Power, Equity, and Building Better Robots

As HRI researchers, designers, and developers we need to reflect on the ways that power pervades the social contexts we're designing for and in. What can we do, with the power we have as designers, to produce more equitable HRI?

By Katie Winkle

DOI: 10.1145/3618303

OPEN ACCESS

What does it mean to talk about power and equity in the context of HRI? First, we need to define what we mean by power. I find D'Ignazio and Klein's framing a really useful one, specifically because it links the concept of power to system design: "We use the term power to describe the current configuration of structural privilege and structural oppression, in which some groups experience unearned advantages—because various systems have been designed by people like them and work for people like them—and other groups experience systematic disadvantages – because those same systems were not designed by them or with people like them in mind" [1].

I would argue most of us working in HRI are system designers of one sort or other. Some of us are working to design and implement systems which can better understand human behavior, others are trying to figure out a particular appearance or behavior designs that make robots most appealing and acceptable (or not) to potential users, others still are working to design robot-based interventions that fit within broader healthcare or educational systems. Fundamentally, we're all looking to contribute to the design, development and application of robot systems which work well for people. But which people? All people, equally? Is that even possible?

Similarly, we all exist within a number of systems for which D'Ignazio and Klein's framing of power is equally applicable. Perhaps most obvious to our HRI work might be the educational system(s) we have progressed through and/or are embedded in, but we might also think about the political and economic system(s) we reside within, whether that's at the local, national and/or international level. Thinking about your own university, do all students get the same educational experience? Do systems of financial aid or educational assessment work equally well for everyone? How does the student/researcher experience at your university vary from one in a neighboring city, state, or country?

As HRI system designers, we all have power. Varying amounts of power, generally influenced also (unfortunately) by factors like race and gender, but power nonetheless, and as the old adage goes: "with great power comes great responsibility." We get to shape what HRI research gets done, and hence influence the ways in which robots are likely to

Illustration by Moor Studio / Shutterstock.com

impact society, but also the ways in which such research is done. In our recent work on Feminist HRI [2], myself and my collaborators try to provide a practical guide for examining and challenging power in and with HRI. We believe doing so offers a route to better, more ethical HRI "every day," as well as generating new and interesting research directions for the field.

EXAMINING AND CHALLENGING POWER IN HUMAN-ROBOT INTERACTION

Power differentials pervade the real-world environments and social contexts that we are designing for. Failing to acknowledge and reflect on these when designing HRI increases the risk that our robot deployments will further increase power disparities and inequities among different types of users. Evidence for this can already be seen in current deployments of warehouse robots. Promised to reduce "dirty, dull and dangerous" work, the deployment of warehouse robots instead correlates with increased worker isolation, pressure on production quotes and a surge in workplace injuries [3].

Let's take another example: Imagine you're working on social robots to be used for STEM education in the classroom. We know that the STEM classroom continues to be a site of racialized, gender-based inequity [4]. Is your robot deployment going to reduce, maintain or even amplify such disparities? Given demonstrated issues of race and gender-based bias in computer vision [5] and voice detection systems [6], for example, it's not hard to imagine social robots "failing to notice" those children already most marginalized in the classroom, What message does that send, and reinforce, about whose voices are (not) valued in STEM?

What could HRI designers and developers do to avoid such outcomes? Reflecting on the power relationships between different actors within your use case environment is a vital first step. Knowing about race and gender-based inequities in the STEM classroom you might identify different or additional benchmarking and validation tests you want to run before any real-world deployment—something akin to an ethical risk assessment of your system. You might even re-visit key parts of your system design: What type of sensor technologies you're using, what the robot "looks like" if it is particularly anthropomorphic (thinking about potential for racialization and gendering), how it is intended to influence classroom dynamics, etc. You might even wonder whether your robot should actually attempt to interact an equal amount with each child in the classroom, or instead specifically target increased engagement with those children who are typically most marginalized, as an exercise in more equitable[a] design and advocacy—depending on where you want, and are able, to position your work on a spectrum from "examining and reflecting on power" to "actively challenging power with design."

