Robots for Social Justice (R4SJ): Toward a More Equitable Practice of Human-Robot Interaction

Yifei Zhu
Colorado School of Mines
Golden, CO, USA
zhu1@mines.edu

Ruchen Wen
Univ. of Maryland, Baltimore County
Baltimore, MD, USA
rwen@umbc.edu

Tom Williams
Colorado School of Mines
Golden, CO, USA
twilliams@mines.edu

ABSTRACT

In this work, we present *Robots for Social Justice (R4SJ)*: a framework for an equitable engineering practice of Human-Robot Interaction, grounded in the *Engineering for Social Justice* (E4SJ) framework for Engineering Education and intended to complement existing frameworks for guiding equitable HRI research. To understand the new insights this framework could provide to the field of HRI, we analyze the past decade of papers published at the ACM/IEEE International Conference on Human-Robot Interaction, and examine how well current HRI research aligns with the principles espoused in the E4SJ framework. Based on the gaps identified through this analysis, we make five concrete recommendations, and highlight key questions that can guide the introspection for engineers, designers, and researchers. We believe these considerations are a necessary step not only to ensure that our engineering education efforts encourage students to engage in equitable and societally beneficial engineering practices (the purpose of E4SJ), but also to ensure that the technical advances we present at conferences like HRI promise true advances to society, and not just to fellow researchers and engineers.

CCS CONCEPTS

• **Social and professional topics** → *Codes of ethics*; • **Computer systems organization** → **Robotics**.

KEYWORDS

Social Justice, Engineering, Robot Design, Human-Robot Interaction

ACM Reference Format:
Yifei Zhu, Ruchen Wen, and Tom Williams. 2024. Robots for Social Justice (R4SJ): Toward a More Equitable Practice of Human-Robot Interaction . In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634944

1 INTRODUCTION

1.1 Equity in Technology

Like all artifacts, technologies are laden with values and virtues, preferences and politics [68]. In particular, technologies are especially laden with the values, virtues, preferences, and politics of their *designers* [18]. This gives designers incredible power to reinforce or rewrite the patterns, systems, and structures that define and govern society, especially for technologies deployed and used broadly[50]. While the systemic impacts of *non-embodied* computing technologies (like loan disbursement, recidivism prediction, and benefit approval algorithms) have received the most attention in the literature [23], robotic technologies may similarly impact our society if deployed as broadly.

Critically, the risks of widespread *negative* systemic impacts are critical to consider even for technologies that aim to yield objectively positive benefits to their users. Some scholars have recently discussed, for example, how educational technologies like smartboards (which improve student learning outcomes and increase classroom efficiency) have led to greater social inequality. Since smartboards were disproportionately deployed into high-resourced schools, they disproportionately benefited students with high socioeconomic status, thus further increasing economic disparities along class lines [10, 60]. Many robotic application areas dear to the field of Human-Robot Interaction (HRI), including education, healthcare, and therapy, run the risk of similarly exacerbating class and race-based inequality depending on where and how they are deployed. Although an educational robot might "do good" at an individual level, if its price tag renders it accessible only to highly resourced school districts, it may very well "do harm" at a societal level, widening the existing gap between resources available for different school districts.

These concerns are further magnified for the language-capable robots centered in much HRI research, whose speech itself reflects a set of implicit values, virtues, preferences, and politics [65]. Because robot speech may lead users to attribute robots with intelligence [6, 14], social agency [28], moral agency [16, 52, 57], and the uniquely potent combination of these factors [27], robots may be accepted as informed community members (although cp. [7]), granting them an outsized potential for persuasion and influence [26, 50, 53]. Depending on who designs these robots – and the community into which they are deployed – this provides opportunities either to reinforce the customs and norms of those communities or for robots' designers to exert a colonizing influence from afar by imposing colonial ideologies through robotic technologies [25].

Because of this potential for societal and interpersonal influence, it is critical for robot designers and researchers to consider not only the direct impacts that our technologies have on individual users but also how our robots impact entire communities – and how these community-level impacts can be anticipated and steered through equitable robot design processes.

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634944

1.2 Equitable Human-Robot Interaction

To grapple with these societal and ethical dimensions of robot design, a number of researchers have recently presented guidelines for designs grounded in different equitable design and social frameworks from beyond robotics. For example, Ostrowski et al. [43] presented the HRI Equitable Design framework, which adapts Costanza-Chock's Design Justice framework [8] to the field of HRI, expanding on Costanza-Chock's original design questions through six additional questions for robot designers. To motivate this extension, Ostrowski et al. analyze the ways that equity and social justice have (and have not) been incorporated in past papers published at HRI venues, and show that societal impact is rarely considered or discussed. Their resulting framework thus serves to push designers to center equity by asking questions regarding *Who is included* and *Who gets to design and benefit* [43].

As a second example, Winkle et al. [67] presented the *Feminist HRI* framework, which adapts D'Ignazio and Klein's *Data Feminism* [12] to the field of HRI, raising the conversation of who designs and benefits to a higher societal level through power structure analysis. Winkle et al. highlight that designers are not simply makers; they also hold a power within our society that is wielded through and embodied in their robotic designs. *Feminist HRI* provides guidelines for how robot designers can *examine and challenge* this power structure both in their robot designs and in their research practices.

Finally, Tanksley presented an *Abolitionist pedagogy*, which leverages theories from Critical Race Theory to actively engage with anti-black racism in STEM education, especially with respect to robotics and AI [59]. Specifically, Tanksley argues that by incorporating and prioritizing abolitionist ideologies in classrooms, STEM students, who are future engineers and designers, can approach technology more thoughtfully, and more importantly, they will be equipped with knowledge and tools to challenge racism and inequality that are manifested in existing systems [59]. This framework aims to facilitate a justice-oriented infrastructure, and encourages educators to design curricula that can guide students to critically evaluate and probe the technology they might encounter.

