robot before they could reach a known exit. The robot, in the different experimental conditions, was programmed to direct them in an alternate direction from which they originally came and in conflict with a standard emergency evacuation sign. 95% of the participants either followed the robot's guidance or stood by the robot during the emergency. Some participants even followed the robot's guidance after being told during the initial guidance to the meeting room, "I apologize. The robot is broken again. Please proceed to the room by going down the hallway and taking a left" by one of the study personnel. After the study, most participants stated they followed the robot because they felt the robot knew more than they did, or that it would not or could not be programmed to lead people astray.

Booth and colleagues have examined how overtrust of robots can compromise physical security.[1] Their scenario involved a robot presenting itself at the locked door of a dormitory and asking unsuspecting individuals or groups for assistance in entering the building. The robot was allowed entry about 70% of the time by groups; when the robot was carrying food, individuals granted it access about 80% of the time. Fifteen participants suspected the robot could be a potential bomb threat; yet 13 of the 15 still provided entry to the robot anyway.

Overtrust of Automated Driving Systems

As various versions of autonomous driving systems are being evaluated for near-term deployment, companies like Tesla have created an "autopilot" mode for some models of its cars. Tesla warns users to remain attentive and vigilant while the "autopilot" is in control of a vehicle's operation.[8] Yet this may place drivers in a difficult psychological situation, as their minds might naturally wander and their attentiveness diminish. Videos depicting people sitting in the back seat of their cars while an automated system steers the car at highway speeds abound: these drivers appear to trust the system implicitly, even though this behavior jeopardizes their safety.

People might assume a robot has knowledge it does not possess, viewing a robot's actions, at times mistakenly, as a direct reflection of the intentions of the developer, when in fact the robot may be malfunctioning or users may be misinterpreting its abilities. Even when presented with evidence of a system's bad behavior or failure, such as in the example of the robot guide described earlier, users may still defer to it. If this holds true for autonomous driving systems, drivers may underestimate the likelihood of a crash. Perhaps more importantly, the driver may mischaracterize warnings of an autonomous driving accident. At least one accident involving an autonomous driving system was an all-or-nothing event, in which the autopilot failed to recognize another vehicle and hit it at full speed; the accident resulted in the death of the driver.[3]

Proponents of self-driving cars suggest the accident rate will be substantially lower if human-driven cars are replaced. What remains to be seen is the

qualitative nature of those accidents. Imagine a scenario in which an autonomous car fails to perceive obstacles in its path due to a sensor failure. Such a failure might cause the system to run into, over, and through items until the accumulated damage to the system is so great the car can no longer move. Consider the magnitude of harm if the case involved an autonomous commercial truck driving into and through a shopping mall.

Overtrust influences people to tolerate risks they would not normally accept and may exacerbate problematic behavior such as inattentiveness while driving. The availability of an autopilot may incline people to eat, drink, or watch a movie while sitting behind the wheel, even if the system is incapable of dealing with an emergency should one arise. Parents may send their kids without supervision for a ride to a grandparent's house. These may be reasonable actions if the chances of a driving accident are extremely low. But that is unlikely to be a safe assumption at the present time.

Recommendations for Mitigating Overtrust

As the adoption of robotic technologies increases, methods for mitigating overtrust will require a multifaceted approach beginning with the design process. Since users might not utilize the technology in the ways designers intend, a recommendation to consider, at least in some cases, is to avoid features that may nudge users toward anthropomorphizing robots. Anthropomorphization can induce a false sense of familiarity in users, resulting in the expectation of human-like responses when in fact the associated risk may be much higher.

Mitigating overtrust may require the robot to have the ability to model the behavioral, emotive, and/or attentional state of the person with whom it interacts. For certain types of robots, potentially including some brands of self-driving cars, the system may need the ability to recognize if the user is paying attention or is distracted. Robots entrusted with the safety of human lives may also need to be able to detect certain characteristics about those lives. This can include whether the user is a child, or whether the user has any physical or mental impairment that may increase the risk in the current situation. For example, if a young child is left alone in a self-driving car, the system might need to be diligent and proactive about preventing certain kinds of harms, such as by monitoring the temperature of the interior cabin or warning an adult if the child is left alone for too long.

