DOI:10.1145/3604632

T. Williams et al.

Computing Ethics

Voice in the Machine: Ethical Considerations for Language-Capable Robots

Parsing the promise of language-capable robots.

Language is often viewed as a distinctly human capability, and one at the heart of most human-human interactions. To make human-robots natural and humanlike, roboticists are increasingly developing language-capable robots. In socially assistive contexts, these include tutoring robots that speak with children to guide and encourage them through educational programming, assistive robots that engage in small talk to provide companionship for the elderly, and robots that recommend physical activities and healthy eating. In field contexts, these include robots for search and rescue and space exploration; that accept verbal commands for navigation, exploration, and maintenance tasks; and may verbally ask questions or report on their success or failure.

This emerging trend requires computer scientists and roboticists to attend to new ethical concerns. Not only do language-capable robots share the risks presented by traditional robots, (such as risks to physical safety and risks of exacerbating inequality) and the risks presented by natural language technologies such as smart speakers (such as encoding and perpetuation of hegemonically dominant white heteropatriarchal stereotypes, norms, and biases[6] and climate risks),[1] but they also present fundamentally new and accentuated risks that stem from the confluence of their communicative capability and embodiment. As such, while roboticists have a long history of working to address safety risks, and while computational linguists are increasingly working to address the bias encoded into language models, researchers who hope to work at the intersections of these fields must be aware of the new and accentuated risks—and the responsibility to mitigate them—that arise from that intersection.

Fundamental trust and influence concerns arise for language-capable robots.

In this column, we explore three examples of the unique types of ethical concerns that arise with language-capable robots—influence, identity, and privacy—requiring consideration by researchers, practitioners, and the general public, and needing unique technical—and social—responses. We then use these examples to provide recommendations for roboticists toward designing, developing, and deploying language-capable robot technologies.

Three Illustrative Issues

Trust and Influence. First, fundamental trust and influence concerns arise

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK

for language-capable robots. Mere embodiment promotes trust and compliance, and mere language capabilities promote perceptions of human-likeness and intelligence. This intersection means we are more likely to listen to what robots have to say, even if they are not truly trustworthy or have no true competence in the topics of their conversation.

This creates an inherent risk of overtrust and overreliance on language-capable robots, that will go far beyond that of other technologies like smart speakers or virtual agents. And because the type of overtrust developed in these robots is likely to include ethical trust rather than mere capacity trust, language-capable robots may be uniquely capable of using this trust (intentionally or unintentionally) to exert influence over human morals, for better or for worse.[4] For example, people may read into a language-capable robot's weak response (or lack thereof) to observed sexism as indication that the observed violation is not serious or even as tacit approval of the violation. On the other hand, though, robots that intentionally and carefully respond to observed sexism may be able to make it clear that it should be taken seriously and help exert positive influence on their teammates' moral ecosystem.

Identity. Language-capable robots' unique status as anthropomorphized perceivers and communicators begets unique responsibilities. Language-capable robots are likely to be gendered and racialized in ways that are co-constructed in terms of how they are embodied (that is, their physical morphology) and how they speak (for example, their voice pitch, accent, word choice, and norm adherence). Moreover, robots' embodiments will shape how their speech is perceived, and vice versa. The default identity perception of a language-capable robot is likely to be one grounded in white heteropatriarchy (because it reflects that identity, is designed for the gaze of that identity, or is designed according to the assumptions, mores, and aesthetics of that identity), unless robot social-identity performance is explicitly attended to and monitored. Contrary to some suggestions in the literature, "neutral" gender or race performance may not be a realistic option for language-capable robots due to humans' pervasive application of gendered and racialized norms to speech patterns.

This is a critical design challenge for multiple reasons. First, people's biases carry over into interactions. For example, people interpret and judge robot harshness according to the same gendered norms and stereotypes used to interpret and judge human politeness. Second, robot designs make claims about the roles that roboticists see *humans* of different identities playing within society and make claims about who robots are designed for. For example, female-gendered robots used in service roles make claims about how those roles ought to be gendered. Third,

robot designs reinforce and perpetuate biases.[7] For example, robots that by design attempt to recognize and assign a label of male or female to interactants makes a political claim asserting a binary nature of gender; and when these labels are used or communicated, that claim is perpetuated. All of these factors will influence who sees themselves as future roboticists, reinforcing harmful education and employment trends. More representation is needed in both our robots and our roboticists, and robot designers must pay more attention to identity considerations and obtain more buy-in and engagement from stakeholders during both design and interaction.