The *HRI Equitable Design Framework*, *Feminist HRI*, and *Abolitionist Pedagogy* provide crucial guidelines and tools for researchers and educators as to how to interrogate and challenge who has power, who designs, and who serves. Indeed, power structures shape the needs of communities, and any framework that hopes to meaningfully approach robotics from a social justice perspective needs to examine those power structures as the underlying factor for social justice concerns. Yet, existing frameworks leave open key questions regarding the practice and praxis of equitable robotics. Specifically, we argue that the research and design of robots cannot be separated from their practical *engineering*. That is, the questions of who designs robots, who benefits from robots and what ideology or social hierarchy permeates robot designs, cannot be separated from the question of how the process of engineering robots is enacted in order to address the specific needs of specific communities. In this work, we thus leverage a recent framework from Engineering Education, *Engineering for Social Justice* (E4SJ), to present a fourth framework for equitable robotics, *Robots for Social Justice* (R4SJ), which complements the *HRI Equitable Design Framework*, *Feminist HRI*, and *Abolitionist pedagogy* from an Engineering perspective.

Presented in Leydens and Lucena's groundbreaking book "Engineering Justice" [35] (see also [32, 34]), E4SJ presents specific criteria for engineering practice that centers *Enhancement of Human Capabilities in Communities*. Leydens and Lucena demonstrate how these principles can be integrated into real-world engineering practice, and the concrete opportunities and challenges for instilling these principles into engineering practice, from an Engineering Education perspective.

In this work, we review the past ten years of work published at the HRI conference through the lens of E4SJ. In doing so, we identify key gaps in HRI research practice that become visible through this lens. By taking an E4SJ perspective, our results show how some communities are overlooked, what capabilities are not being actively engaged that could be enhanced, and what elements of E4SJ remain absent from HRI practice. Overall, our results show how an engineering practice grounded in *Robots for Social Justice* (R4SJ) could lead to better engineering of intelligent, interactive robotic technologies of true societal benefit by considering not only the communities served, but also by directing engineering practice to best enhance the capabilities of those communities.

2 ENGINEERING FOR SOCIAL JUSTICE

2.1 Engineering practices and communities

The field of engineering centers and relies on technical and scientific knowledge and skills to solve problems. Yet this techno-scientific lens often results in engineering efforts being isolated from the societal context in which the products of engineering are deployed [15, 61]. For engineering efforts to be successful, it is paramount that engineers thoughtfully engage with communities and society. Armstrong and Baillie [2] argue this by emphasizing the importance of cultural relativism in engineering practices, noting that an equitable relationship between engineers and the community they interact with is essential for the sustainable success of engineering solutions [2, 3]. Moreover, engineers must genuinely understand and respond effectively to changing social context [29], and responsibly engage community members in response to those changes [13, 49]. If engineering efforts become de-contextualized from their socio-technical environment or detached from the communities they impact or intend to serve, they may ultimately and inadvertently exacerbate the social injustices that impact, constrain, and structure those communities [37].

Engineering is fundamental to the field of HRI. The physical robot bodies that researchers use, the sensors and actuators that carry out interactions between robots and people, and the algorithms and software that determine robot behaviors, are all products of engineering efforts. Consequently, research in HRI embodies and extends these engineering practices. While some research objectives, such as producing new knowledge, might seem distant from community members, research efforts have the crucial role of informing, guiding, and shaping future engineering efforts, and are therefore intrinsically intertwined with community-centered engineering practices. Thus, in this work, we explore how the practice of HRI research and engineering could be pursued more ethically and responsibly by borrowing from and building on engineering education frameworks that center the relationship between engineers and community members.

2.2 The E4SJ framework and HRI

Engineering for Social Justice (E4SJ) is a set of engineering practices that strives to enhance human capabilities through equitable distribution of opportunities and resources while reducing imposed risks and harms within specific communities [35]. Specifically, when viewed through the lens of the E4SJ framework, the central goal of engineering is that engineers should work alongside communities to develop engineering solutions that **enhance human capabilities** in a way that aligns with community priorities. The way that this goal should be pursued is guided by five key principles:

(1) **Listening contextually** – at the basis of Engineering practice is listening to and empathizing with different communities' perspectives and their constituent struggles, concerns, desires, and preferences.
(2) **Identifying structural conditions** – these perspectives must be understood through the lens of the structural conditions (e.g., racial, gendered, socioeconomic) that constrain those communities' opportunities, desires, and aspirations, as well as the structural conditions that constrain the engineers' own opportunities, desires, and aspirations.
(3) **Acknowledging political agency / mobilizing power** – engineers must understand how communities' political power and agency (as well as their own) can be mobilized and leveraged when developing engineering solutions.
(4) **Increasing opportunities and resources** – engineers should work with communities to identify the opportunities (e.g. health, education, housing, and employment) that could be improved by leveraging and mobilizing political power, engineering solutions, and other resources, as mediated by structural conditions.
(5) **Reducing imposed risks and harms** – engineers should work with communities to identify how to leverage the identified resources to develop solutions in a way that is sensitive to how the solution's potential risks could be distributed across the community.

This framework, while originally designed to help cultivate more equitable practice in the broad context of engineering education, may be applied to the fields of Robotics as a form of third-wave AI Ethics [4, 30] (i.e., focused not solely on moral philosophy or fairness, accountability, and transparency, but moreso on issues of power and social justice). E4SJ requires a community-focused approach, in which the engineer (or designer, or researcher) specifies and focuses on a particular community that is structurally disadvantaged in particular ways, such as children, AAC device users, farmers, people with disabilities, LGBT+ people, Black people, women, immigrants, incarcerated people, and so forth. A community-engaging approach has been shown to be effective in helping engineering students shift their mindset from charity to thinking for the community, to advocate for social justice, and to develop meaningful empathy for specific communities [48, 66].

Taking this community-focused approach is critical to the development of technology that can truly benefit communities. Technologies need to be designed with specific communities in mind because different communities have inherently different *axiologies* – that is, they have different values, and thus different goals. Navigating this tension requires one to clearly specify the community that's being designed for, to have an awareness of this specific community, and to develop a local understanding of that community (or, at least, explicitly building on the findings of others who have developed and documented this level of understanding). Failure to do so risks technology only meeting the values and needs of the technology's developers and engineers. Similarly, developing this local understanding forces the technology developer to ensure that the technology fits into the lives of the actual people who are expected to use it.

