Philosophers Can Help Engineers Describe Their Systems' Capabilities (and Limits)

By Paul Goldberg

The 2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO) in Berlin, Germany featured important research on the ethical, economic, and social implications of robotics: talks discussed tradeoffs between automation, egalitarianism, and growth; public trust in robotics; understanding occasional human cruelty toward robots; and human–robot collaboration on the job, whether through robotic exoskeletons, supernumerary limbs, or in hand-off tasks on job sites.

The concluding talk by Eleanor Watson, titled "Ethics Challenges and Astounding Opportunities: Prompt-Driven Multimodal AI in Science and Education" considered many astounding examples of how the capabilities of artificial intelligence (AI) and robotics technologies are outpacing our expectations and scrambling longstanding discursive, social, and political practices. Our question about technology, Watson incisively summarizes, needs to shift from "Can we do it?" to "Dare we do it?"

Watson's talk displayed remarkable synoptic vision about recent, ethically salient advances in AI research. But I want to single out three of Watson's claims, which she advanced on the basis of recent AI research.

1) AI can decode our thoughts using functional magnetic resonance imaging (based on [1]).
2) Facial recognition technology can reveal our political orientations (based on [2]).
3) Deep learning models can detect a patient's race from chest X-ray scans (based on [3]).

I'm deeply skeptical about each of these claims, and I'll spend the rest of this column explaining why. But let me begin by explaining just what kind of skepticism I intend.

> "WATSON'S TALK DISPLAYED REMARKABLE SYNOPTIC VISION ABOUT RECENT, ETHICALLY SALIENT ADVANCES IN AI RESEARCH."

One might be skeptical of these claims because one doubts that current technology is advanced enough to decode our thoughts, reveal our political orientations from a facial scan, or detect a patient's race. Importantly, however, this kind of skepticism is open to there being a future, more sophisticated set of technologies that could achieve these things.

My skepticism is distinctly philosophical. It isn't based on an empirical assessment of the state of current technologies. Rather, I think the concepts of *thought*, *political orientation*, and *race* operative in the three claims are either mistaken or used in a misleading way. In other words, once we recognize that thought, political orientation, and race are complex and contested concepts, we'll be forced to concede that these three claims almost certainly can't be true when taken at face value, and that insofar as they might be true, they are misleadingly expressed. These claims thus represent object lessons in what engineers can learn from philosophers, whose stock-in-trade is conceptual analysis: clarifying the meaning and implications of our concepts, especially those most foundational (and hence, taken for granted) in our everyday dealings.

On to the claims. The issue with the first one is that "thought" is among the most contested notions in the history of, well, *thought*. Hence, this claim must be relying on one of those (contested) notions, and thus, the claim is just as contestable as that notion upon which it relies. Before continuing, it's worth noting that the authors of the academic paper themselves use the relatively more precise and circumspect notions of "perceived speech" or "imagined speech" rather than "thought" as the objects of the decoding. I'll set this aside, since more sensationalist descriptions of the study (e.g., "mind-reading" or "unspoken thoughts") show up in popular media, most especially in attention-grabbing headlines [4], [5].

We might think that a thought is akin to an inner voice, something like a running internal monologue, or perhaps

Digital Object Identifier 10.1109/MRA.2023.3323101
Date of current version: 12 December 2023

1070-9932/23©2023IEEE

a more or less chaotic "stream of consciousness." But how precise or accurate really is this notion? If thought consists in language, then how do we account for the intuition that people who speak different languages can share a thought—e.g., that "the sky is blue" and "Der Himmel ist blau" express the same thought? If we accept this intuition, then "the sky is blue" expresses a thought but is not itself the thought. Moreover, although surely when I'm thinking, words and sentences sometimes flow through my mind, so too do images, sounds, smells, and so on. And thought is ever modulated by emotions: frustration, boredom, anticipation, and so on. Furthermore, how do we square the fact that thoughts are hazy with the fact that phrases and sentences are (relatively) clear? (Consider how writers daily experience the frustration of producing sentences that fail to capture their thoughts.)

Finally, we might ask what thought *does*: Does it represent states of affairs in the world or does it dispose us to act in the world? To consider the range of possible concepts of thought is to confront the fact that a verbal string (of the type produced by the AI decoder) fails to capture most of what is occurring when we're thinking. Much more plausible—and still astounding, though less sensationalist—is the more precise claim that the decoder can reliably transcribe imagined but unspoken speech.