Language-capable robots also communicate assumptions about the nature of human identity in unique ways. They can mis-gender or mis-racialize those perceived, either because designers attempt to automatically identify gender or race from sensor data, or because they use language models that use racialized or gendered descriptive cues. Mischaracterizing identity can be traumatic and reinforce stereotyping behavior. These factors also intersect with surveillance and privacy concerns in insidious ways.[1] Robot designers should more carefully attend to identity considerations in robot language design and reconsider where we deploy language-capable robots.

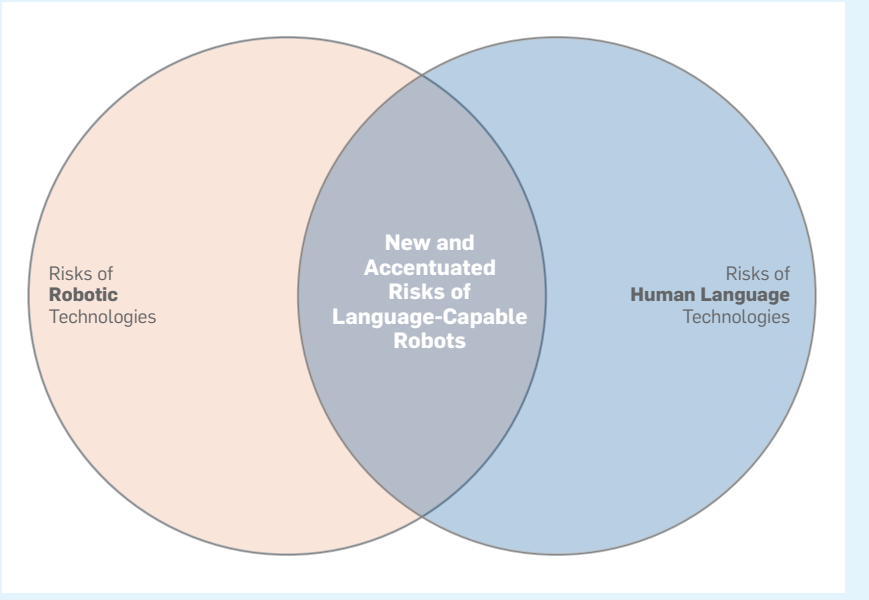
Language-capable robots are privy to both what is visually perceivable and to any nearby conversation, which may be recorded, interpreted, and stored.

Privacy. Any robot perceptual capability raises privacy concerns because it can be used as a mobile surveillance tool. But language-capable robots are privy to both what is visually perceivable and to any nearby conversation, which may be recorded, interpreted, and stored. This combination of physical (and typically mobile) embodiment, and language-based capability to both perceive and to communicate, leads to new and accentuated privacy risks. This is particularly concerning in the sensitive domains where language-capable robots are being proposed, like healthcare contexts or contexts with vulnerable populations like children. And these concerns are exacerbated for language-capable robots in particular due to their ability to pass on what they observe through conversation. Robots that learn from but do not track the provenance of what they are told or overhear pose an outsized risk of privacy breach.

These new risks require new forms of transparency. Users must know what information will be collected, stored, and used, how it will be collected, where it will be stored and for how long, who it will be accessed by and for what purpose, and what could be done with it in the future. They must also know what *they* can do technically and legally to redact their data. Robots must be transparent to users about when they are being listened to (and by whom).

Moreover, roboticists must consider different transparency solutions for different contexts. Users might be provided with privacy knobs that can control who, what, where, when, why, and how data are collected, stored, used, and communicated to others. Robots' language capabilities may need to be leveraged to explain behaviors to patients and to learn privacy policies online. Conversely, designers may need to change user expectations and intuitions, making clear how robots' capabilities for hearing, storing, forgetting, and communicating differ from those of both other voice-interactive technologies (such as Alexa), and other language-capable agents (for example, humans). Data literacy efforts may be needed to increase user understanding of robotic systems and data practices. Researchers may need to help develop community privacy norms and standards, or push for regulatory action mandating privacy-respecting design. Finally, roboticists' decisions regarding all of these considerations—as well as whether to deliberately avoid privacy-sensitive domains altogether—must be made based on factors like the culture in which our robots are deployed, and the populations and communities we are designing for and with.

Language-capable robots create new and accentuated risks different from those presented by (non-communicative) robotic technologies or by (non-embodied) human language technologies.