When viewed through this lens, two key elements stand out to us as immediately applicable to research practice in HRI: (1) the research we perform should be motivated by the needs of **specific communities** in order to help those communities achieve an equitable distribution of opportunities and resources that they otherwise lack due to structural conditions (cf. [36]), and (2) the types of solutions proposed in our research should seek to achieve this goal specifically through means that advance the key human capabilities differentially valued by communities. In particular, Leydens and Lucena suggest that engineers should strive to, through the design process, advance one or more of the 10 capabilities delineated by Martha Nussbaum, which we discuss in the following section.

2.3 Human Capabilities

Nussbaum introduced a human development paradigm, which applies the capabilities approach, and generated a list of central human capabilities from the following ten aspects [42]:

(1) **Life** – Ability to live to the end of a life of normal human length without it being cut short.
(2) **Bodily health** – Ability to live in good health, with access to adequate nutrition and shelter.
(3) **Bodily integrity** – Ability to move freely, be free from assault, and have sexual satisfaction and reproductive choice.
(4) **Senses, imagination, and thought** – Ability to use senses to imagine, think, and reason, informed by an adequate education, to produce and experience works, events, political and artistic speech, and religious exercise, of one's own choice; and the ability to have pleasurable experiences and avoid pain
(5) **Emotions** – Ability to experience and explore positive and justified negative emotions and feelings towards others.
(6) **Practical reason** – Ability to develop and engage in reflection as to what is "good", and use the resulting personal axiologies to engage in goal-driven self-reflection.
(7) **Affiliation** – Ability both to (a) live, engage socially, and empathize with others; and (b) be treated with dignity and respect, and avoid discrimination.
(8) **Other Species** – Ability to live with and experience concern for other species (e.g. plants and animals) and the natural world in general.
(9) **Play** – Ability to laugh and play.
(10) **Control over one's political and material environment** – Ability to participate effectively in political processes affecting one's life, hold property, seek employment and work in a human-like, goal-driven, and social way.

There are important questions that can be raised about the ontology (what "is") of capabilities, such as who gets to decide what capabilities belong in such a taxonomy. There are important questions that can be raised about the epistemology (how the "what is" "is known") of capabilities, such as how we adjudicate "levels" of capabilities when those levels have been reached, and who gets to decide on those levels. And, there are important questions that can be raised about the axiology (what "is valued") of capabilities, such as how different capabilities might be differentially valued across different communities and in different cultures.

Nevertheless, these capabilities serve as a starting point for a capability-directed discussion of our own field's advancement of human capabilities. Moreover, by centering *axiology*, this framework provides a critical first interrogation of the most basic questions surrounding what our field chooses to value, and how this is reflected in our research. Finally, while the E4SJ method is just one theoretical framework within which engineers can pursue the advancement of these capabilities, we argue that it provides a productive and nuanced way to discuss those capabilities. Outside the context of the E4SJ framework, for example, one might be able to motivate the development of certain technologies through their ability to enhance affinity, even if the target users whose affinity would be enhanced would be a group doing demonstrable societal harm, such as organized white supremacists. In contrast, operating within the context of the E4SJ framework encourages engineers to specify *whose* affinity is being enhanced. And conversely, operating outside the context of E4SJ, one might be able to motivate which community the technology is being designed for and with; while operating within the context of the E4SJ framework encourages engineers to specify precisely how the technology is benefiting that community in terms of that community's priorities over different capabilities, and the ways that that community's capabilities are uniquely constrained by structural conditions.

As such, in this paper, we use these capabilities, and the specific *ways* in which the E4SJ framework suggests engineers seek to advance those capabilities, as a lens for analyzing the state of Human-Robot Interaction research. As we will show, while *all* of the capabilities laid out by Nussbaum stand to align with HRI solutions, in practice most HRI research (at least based on what is reflected at the ACM/IEEE International Conference on Human-Robot Interaction) is not explicitly motivated by these sorts of capabilities, and the research that *is* capability-motivated seeks to advance a narrow set of capabilities, such as preventing harm, promoting social engagement, caring for human needs, providing education, and promoting good health. This suggests that the space of capability-focused solutions explored by HRI researchers may be overly focused on a few goals at the expense of others due to (1) the particular axiologies and lived experiences that are common to HRI researchers and (2) the current HRI research funding ecosystem.

Moreover, as we will show, the *way* in which capabilities are typically advanced is misaligned with the community-focused approach that is proposed by the E4SJ framework. Even when HRI researchers produce technical advancements oriented around facets of, say, interaction, their solutions do not typically directly focus on the interaction needs of particular communities that are otherwise inequitably stymied by structural forces. Finally, while we will not deeply discuss it in this work, researchers tend to not include members of those communities in their research teams or even (explicitly) build off of work that does include and engage with members of those communities.

Accordingly, we believe the E4SJ approach stands to address these shortcomings in the HRI field's research practice. Specifically, we believe that the key human capabilities delineated by Nussbaum, when paired with a community-centered view of engineering as suggested by the E4SJ framework, can serve as guiding principles for the field, helping us to better gauge the promise of solutions being suggested by ourselves and others in our community from a social justice perspective.

3 METHOD

To understand the extent to which the HRI community is engaging in research practices aligned with the principles of E4SJ, we analyzed the papers published at HRI from 2012 to 2022[1]. While this obviously does not encapsulate the entirety of the Human-Robot Interaction and Social Robotics communities, this set of papers was selected due to the balance it strikes between comprehensivity (all papers from this conference from this decade are covered) and concision (the small, single-track nature of HRI leads to a manageable number of papers). In the rest of this paper, we discuss the results of this analysis and use it to motivate a vision for a more equitable future of HRI research practice.

The analysis follows these steps: For each paper, the first author assessed whether it (1) identified a specific community, aside from "engineers", "HRI researchers", or "people who happen to be interacting with a robot in some un(der)specified and mysterious circumstances", (2) expressed a motivation aligned with one of the ten key human capabilities, and (3) expressed a motivation to perform some task *in a way* that aligned with *enhancing* one of the ten key human capabilities, even if the main purpose of the work did not. From the set of papers that met all three criteria, we then further apply a fourth criterion to only include those that (4) either included an analysis of the explicit needs of their selected community, or cited such an analysis from some other work. All four criteria are inspired by and closely adhere to the E4SJ principles.