A feminist approach to HRI design would invite us to recognize and remember that our end users have minds and bodies which experience pleasure, pain, desire, and emotion.

Here we must be cautious, however. Firstly, would targeted engagement actually realize the underlying goal of reduced marginalization in the classroom? Maybe it would actually result in increased (unwelcome) attention, alienation and awkwardness for the targeted students, or maybe teachers will start interacting less with those students (consciously or not) as they know "the robot's going to take care of it." Secondly, while "HRI design as advocacy" has a lot to offer both in terms of improved HRI design and new research directions, we must be cautious of imposing our own values on users. Shaowen Bardzell summarizes this ethical dilemma succinctly in her work on Feminist HCI. Specifically, she draws attention to the fact that design based on empirical research risks working always within the status quo, and hence propagating regressive or harmful practices under the guise of usability. However, designers taking it upon themselves to challenge these practices and produce progressive design solutions, without engagement from key end users or affected communities, can also be problematic. We must question our own position in asserting "what an "improved society" is and how to achieve it" [7]. Particularly, designers in a position of relative privilege, often with little to no lived experience relating to the application context.[b] Setting design or research agendas from this position through a "desire to help," without proper engagement of targeted communities, is reminiscent both of techno-determinist and paternalistic, colonialist savior narratives that ignore work already happening within those communities and minimize their agency: "Don't worry! We're finally here to solve your problems with our superior technology!"

Participatory approaches, which look to center and engage target users throughout design and development processes, provide one obvious way to avoid this.

[a] Compared to equality (treating everyone the same) equity recognizes that certain people, under certain circumstances, need to be treated differently in order to achieve meaningful equality of opportunity.

[b] An example from my own work: I am currently working on a project exploring the use of social robots to increase reach of and engagement with healthy living activities in traditionally disadvantaged suburbs of Uppsala. In comparison, I live in a gentrified area of Stockholm, enjoy relatively good socio-economic stability and am unlikely to face the same barriers to traditional healthcare provision that individuals from these communities are often subject to encounter. Who am I to decide, by myself, the best way to use and design robot technologies for this use case? What I can do however, is provide HRI design expertise and lead technical HRI development within a collaborative co-design project working together with members of the community and healthcare researchers. Previous work has identified grassroots work tackling improved health already happening within the communities—community walking groups, women's groups etc. It is vital to consider how any possible robot-based intervention can further support and compliment these already-working activities.

PARTICIPATORY HRI: SHARING POWER AND BUILDING BETTER ROBOTS

Participatory design and co-design approaches represent ways for designers to engage users in research, design, and development processes. The terms are sometimes used interchangeably, but co-design specifically requires participants have equal authority to the designers, driving the overall what, why and how rather than working to deliver on designers' pre-defined agenda—not always possible given constraints of project-based academic funding and research requirements to deliver particular research outputs. Participatory design still aims to empower potential end-users to contribute to and lead to particular design decisions, but the design space possibilities might be somewhat reduced; consider the questions: "(How) can we use robots to improve community healthcare" versus "how should this robot health coach be personalized to this particular community?" Note a valid answer to that first question on "how can/should we use robots in this application" might be "we cannot." Are we ready to write up and report that as our result? Part of reckoning with power in HRI means recognizing and being transparent about the extent of our own power. The line between co-design and participatory design is fuzzy but if, in reality, we already know we are going to engage in design of a robot health coach because that's what we're funded to do, what your supervisor wants you to do or what you need to do in order to get published and graduate, perhaps participatory design (PD) is a more appropriate and achievable approach—and that's OK! It is still possible to challenge power and engage in more equitable design practices via PD, we should just be open and transparent as to our research agenda and the limits of our design exploration space.