Overtrust influences people to tolerate risks they would not normally accept and may exacerbate problematic material.

Future systems and contemporary research have begun to focus on robots that recognize and react to human behavioral, emotive, and attentional states. Softbank Robotics, for example, claims that its Pepper robot can recognize emotions and facial expressions and use this information to determine the mood of the person with whom it is interacting.[11] Presumably the same or a similar kind of approach could be applied to high risk situations. Future robots might, and perhaps should, be able to generate information about the person's attentive state and make behavioral predictions. While such predictions can of course be mistaken, this kind of information could be used to detect and, ideally, help prevent overtrust.

Transparency about how robots function is also critical for preventing overtrust. In order for people to be informed users, they need the opportunity to become familiar with the ways in which a robot may fail. DARPA and other entities have made significant investments in research projects (such as Explainable AI) that focus on creating systems that can explain their behavior to people in an understandable way.[4] Applied to autonomous vehicles, for example, the system would be able to warn users of driving situations that it may not be able to handle or has little experience handling.

Overall, we believe that significant research in many areas, including on mental modeling and theory of mind, could confront the problem of overtrust, resulting in robots that are more transparent—allowing people to more fully understand and learn how the technology will behave. Mental modeling research may also provide insight into techniques that facilitate better communication between robots and humans, and thereby allow each party to more accurately calibrate the risks associated with the interaction. For example, an alert could inform human drivers of autonomous vehicles that there is increased uncertainty emerging from an upcoming traffic condition, such as a left-hand turn, and suggest they deactivate the autopilot mode. It is a type of design pathway that some car companies are already exploring.[5] C

References

1. Booth, S. et al. Piggybacking robots: Human-robot overtrust in university dormitory security. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, ACM, New York, NY, USA, 2017, 426–434.
2. Borenstein, J., Howard, A., and Wagner, A.R. Pediatric robotics and ethics: The robot is ready to see you now but should it be trusted?" *Robot Ethics 2.0*, P. Lin, K. Abney, G. Bekey, Eds., Oxford University Press, 2017.
3. Boudette, N.E. Tesla's self-driving system cleared in deadly crash. *The New York Times* (Jan. 2017).
4. Gunning, D. Explainable artificial intelligence (XAI). Defense Advanced Research Projects agency; https://bit.ly/2x2sS3P
5. Lee, T.B. Car companies' vision of a gradual transition to self-driving cars has a big problem. *Vox* (July 5, 2017); https://bit.ly/2tLUAzp
6. Mosier, K.L., Palmer, E.A., and Degani, A. Electronic checklists: Implications for decision making. In *Proceedings of the Human Factors Society 36th Annual Meeting*, Human Factors Society, Santa Monica, CA, 1992, 7–11.
7. Parasuraman, R. and Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Human Factors 39*, 2 (Feb. 1997), 230–253.
8. Plummer, L. Tesla could stop you using autopilot in its cars—But only if you take your hands off the wheel. *Mirror* (Aug. 30, 2016); https://bit.ly/2uJb3D6
9. Robinette, P., Howard, A. and Wagner, A.R. A conceptualizing overtrust in robots: Why do people trust a robot that previously failed? In *Autonomy and Artificial Intelligence*, W. Lawless, R. Mittu, D. Sofge, and S. Russell, Eds., Springer, 2017.
10. Schneider, D. Robin Murphy: Roboticist to the rescue. *IEEE Spectrum* (Feb. 1, 2009); https://bit.ly/2L74Z2a
11. SoftBank Robotics. Who is Pepper?; https://bit.ly/2wZzGzZ

Alan R. Wagner (alan.r.wagner@psu.edu) is an assistant professor in the Department of Aerospace Engineering and a research associate in the Rock Ethics Institute at The Pennsylvania State University, University Park, PA, USA.

Jason Borenstein (borenstein@gatech.edu) is the Director of Graduate Research Ethics Programs and Associate Director of the Center for Ethics and Technology within the School of Public Policy and Office of Graduate Studies at the Georgia Institute of Technology, Atlanta, GA, USA.

Ayanna Howard (ah260@gatech.edu) is Professor and Linda J. and Mark C. Smith Endowed Chair in Bioengineering in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, GA, USA.