To apply the first criterion, the first author focused on the introduction and related work sections to look for the motivations articulated in each paper. For example, one work explored how drones could convey emotions through flight paths [5], however, because no specific user population was mentioned other than people who might be interacting with drones, this paper was excluded from further analysis. For papers not excluded in this way, the second criterion was then applied. For instance, one paper specifically focused on people who need assistive feeding and articulated a motivation to improve the feeding experience for them [19]. This paper was therefore retained after considering the second criterion. For papers meeting both the first and second criteria, the third criterion was then applied by examining each paper's results. For example, one work explicitly mentioned children who use robots as playmates and conducted experiments with children to investigate differences in engagement under different prosodic synchronicity, finding the proposed system improved engagement for children

[1]Papers from HRI 2023 were not included because these efforts began prior to the public release of papers from HRI 2023 into the ACM Digital Library

during play [51]. This work was therefore retained after considering the third criterion.

Finally, the first author examined the papers remaining after applying the first three criteria, and looked for discussions and considerations of concrete user needs. For instance, a paper that was motivated by helping dementia caregivers *(a specific community)* specifically expressed the goal of better understanding how robots can emotionally support caregivers *(enhancing a specific human capability, i.e., emotion; and expressing motivation to enhance that capability, i.e., through design guidelines for robots to be developed to help caregivers)*, explicitly discussed the emotional diligence required from dementia caregivers, and explicitly discussed the need to ease this burden *(discussion of the concrete needs of this community)* [39]. Because this paper met all four criteria, it was included in our final paper subset for further analysis. While these decisions were made subjectively, the first author did their best to be systematic and generous, looking for any indication that could be used to justify inclusion on the basis of the criteria above. Even with the subjective nature of the analysis, the results discussed in the following sections establish a general overview of the state of the HRI field and initiate a discussion on how the field of HRI could more intentionally advocate for social justice and equity.

4 RESULTS AND DISCUSSION

Before discussing our result, we acknowledge that many papers published at HRI are motivated by research questions surrounding fundamental dimensions of interactivity, such as transparency and trustworthiness, whose insights can be applied broadly across different domains. Our intent is not to critique these basic research approaches, but rather to highlight gaps in the HRI literature. However, as we will argue later, we do believe that even in these types of fundamental works, researchers should strive to articulate the communities that *could* be ultimately served, and the capabilities *could* be ultimately enhanced, by their research findings.

4.1 Do HRI researchers clearly aim to meet the needs of specific communities?

521 research papers were presented at HRI from 2012 to 2022, with the Alt.HRI session added in 2016. Of these accepted papers, we identified 99 papers that specified a user population, and with further analysis, we identified 90 papers as clearly considering the user population's specific needs, instead of merely mentioning the intended beneficiary communities. The majority of the papers that *did* specify intended beneficiary communities (also illustrated by different sessions of HRI each year) presented technical advancements designed to help one of the following user groups: (1) Educators (2) Children with ASD (3) People with certain disabilities (blindness, hearing impairment etc.) (4) Medical personnel supporting patients suffering from Dementia, Parkinson's disease or other medical conditions (5) People in need of comfort, encouragement, and companionship.

While researchers should avoid designing for particular communities merely for the sake of novelty, this does suggest that HRI researchers might be overly focused on a relatively narrow subset of communities, and that it may be worth broadening the field's horizons to consider a broader range of possible sites of use. In particular, we note that the communities that are commonly centered are those that align with overarching domains or use cases, such as education and health. In contrast, the communities most systematically disadvantaged by structural conditions (at least in American society) – i.e., communities of color and low-income communities – are notably missing from this list.

This lack of attention to communities that are most structurally disadvantaged highlights the need to attend to power differences within and between communities. Like Winkle's *Feminist HRI* framework [67], E4SJ emphasizes a need to analyze who has power and influence, and who is most likely to be affected by technology, as captured by E4SJ's third principle. Because E4SJ is a framework from the Engineering Education community, however, the way that power is analyzed is slightly different under E4SJ than under other recent frameworks. Leydens and Lucena, for example [33], highlight key education tools like *Rainbow Diagrams* and *Privilege by the Numbers Activities*, that can be used to practically analyze power differentials within communities and between engineers and the communities they hope to serve. Moreover, E4SJ encourages engineers to find ways to leverage both their own power and the power sources that exist within communities in order to make their engineering solutions effective. These considerations were essentially ignored within the papers that we analyzed. This is important because engineers themselves do not exist outside of power structures, and by engaging with community members, a power relationship is established between engineers and community members.

4.2 Do HRI researchers clearly aim to enhance human capabilities within those communities?

Of these 90 papers, 79 engaged with and sought to enhance at least one of Nussbaum's 10 human capabilities (Fig. 1), and 10 addressed multiple capabilities.

- 35 papers presented technologies that either directly or indirectly facilitated *Sense, imagination and thought*, by helping users retain information, learn new knowledge, or improve creativity. For example, researchers considered how adaptive tutoring robots could assist in child second language acquisition [11],
- 16 papers presented technologies that facilitated *affiliation* through helping the user develop or improve certain social skills, and moderate and improve team dynamics. For example, researchers considered how a robot mediator can help resolve interpersonal conflict [58].
- 21 papers presented technologies that facilitated *Bodily health* by offering assistive robots and systems to people who are in need of rehabilitation, therapy, or companionship, or indirectly offering assistance and improvements through supporting medical care professionals. For example, researchers considered how to improve the teleoperation of assistive robotic arms for patients with upper extremity disabilities [24].
- 8 papers sought to facilitate *emotion* by designing robots that express emotion, or robot systems that can perceive user's emotion. For example, researchers considered how families interact with the Cozmo robot to better understand