Typical PD methods include focus groups, workshops, and maybe even hackathons, but generally it's still up to the roboticists, engineers, programmers, etc. to go away and implement the co-designed requirements, ideas, or designs—particularly when it comes to the automation of robot behaviors via some sort of machine learning or AI. Participatory automation, done well [8] can allow us to go one step further. In my own work, I'm interested in setups that allow domain expert(s) to "teach" robots how to behave in-situ. I believe such setups represent a way of simultaneously tackling the technical and societal challenges associated with building socially intelligent and/or expressive robots. The socio-emotional intelligence involved in responding appropriately, e.g., to a de-motivated physiotherapy patient, is tacit and intangible. It is not possible, even with the help of domain experts, to write simple heuristics for governing how social robots ought to deal with such situations—I know, because I have tried. Yet, when working with people such domain experts are able to intuitively "do the right thing" for the person they're working with, because they have a good "sense" of what's likely to work best. This intuition is likely informed by their prior experiences with a variety of patients, their knowledge, understanding of, and relationship with this specific patient as well as their real-time assessment of things like the patient's mood and energy levels. But none of this seems to be particularly explicit in their mind at the moment of interaction. They just know what to do. For this reason, human-robot "teaching" setups that allow domain experts to program and/or personalize robot behaviors in real-time seem like an appropriate way to address the technical challenge of generating autonomous, contextually-aware socially intelligent behavior.

Embedded within a broader PD process, such approaches provide a way for domain experts to lead the design, automation, and evaluation of robots [9]. During my Ph.D., I took this approach to development of a "fitness coach" robot; a robot that would guide and motivate users through the kind of boring exercise program that people often give up after a few sessions. I worked very closely with a (human) fitness instructor from my university gym to first identify what the robot should be able to do (its action space), as well as the kind of information it should be using to inform its decision making (its input space). We also worked together to design a "teaching tablet" he could use to control the robot in real-time; providing action exemplars the system could learn from, and, as the system started learning, accepting, or rejecting suggested robot actions as another form of teaching feedback. When it came to putting the robot in to the gym, it was also the fitness instructor who led on deciding where exactly the robot should be placed, and how and when he would "handover" clients to/from the robot fitness coach when they came to exercise. For example, as participants progressed through the program, he elected to guide them through some additional, post-session stretching exercises to supplement their workout with the robot. In summary, I tried to share my power as the researcher/designer, by sharing decision making with (and yielding to) the fitness instructor wherever possible, such that he had a central role in design, deployment, automation, and evaluation of the system.

We get to shape what HRI research gets done, and hence influence the ways in which robots are likely to impact society, but also the ways in which such research is done.

Toward the end of deployment, we tried completely running the robot autonomously based on what it had learned. We demonstrated the robot out-performed a heuristic-based system—an "if this then that" rule-based system using rules also written by our fitness instructor—and delivered personalized motivation to each participant, evidencing the "technical" success of our approach. On the social side, the instructor's presence and involvement in the process clearly impacted participants' engagement with and trust of the robot-based program. Reflecting on the program afterwards, some particularly drew attention to the distinct, but complementary, roles taken by the robot and the instructor, noting they appreciated the combination of the one-on-one exercise time with the robot with the pre- and post-session interactions with the instructor. In summary, empowering the instructor to take such a leading role in the robot's design and development led to technical and social success that I doubt I'd have come anywhere close to if acting alone. Sharing my power to design and develop HRI yielded a better result across all of our typical success measures.

But what if the machine learning approach hadn't outperformed the heuristic system? Or, what if, after months of labor-intensive robot-teaching from the instructor required to train our robot coach, DeepMind or OpenAI released some new pre-trained social chatbot that we could plug into our robot in order to achieve basically the same thing, straight "out of the box"? Such a robot coach might score equally well on our traditional measures of success, but intuitively it seems obvious people might feel differently about a social agent downloaded straight from the web versus one trained for them, with them, and by their personal fitness trainer. Similarly, hands-on involvement in teaching, monitoring, and improving the robot's behavior seems likely to provide more job satisfaction for the fitness instructor—utilizing and valuing their professional skills as made visible via the robot-teaching process. To me, these seem like valid reasons enough to pursue these more participatory approaches. The difficulty is, in a world seemingly driven by increased "efficiency," by a need to do more things more quickly more cheaply, how often are we thinking about, let alone measuring and prioritizing things like job satisfaction, or that "fuzzy" feeling of preferring a locally, taught-by-someone-you-know robot versus over the web-downloaded system, even as they seemingly do the same job equally well?