The problem with the second claim is that our political orientations are intrinsically dynamic and multidimensional. Our stated political orientations can and do change over time (e.g., today's conservative can become tomorrow's liberal). Moreover, the meaning of the categories by which we express our political orientations change (e.g., a "liberal" today generally has far different commitments from a "liberal" of the 1960s, let alone a "liberal" of the 18th or 19th centuries). Relatedly, our political categories carry a wide range of meanings, many of which are mutually inconsistent. Indeed, proponents of these categories often fight over which meaning ought to be dominant (e.g., political libertarians often claim to be "liberals" but favor economic policies that are typically classed as "conservative," and Ron Paul, perhaps the most famous contemporary American libertarian politician, ran for the Republican Party presidential nomination; famously, many more Americans claim to be conservative than tend to vote for the Republican Party). It's thus up for debate which political category best describes one's own orientation.

Finally, there are vastly more dimensions by which to describe one's political orientation than "liberal" and "conservative," the categories referenced in the study (e.g., consider that some on the far left are as hostile to "liberalism" as "conservatism"; we might also distinguish secularists from theocrats, nationalists from cosmopolitans, and much, much more). Indeed, the very subject matter of politics is trying to persuade masses of people that their political beliefs and interests better align with this orientation rather than some other, or to persuade them to discard some of their current commitments for different ones. Therefore, while facial recognition technology might be able to predict, with surprising success, whether an individual at a given time self-describes as "liberal" or "conservative," this information is far less informative than it initially appears to be. It doesn't tell us, e.g., what they mean by "liberal" and "conservative," whether other labels are more appropriate, whether or how their self-description maps on to their policy preferences, or what their political commitments will be in the future. (It's worth noting that Kosinski discusses some of these concerns; but to my mind he does not satisfactorily address them) [2].

Finally, the problem with the third claim is that the traditional, still all-too-prevalent notion of *race*, or of what Kwame Anthony Appiah and Ron Mallon call *racialism*, has been shown to fail decisively [6], [7]. On this notion, *races*—defined as discrete classes of people who share physical, intellectual, and moral characteristics, which are jointly determined by heritable biological traits—exist. But this notion has proven to be a farce. Nature contains animals, minerals, and vegetables; planets and galaxies; electrons and quarks, but it does not contain races. The supposed physical markers of race (e.g., skin tone, hair color and texture) vary continuously rather than discretely across the human population; and as is now well known, "genetic variation within racially identified populations is as great or greater than diversity between populations" [7]. Race, in other words, is not a genuine biological category; it's a pseudobiological category, i.e., a social construction masquerading as a biological category. Nevertheless, people indeed exist within a socially defined racial taxonomy that carries sharp social, economic, and political consequences.

> "IMPORTANTLY, HOWEVER, THIS KIND OF SKEPTICISM IS OPEN TO THERE BEING A FUTURE, MORE SOPHISTICATED SET OF TECHNOLOGIES THAT COULD ACHIEVE THESE THINGS."

The foregoing makes the third claim highly fraught. Indeed, in the study itself, the authors acknowledge that race is a social construct rather than a legitimate biological category, and they are careful to point out that their results show that AI deep learning models can accurately predict a patient's self-reported race [3]. But the trouble is that given the continued prevalence of racialism, unless researchers assiduously emphasize that crucial qualification, then their precise, relatively narrow claim will be easily conflated with the deeply misleading claim that such models can accurately "predict a person's race" [8]. Given that there simply are no races at all, the latter claim is false when taken at face value, i.e., as referring to the ordinary, biological concept of race. And given racialism's continued prevalence, it's crucial that researchers take pains to avoid unwittingly suggesting their support for it, whether in their publications,

in presentations, or in communications with media outlets.

Technologies are indeed outpacing our discursive norms and public policies, and they can be deployed to consequential ends by businesses and states. Moreover, the incentive in media and industry will generally be to amplify sensationalist claims and to dampen important nuances. For all of these reasons, engineers must think and communicate with care and precision about the capabilities and limitations of their systems. The presenters at ARSO 2023 impressively exhibited that care on a wide range of topics. But when matters get especially sticky, consider reaching out to a philosopher. We're trained to help.

REFERENCES

[1] J. Tang et al., "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nature Neurosci.*, vol. 26, no. 5, pp. 858–866, May 2023, doi: 10.1038/s41593-023-01304-9.

[2] M. Kosinski, "Facial recognition technology can expose political orientation from naturalistic facial images," *Scientific Rep.*, vol. 11, no. 1, 2021, Art. no. 100, doi: 10.1038/s41598-020-79310-1.

[3] J. W. Gichoya et al., "AI recognition of patient race in medical imaging: A modelling study," *Lancet Digit. Health*, vol. 4, no. 6, pp. e406–e414, Jun. 2022, doi: 10.1016/S2589-7500(22)00063-2.