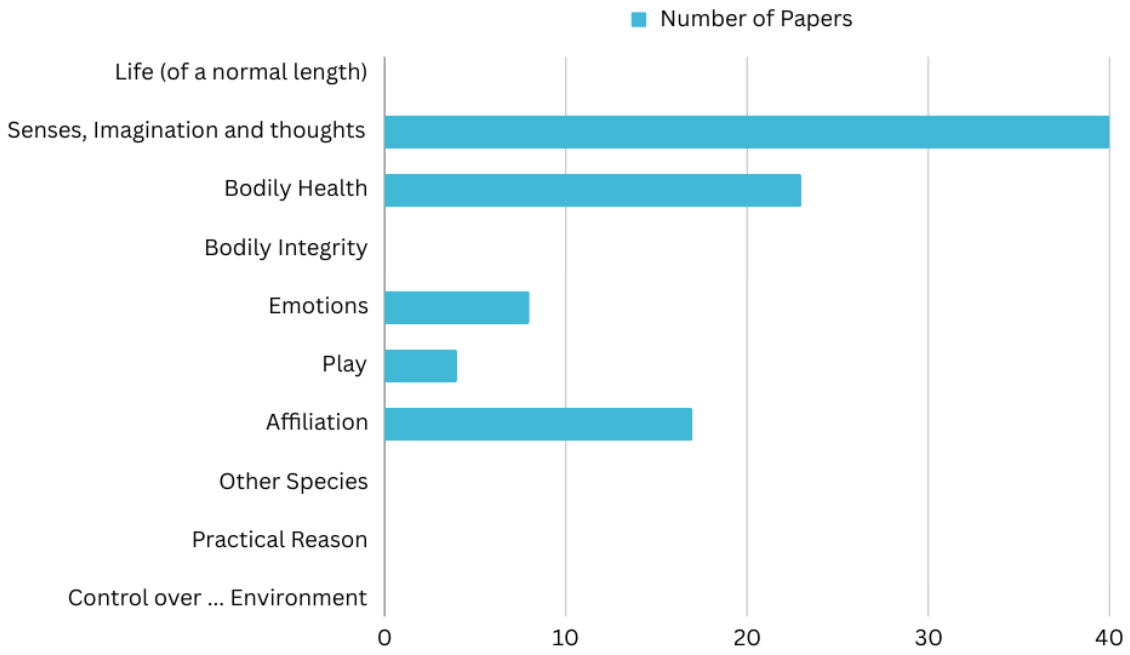

Figure 1: Papers directly or indirectly engaging with each of Nussbaum's 10 human capabilities. Papers engaging with multiple capabilities were counted multiple times.

the nuanced role that sadness plays in HRI to better future designs and offer more thoughtful interactions [46].

- 4 papers presented technologies that facilitated *Play* by offering robotic assistance to entertain or play with a demographic. For example, researchers considered how the skills of stand-up comedians could be used to improve the humorous aspects of human-robot interactions [63].

As mentioned above, while researchers should avoid designing to enhance particular capabilities merely for the sake of novelty, this analysis again suggests that HRI researchers might be overly focused on a narrow range of capabilities, and that it might be worth broadening the field's horizons to consider a broader range of possible benefits, i.e., to expand the range of what is considered valued in HRI. In particular, we note that the capabilities that are commonly centered, namely, *Sense, imagination and thought* and *Bodily health* are again those that align with the overarching domains or use cases of education and health. In contrast, there may be unexplored opportunities when it comes to possible ways for robots to enhance the ways that the need for moral reasoning, play, avoidance of premature death, bodily integrity, and harmony with other species, uniquely manifest in different communities.

4.3 What other motivations comprise the axiology of HRI research?

Now that we have discussed the ways in which the axiology of HRI research papers *have* aligned with Nussbaum's capabilities, let's consider the axiology of the remaining papers. Virtually all of the remaining 417 papers had concrete motivations not captured by the E4SJ framework. These motivations include:

(1) an abstract desire for explainability and understanding of a robotic system (i.e., Leutert et al. [31])
(2) better understanding of trustworthiness (i.e., Sebo et al. [54])
(3) improving the efficiency of an existing system (i.e., Milliez et al. [38])
(4) improving robot perception, cognition, and behavior modeling (i.e., Murakami et al. [41])
(5) development of novel algorithmic capabilities (i.e., Mohseni-Kabir et al. [40])
(6) better understanding of robots' role in our society and how we perceive robots in social contexts (i.e., Paepcke and Takayama [44])

While all of these types of approaches have the potential to help create a more equitable society, the lack of articulation and discussion of an intended beneficiary community (and thus, subsequently, a lack of specification for how that community was intended to be helped by the technology) evokes a dangerous perspective in which these advances are cast as beneficial in and of themselves. While more theoretical and fundamental work often aims to expand knowledge and information, it is still possible to envision future applications and use cases, and subsequently, identify an intended community to serve.

It is especially important for HRI researchers to be clear about the specific communities they intend to benefit due to the previously

discussed ways that well-meaning technologies can increase inequity, especially in the exact types of domains that HRI frequently centers around, like education. Further, while it is true that *explainability* and *trustworthiness* are admirable goals in some contexts (e.g., a robot that shares critical systemic knowledge with undocumented communities ought to be trusted in order to actually help the community), these principles can be rendered dangerous when re-contextualized into domains in which explanation-generation and trust-building mechanisms are deployed in order to coerce compliance with existing state power structures (e.g., robots deployed by corporate or state actors such as police for the purpose of surveillance or oppression should not be blindly trusted [64]).

Similarly, *efficiency* can be an admirable goal in the context of making robots affordable for low-income communities, serving a greater number of hospitalized children, enhancing disabled users' mobility, or reducing energy use in all these domains. But in many of the domains described in the analyzed papers, increased efficiency would primarily stand to benefit the wealthy executives and shareholders who may be exploiting the labor of those interacting with the robot. For instance, a social justice-oriented approach to increasing efficiency in a work environment would need to be motivated by a community-provided efficiency concern grounded in one of Nussbaum's 10 human capabilities. That is, the beneficiary community would be *workers*, instead of the corporations; the approach would enhance some aspects of the *worker's capabilities*, instead of merely maximizing output for corporations. Through this approach, we can rethink who is considered a stakeholder [17], and who should we prioritize and serve.

Moreover, there is good reason to be skeptical of a *complete* emphasis on metrics such as efficiency, effectiveness, and transparency, which are traditionally centered by neoliberal axiologies and theories of value [47, 62].