TOWARD MORE EQUITABLE HRI: REDEFINING WHAT "GOOD" LOOKS LIKE

As is perhaps obvious by now, I really believe (informed by all of my research experience thus far) that power-sensitive and participatory approaches will only improve HRI. That said, I want to avoid any possible suggestion that these approaches are only worth doing because they might yield better results on typical evaluation metrics we see in HRI research papers. Even if they didn't, other proponents of more ethical HRI and I would argue for them anyway, as we need to think more about the pros and cons, the costs and benefits of the processes we are engaging in, rather than just the "end results." In fact, designing more equitable HRI might even require us to re-visit our evaluation metrics or measures of success to re-define what "good" looks like. Take the warehouse robot example discussed earlier in this article. Typical HRI measures of task performance or speed clearly correlate with what operations managers or corporate entities might want from deployment of such robots. What sort of measures would represent an improvement in the workers' experience? A feminist approach to HRI design would invite us to recognize and remember that our end users have minds and bodies which experience pleasure, pain, desire, and emotion. We might think about evaluating (and designing for) additional success measures relating to user comfort and experience. We might also think about designing robots and interactions that help our end-users achieve a sense of competence and meaning in their robot-assisted work, interactions that positively foster a sense of relatedness whether that's to the robot itself but also ideally to other humans as well.

Part of reckoning with power in HRI means recognizing and being transparent about the extent of our own power.

As a research field, we in HRI know how to do some of these things really well. We've seen work on hugging robots that tackle isolation, social robots that bring different human users together, and animation-inspired design that is a delight to engage with. We also, of course, have expertise in collaborative and industrial setting HRI, so the question going forward is how can we bring these together? How can we work together across specialisms to tackle some of the new and exciting research gaps that emerge as we try take more power sensitive, participatory and/or feminist approaches to HRI? Having had so many interesting and exciting conversations with students and other young researchers at this year's HRI conference, I am confident that we're going to see an increasing amount of research and design tackling these questions, and I, for one, cannot wait to see the results.

References

[1] D'ignazio, C. and Klein, L.F. *Data Feminism.* MIT Press, 2020.

[2] Winkle, K., McMillan, D., Arnelid, M., Harrison, K., Balaam, M., Johnson, E. and Leite, I. Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, New York, 2023, 72–82.

[3] Engstrom, E. and Jebari, K. AI4People or People4AI? On human adaptation to AI at work. *AI & SOCIETY* 39, [2022].

[4] Kuchynka, S.L., Eaton, A. and Rivera, L.M. Understanding and addressing gender-based inequities in STEM: Research synthesis and recommendations for US K–12 Education. *Social Issues and Policy Review* 16, 1 [2022], 252–288.

[5] Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in PMLR 81:77-91. 2018.

[6] Parreira, M.T., Gillet, S., Winkle, K. and Leite, I. How did we miss this? A case study on unintended biases in robot social behavior. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, New York, 2023, 11-20.

[7] Bardzell, S. Feminist HCI: Taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New York, 2010, 1301–1310.

[8] Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M.C., Gabriel, I. and Mohamed, S. Power to the people? Opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization [EAAMO '22].* ACM, New York, 2022.

[9] Winkle, K., Senft, E. and Lemaignan, S. LEADOR: A method for end-to-end participatory design of autonomous social robots. *Frontiers in Robotics and AI*, 8 [2021].

Biography

Katie Winkle is an assistant professor in social robotics at Uppsala University in Sweden. Originally studying engineering [and still thinking of herself as an engineer first and foremost] her work draws from design, computer science, psychology, and the social sciences to tackle both technical and societal challenges in delivering trustworthy human-machine interaction.