[4] H. Devlin, "AI makes non-invasive mind-reading possible by turning thoughts into text," *The Guardian*, May 2023. [Online]. Available: https://www.theguardian.com/technology/2023/may/01/ai-makes-non-invasive-mind-reading-possible-by-turning-thoughts-into-text

[5] P. Dahr, "Can AI and fMRI 'hear' unspoken thoughts?" *IEEE Spectr.*, May 2023. [Online]. Available: https://spectrum.ieee.org/mind-reading-ai

[6] K. A. Appiah, "Race, culture, identity: Misunderstood connections," in *Color Conscious: The Political Morality of Race*, K. A. Appiah and A. Guttman, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1996, pp. 30–105.

[7] R. Mallon, "'Race': Normative, not metaphysical or semantic," *Ethics*, vol. 116, no. 3, pp. 525–551, Apr. 2006, doi: 10.1086/500495.

[8] H. Bray, "MIT, Harvard scientists find AI can recognize race from X-rays — And nobody knows how," *The Boston Globe*, May 2022. [Online]. Available: https://www.bostonglobe.com/2022/05/13/business/mit-harvard-scientists-find-ai-can-recognize-race-x-rays-nobody-knows-how/

FROM THE GUEST EDITORS *(continued from page 9)*

livestock farming systems. Deformable body posture, irregular movement, and the complex farming environment render individual animal tracking in a herd challenging. The authors improved a face-based cow tracking system using You Only Look Once v5 with coordinate attention. A vision transformer was embedded in the reidentification network DeepSORT to enhance feature matching and tracking accuracy. The system was tested on a dataset with multiple cows collected on a commercial farm.

Finally, the article of Car et al. [A7] proposes a fully autonomous robotic indoor farming system. The system, called SpECULARIA, consists of multiple mobile robots and plants grown in moving containers. The work cell is structured such that the system can plan and execute procedures to control every plant's growth and hygiene from seed to harvest. The study benchmarks the proposed setup against a classical mobile manipulation approach to demonstrate its feasibility.

We would like to thank the editor-in-chief of *IEEE Robotics and Automation Magazine*, Yi Guo; the associate editors; and the many anonymous reviewers for their support while creating this special issue. We hope the special issue will motivate researchers and practitioners to develop robotics and AI technologies for agriculture. There are plenty of opportunities, but there are still quite some technical and nontechnical challenges to create a proof of concept and bring it to a marketable and accepted product. Any additional pair of hands is most welcome during this endeavor. You are all most welcome to join this field!

APPENDIX: RELATED ARTICLES

[A1] A. You et al., "Semiautonomous precision pruning of upright fruiting offshoot orchard systems: An integrated approach," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 10–19, Dec. 2023, doi: 10.1109/MRA.2023.3309098.

[A2] F. Esser, R. A. Rosu, A. Cornelißen, L. Klingbeil, H. Kuhlmann, and S. Behnke, "Field robot for high-throughput and high-resolution 3D plant phenotyping: Towards efficient and sustainable crop production," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 20–29, Dec. 2023, doi: 10.1109/MRA.2023.3321402.

[A3] S. Vijh et al., "USMA-BOF: A novel bag of features algorithm for classification of infected plant leaf images in precision agriculture," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 30–40, Dec. 2023, doi: 10.1109/MRA.2023.3315929.

[A4] C. Geckeler, S. E. Ramos, M. C. Schuman, and S. Mintchev, "Robotic volatile sampling for early detection of plant stress: Precision agriculture beyond visual remote sensing," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 41–51, Dec. 2023, doi: 10.1109/MRA.2023.3315932.

[A5] A. Dechemi et al., "Robotic assessment of a crop's need for watering: Automating a time-consuming task to support sustainable agriculture," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 52–67, Dec. 2023, doi: 10.1109/MRA.2023.3321391.

[A6] A. Guo et al., "Vision-based cow tracking and feeding monitoring for autonomous livestock farming: The YOLOv5s-CA+DeepSORT-Vision transformer," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 68–76, Dec. 2023, doi: 10.1109/MRA.2023.3310857.

[A7] M. Car, B. A. Ferreira, J. Vuletiç, and M. Orsag, "Structured ecological cultivation with autonomous robots in agriculture: Toward a fully autonomous robotic indoor farming system," *IEEE Robot. Autom. Mag.*, vol. 30, no. 4, pp. 77–87, Dec. 2023, doi: 10.1109/MRA.2023.3315934.