5 RECOMMENDATIONS: ROBOTS FOR SOCIAL JUSTICE

Based on our analyses in the previous sections, we propose the following high level recommendations, which reflect a *Robots For Social Justice* engineering process (Fig. 2).

- **Recommendation 1:** HRI Researchers should clearly specify the communities their research is intended to benefit. Even for highly theoretical research, we argue that researchers should be able to identify some community that would ultimately benefit.
- **Recommendation 2:** HRI Researchers should clearly specify the human capabilities their research is intended to enhance for those communities. Even for highly theoretical research, we argue that researchers should be able to identify some human capabilities their research would enhance.
- **Recommendation 3:** HRI Researchers should provide clear justification, grounded in close, contextual listening (by themselves or others), for claims regarding the value and prioritization placed on those capabilities by those communities.
- **Recommendation 4:** HRI Researchers should clearly specify the relevant structural conditions that motivate, constrain, and shape those values and priorities.

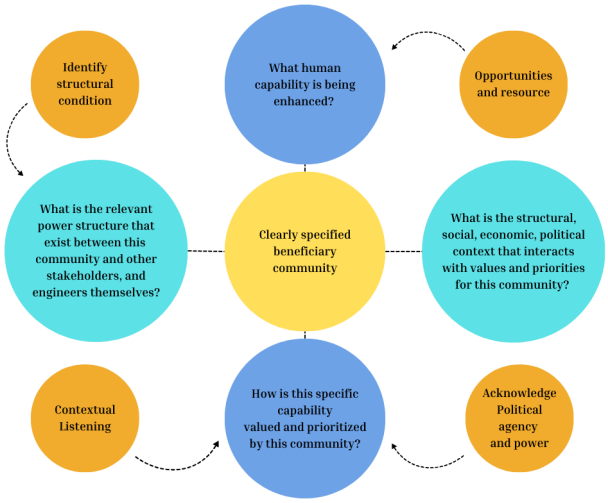

Figure 2: 5 R4SJ recommendations and E4SJ criteria

- **Recommendation 5:** HRI Researchers should map out the power structures within the communities they wish to serve; between the community and the other stakeholders with which they interact; and between the engineers themselves and these communities and other stakeholders.

Building on the insights of the *Design Justice*, *Abolitionist Pedagogy* and the *Feminist HRI* frameworks, *Robots for Social Justice (R4SJ)* argues that if we want to ensure that our technologies are actually helping to build an equitable future, rather than simply helping those who are already socially and economically empowered, we should cultivate a culture of careful reflection in which we do our best to thoughtfully articulate answers to key engineering questions such as: Who is our technology actually intended to help? Whose capabilities (and which of their capabilities) are prioritized by our research efforts? Do these align with the priorities of the communities we seek to help? Do our technologies actually help advance those capabilities? And what risks and harms are imposed by our technologies? Moreover, E4SJ encourages us to ask these questions in specific ways grounded in engineering practice, leading to additional questions that are less commonly asked in the HRI literature, which we list here for consideration:

(1) How do our technologies increase opportunities and resources for our intended beneficiary communities?
(2) How do our technologies politically empower communities?
(3) What are the structural conditions that constrain the opportunities, desires, and aspirations of their intended beneficiary communities (both in terms of why technologies for *those* communities are well-justified, and in terms of how our technologies stand to subvert those limitations)?
(4) And finally, how are our proposed technologies grounded in contextual listening to communities' stories, values, and desires?

It is our hope that these recommendations and questions can serve as a starting point for HRI researchers to actively engage with communities, and to boldly interrogate the true impact work in our

field could have. Rather than limit the types of HRI research that the community pursues, we intend suggest additional questions that should be explicitly answered during the research process. It must be reiterated that it is not our intention to *criticize* existing work that does not meet the criteria for this analysis, and it is not our intention to minimize the value of and effort behind these papers. Indeed, all of the authors' own work would fail to pass one or more criteria laid out above.

6 LIMITATIONS AND FUTURE WORK

Finally, we acknowledge the limitations of our work. The coding of papers was done by the first author due to the vast scope of our coding efforts. Despite our clear trends, additional confidence could have been gained through the use of multiple coders. There are also several limitations of the E4SJ framework that carry over into R4SJ.

From Leaders to Listeners — First, we acknowledge the nuanced and complex nature of relationships between designers and community members. There is growing advocacy for designers, engineers, and researchers to take on the role of facilitators and collaborators rather than leaders during community engagements. Due to the scope of this analysis, this relationship was not explored. In the future, we would like to explore how this relationship typically plays out for HRI researchers that seeks to enhance human capabilities through contextually listening to and actively collaborating with a specific community (i.e., what roles do HRI researchers typically play? How do different approaches to collaboration affect participation and outcome?).

The Limits of Contextual Listening — Second, and relatedly, it is reasonable to critique E4SJ's clear centering of *contextual listening* to communities. On the one hand, this is critical for much HRI work in the sense that, mere speculations about the potential benefits of one's work to a particular community without talking to that community runs the risk of painting a human-centered veneer over one's research without doing the work of actually assessing the alignment between research and communities' self-expressed needs, values, and priorities (cf. recent critiques of ostensibly human-centered AI initiatives [1, 30]). On the other hand, researchers doing foundational theoretical work cannot be expected to do deep participatory design work with specific communities, and there should be no expectation that their work should be immediately deployable in today's communities. And in fact, some have argued that doing participatory research on technologies that cannot be effectively and immediately deployed could be actively harmful, as at worst it could result in the deployment of technologies that are harmful due to that same nascent status, and at best, could result in wasting the time of communities without helping them. With this consideration, it may actually be preferable for theoretical researchers to tend towards citing the work of others who *have* done the work of documenting communities' needs, values, and priorities, rather than striving for contextual listening themselves. Moreover, as *Design Justice* argues [8], in many cases a truly equitable and de-colonial design practice would not involve listening to and designing for or with communities at all, but would instead involve developing technologies that can be readily hacked and extended by communities, or End-User Programming Interfaces that allow for ready customization and re-programming of mature technologies [45, 55, 56].