Recommendations

To respond to these new and accentuated risks, we make several recommendations to roboticists. First, robotics researchers should avoid the use of large language models for robotic Natural Language Generation. The concerns described in this article are likely to be exacerbated for robots whose speech is generated by large language models, which often simultaneously have high fluency and low accuracy and appropriateness. For example a robot whose speech is generated using a large language model may be even more likely to unintentionally exert negative moral influence, to perpetuate gender stereotypes, or to share private information. Second, robots should be designed to consider the side effects of their speech and actions before selecting behaviors because interactants may read a wide variety of socially contextualized intents, implications, and connotations from even simple robot dialogue patterns.[4] For example, before making a statement, a robot might reason about what might be inferred from that statement, and whether those inferences contain private information. Third, designers should increase transparency to help prevent inaccurate inferences about robots' capabilities and intelligence. Robots' true capabilities (both physical and mental) should be as obvious as possible from their morphology and behavior (both verbal and physical). Roboticists should enhance robots' ability to communicate their levels of expertise in particular areas and levels of certainty for particular claims. Finally, to facilitate these explanatory capabilities, the robotics community should increasingly explore capabilities employed in cognitive systems, such as metacognitive reasoning, argumentation, and explanation-generation, which play key roles in human explanation generation.

In addition to specific algorithmic design decisions, we make more fundamental recommendations. First, social capabilities come with moral benefits and risks so roboticists must be careful, thoughtful, and intentional when enabling task-based versus social language capabilities. Social language comes with increased ethical risk, and not every robot needs to converse on social topics or be able to make small talk. Second, because of robots' potential for inadvertent moral and social influence, roboticists should constrain language-capable robots and their conversations to domains and topics where we can guarantee responsible communication. Similarly, robots may need to intentionally profess ignorance on particular topics, or to make clear they were not programmed to be able to converse about particular issues. Finally, robots pose different benefits and risks to different communities based on how those communities weigh and prioritize benefits and risks. For example, the need to prevent privacy violations and the need to be individually recognized (without additional hardware) may be in opposition, and different communities may prioritize these risks and harms differently. The robotics community's design processes and the decisions we make should be attentive to the specific needs and vulnerabilities of the communities we are designing for and with. Frameworks such as Engineering Justice[5] and Design Justice[3] can be leveraged to this end.

These new risks require new forms of transparency.

Conclusion

Through our three examples—trust and influence, identity, and privacy—we have shown how language-capable robots' physical embodiment and speech capabilities interact to create new ethical risks that require new types of responses. These examples, of course, represent just three possible risks, and the responses we suggest are only a few of the possible ways that researchers might respond to them. But while we do not yet have a comprehensive understanding of the landscape of these risks and responses, even the small area that has been mapped provides guidance on the terrain. Those of us seeking to work in this area have a shared responsibility to address both the traditionally acknowledged risks of robotic and human-language technologies and these new and accentuated risks that arise at their intersection—as well as a responsibility to use our understanding of these risks and their possible responses as a starting point for identifying new risks and new possible responses. Finally, it is our collective responsibility to ensure the domains where we are choosing to employ language-capable robots are not merely those with the greatest potential for technical novelty, but rather, those where their benefits are worth the risks. C

References

1. Bender, E.W. et al. On the dangers of stochastic parrots: Can language models be too big? *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. (2021).
2. Browne, S. *Dark Matters: On the Surveillance of Blackness*. Duke University Press. 2015.
3. Costanza-Chock, S. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press. 2020.
4. Jackson, R.B. and Williams, T. Language-capable robots may inadvertently weaken human moral norms. *ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. (2019).
5. Lucena, J.C. and Leydens, J.A. *Engineering Justice: Transforming Engineering Education and Practice*. Wiley. 2018.
6. Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. 2018.
7. Scheuerman, M.K. et al. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society*. (2021).

Tom Williams (twilliams@mines.edu) is an associate professor of computer science at the Colorado School of Mines, Golden, CO, USA.

Cynthia Matuszek (cmat@umbc.edu) is an associate professor of computer science and electrical engineering, University of Maryland Baltimore County, Baltimore, MD, USA.

Kristiina Jokinen (kristiina.jokinen@aist.go.jp) is a senior researcher, Artificial Intelligence Research Center, AIST Tokyo Waterfront, Tokyo, Japan.

Raj Korpan (rkorpan@iona.edu) is an assistant professor of computer science, Iona University, New Rochelle, NY, USA.

James Pustejovsky (jamesp@brandeis.edu) is TJX Feldberg Professor of Computer Science, Brandeis University, Waltham, MA, USA.

Brian Scassellati (scaz@cs.yale.edu) is a professor of computer science and mechanical engineering and materials science, Yale University, New Haven, CT, USA.

Williams was funded in part by Young Investigator award FA9550-20-1-0089 from the U.S. Air Force Office of Scientific Research. Matuszek was funded in part by NSF awards IIS-2024878 and IIS-2145642. Jokinen was funded in part by Project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan. Pustevovsky was funded in part by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. Scasssellati was funded in part by NSF Award 1955653. The opinions expressed are those of the authors and do not represent the views of these funding bodies.