The Perils of Dual Use — Third, another concern is that researchers could use the type of justifications articulated in this work to highlight potential pro-social uses of proposed technologies while ignoring potentially harmful dual-uses by other communities. This motivates a need for researchers to more broadly and explicitly consider in their papers the wide range of uses potential technologies might have, both positive and negative (cf. recent discussions of such sections in NeurIPS papers [9, 20, 21]).

The HRI Community is a Community Too — Finally, in our analysis, we excluded papers that aim to serve fellow engineers and researchers. We did this in order to closely adhere to the E4SJ framework and investigate the need to design for more diverse communities, and to acknowledge the power that engineers and researchers hold in this specific context. But helping engineers and researchers is obviously important, and HRI researchers, designers, and engineers certainly represent a community of people with shared values and interests [cf. 22], if not a particularly disadvantaged one due to both the nature of academia and due to the prevalence of researchers from the global north in the HRI community. Future work could even analyze the needs of the HRI research community itself through the lens of E4SJ.

7 CONCLUSION

In this paper, we presented *Robots for Social Justice (R4SJ)*: a framework for an equitable engineering practice of human-robot interaction, grounded in the *Engineering for Social Justice* framework. To understand the new insights this framework could provide to the field of HRI, we analyzed the past decade of papers published at the ACM/IEEE International Conference on Human-Robot Interaction, and examined how well current HRI research aligns with the principles espoused in the E4SJ framework. Based on the gaps identified through this analysis, we made five concrete recommendations and highlighted key questions needed to guide the introspection engineers, designers, and researchers. We believe these considerations are a necessary step not only for ensuring that our engineering education efforts encourage students to engage in equitable and societally beneficial engineering practices (the purpose of E4SJ), but also for ensuring that the technical advances we present at conferences like HRI are true advances as far as our society is concerned, not just fellow researchers and engineers. We hope that our work can serve as an additional instrument in the Equitable HRI toolkit to help our field to pursue a collective research program grounded in Social Justice and the advancement of key human capabilities.

ACKNOWLEDGMENTS

This work was supported in part by Young Investigator Award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research and in part by NSF CAREER Award IIS-2044865.

REFERENCES

[1] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

[2] Rita Armstrong and Caroline Baillie. 2012. Engineers engaging with community: negotiating cultural difference on mine sites. *International Journal of Engineering, Social Justice, and Peace* 1 (2012).

[3] Caroline Baillie and Rita Armstrong. 2013. Crossing knowledge boundaries and thresholds: Challenging the dominant discourse within engineering education. *Engineering education for social justice: Critical explorations and opportunities* (2013).

[4] Cynthia L Bennett and Os Keyes. 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing* 125 (2020).

[5] Jessica R Cauchard, Kevin Y Zhai, Marco Spadafora, and James A Landay. 2016. Emotion encoding in human-drone interaction. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[6] Elizabeth Cha, Anca D Dragan, and Siddhartha S Srinivasa. 2015. Perceived robot capability. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

[7] Herbert H Clark and Kerstin Fischer. 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46 (2023).

[8] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

[9] K Crawford and M Whittaker. 2019. AI in 2019: A Year in Review. *AI Now* (2019).

[10] Nicole Darmawaskita and Troy McDaniel. 2021. Analysis of the Impact of Educational Technology on Social Inequity in the United States. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments: 15th International Conference, UAHCI 2021*. Springer.

[11] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction(HRI)*. Association for Computing Machinery.

[12] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

[13] Gary Lee Downey, Juan C Lucena, Barbara M Moskal, Rosamond Parkhurst, Thomas Bigley, Chris Hays, Brent K Jesiek, Liam Kelly, Jonson Miller, Sharon Ruff, et al. 2006. The globally competent engineer: Working effectively with people who define problems differently. *Journal of Engineering Education* 95, 2 (2006).

[14] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42 (2003).

[15] Wendy Faulkner. 2015. 'Nuts and Bolts and People'Gender Troubled Engineering Identities. *Engineering Identities, Epistemologies and Values: Engineering Education and Practice in Context, Volume 2* (2015).

[16] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14 (2004).

[17] R Edward Freeman. 2010. *Strategic management: A stakeholder approach*. Cambridge university press.

[18] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996).

[19] Daniel Gallenberger, Tapomayukh Bhattacharjee, Youngsun Kim, and Siddhartha S Srinivasa. 2019. Transfer depends on acquisition: Analyzing manipulation strategies for robotic feeding. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[20] Elizabeth Gibney. 2020. The battle for ethical AI at the world's biggest machine-learning conference. *Nature* 577, 7791 (2020).

[21] Elizabeth Gibney. 2020. This AI researcher is trying to ward off a reproducibility crisis. *Nature* 577, 7788 (2020).

[22] KE Hayes and M Kaba. 2023. Let this radicalize you: Organizing and the revolution of reciprocal care.

[23] Jonathan Herington. 2020. Measuring fairness in an unfair World. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[24] Laura V. Herlant, Rachel M. Holladay, and Siddhartha S. Srinivasa. 2016. Assistive Teleoperation of Robot Arms via Automatic Time-Optimal Mode Switching. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction(HRI)*. IEEE Press.

[25] Inês Hipólito, Katie Winkle, and Merete Lie. 2023. Enactive artificial intelligence: subverting gender norms in human-robot interaction. *Frontiers in Neurorobotics* 17 (2023).

[26] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[27] Ryan Blake Jackson and Tom Williams. 2019. On perceived social and moral agency in natural language capable robots. In *2019 HRI workshop on the dark side of human-robot interaction*.

[28] Ryan Blake Jackson and Tom Williams. 2021. A theory of social agency for human-robot interaction. *Frontiers in Robotics and AI* 8 (2021).

[29] Brent K Jesiek, Natascha Trellinger, and Swetha Nittala. 2017. Closing the practice gap: Studying boundary spanning in engineering practice to inform educational practice. In *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE.

[30] Matthew Le Bui and Safiya Umoja Noble. 2020. We're missing a moral framework of justice in artificial intelligence. In *The Oxford Handbook of Ethics of AI*. Oxford University Press.

[31] Florian Leutert, Christian Herrmann, and Klaus Schilling. 2013. A spatial augmented reality system for intuitive display of robotic data. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 179–180.

[32] Jon A Leydens and Jessica Deters. 2017. Confronting intercultural awareness issues and a culture of disengagement: An engineering for social justice framework. In *2017 IEEE International Professional Communication Conference (ProComm)*. IEEE.

[33] Jon A Leydens and Juan C Lucena. 2014. Social justice: A missing, unelaborated dimension in humanitarian engineering and learning through service. *International Journal for Service Learning in Engineering, Humanitarian Engineering and Social Entrepreneurship* 9 (2014).

[34] Jon A Leydens and Juan C Lucena. 2016. Making the invisible visible: Integrating engineering-for-social-justice criteria in humanities and social science courses. In *2016 ASEE Annual Conference & Exposition*.

[35] Jon A Leydens and Juan C Lucena. 2017. *Engineering justice: Transforming engineering education and practice*. John Wiley & Sons.

[36] Jon A Leydens, Juan C Lucena, and Dean Nieusma. 2014. What is design for social justice?. In *2014 ASEE Annual Conference & Exposition*.

[37] Jon A Leydens, Juan C Lucena, and Donna M Riley. 2022. Engineering education and social justice. In *Oxford Research Encyclopedia of Education*.

[38] Grégoire Milliez, Raphaël Lallement, Michelangelo Fiore, and Rachid Alami. 2016. Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 43–50.

[39] Sanika Moharana, Alejandro E Panduro, Hee Rin Lee, and Laurel D Riek. 2019. Robots for joy, robots for sorrow: community based robot design for dementia caregivers. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

[40] Anahita Mohseni-Kabir, Sonia Chernova, and Charles Rich. 2016. Identifying reusable primitives in narrated demonstrations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 479–480.

[41] Ryo Murakami, Luis Yoichi Morales Saiki, Satoru Satake, Takayuki Kanda, and Hiroshi Ishiguro. 2014. Destination unknown: walking side-by-side without knowing the goal. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 471–478.

[42] Martha C Nussbaum. 2009. Creating capabilities: The human development approach and its implementation. *Hypatia* 24, 3 (2009).

[43] Anastasia K Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

[44] Steffi Paepcke and Leila Takayama. 2010. Judging a bot by its cover: An experiment on expectation setting for personal robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 45–52.

[45] Hannah Pelikan, David Porfirio, and Katie Winkle. 2023. Designing Better Human-Robot Interactions Through Enactment, Engagement, and Reflection. In *Proceedings of the CUI@ HRI Workshop at the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI'23)*.

[46] Hannah R. M. Pelikan, Mathias Broth, and Leelo Keevallik. 2020. "Are You Sad, Cozmo?" How Humans Make Sense of a Home Robot's Emotion Displays. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[47] Michael Power. 2003. Evaluating the audit explosion. *Law & policy* 25, 3 (2003).

[48] Brandon Reynante. 2022. Learning to design for social justice in community-engaged engineering. *Journal of Engineering Education* 111, 2 (2022).

[49] Greg Rulifson and Jessica Mary Smith. 2021. Beyond the social license to operate: Training socially responsible engineers to contend with corporate frameworks for community engagement. In *2021 ASEE Virtual Annual Conference Content Access*.

[50] Selma Šabanovic. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010).

[51] Najmeh Sadoughi, André Pereira, Rishub Jain, Iolanda Leite, and Jill Fain Lehman. 2017. Creating prosodic synchrony for a robot co-player in a speech-controlled game for children. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*.

[52] Matthias Scheutz, Bertram Malle, and Gordon Briggs. 2015. Towards morally sensitive action selection for autonomous social robots. In *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

[53] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. 2020. Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020).

[54] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–65.

[55] Emmanuel Senft, David Porfirio, and Katie Winkle. 2022. PD/EUP Workshop Proceedings. arXiv:2207.07540 [cs.RO]

[56] Emmanuel Senft, David J Porfirio, and Katie Winkle. 2022. Participatory Design and End-User Programming for Human-Robot Interaction. In *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.
[57] Amanda Sharkey. 2017. Can robots be responsible moral agents? And why should we care? *Connection Science* 29, 3 (2017).
[58] Solace Shen, Petr Slovak, and Malte F. Jung. 2018. "Stop. I See a Conflict Happening.": A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction(HRI)*. Association for Computing Machinery, New York, NY, USA.
[59] Tiera Tanksley. 2023. *Employing an Abolitionist, Critical Race Pedagogy in CS: Centering the Voices, Experiences and Technological Innovations of Black Youth.* Journal of Computer Science Integration.
[60] Andrew A Tawfik, Todd D Reeves, and Amy Stich. 2016. Intended and unintended consequences of educational technology on social inequality. *TechTrends* 60 (2016).
[61] Leland Teschler. 2010. Why engineers shouldn't worry about social justice. *Machine Design* 82, 15 (2010).
[62] Niels Van Doorn. 2014. The neoliberal subject of value: Measuring human capital in information economies. *Cultural Politics* 10, 3 (2014).
[63] John Vilk and Naomi T. Fitter. 2020. Comedians in Cafes Getting Data: Evaluating Timing and Adaptivity in Real-World Robot Comedy Performance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction(HRI)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3319502.3374780
[64] Tom Williams and Kerstin Sophie Haring. 2023. No Justice, No Robots: From the Dispositions of Policing to an Abolitionist Robotics. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 566–575.
[65] Tom Williams, Cynthia Matuszek, Kristiina Jokinen, Raj Korpan, James Pustejovsky, and Brian Scassellati. 2023. Voice in the Machine: Ethical Considerations for Language-Capable Robots. *Communications of the ACM (CACM)* (2023).
[66] Simon Winberg and Chis Winberg. 2017. Using a social justice approach to decolonize an engineering curriculum. In *2017 IEEE Global Engineering Education Conference (EDUCON)*. IEEE.
[67] Katie Winkle, Donald McMillan, Maria Arnelid, Madeline Balaam, Katherine Harrison, Ericka Johnson, and Iolanda Leite. 2023. 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 (HRI)*.
[68] Langdon Winner. 1980. Do artifacts have politics? *Daedalus* (1980).