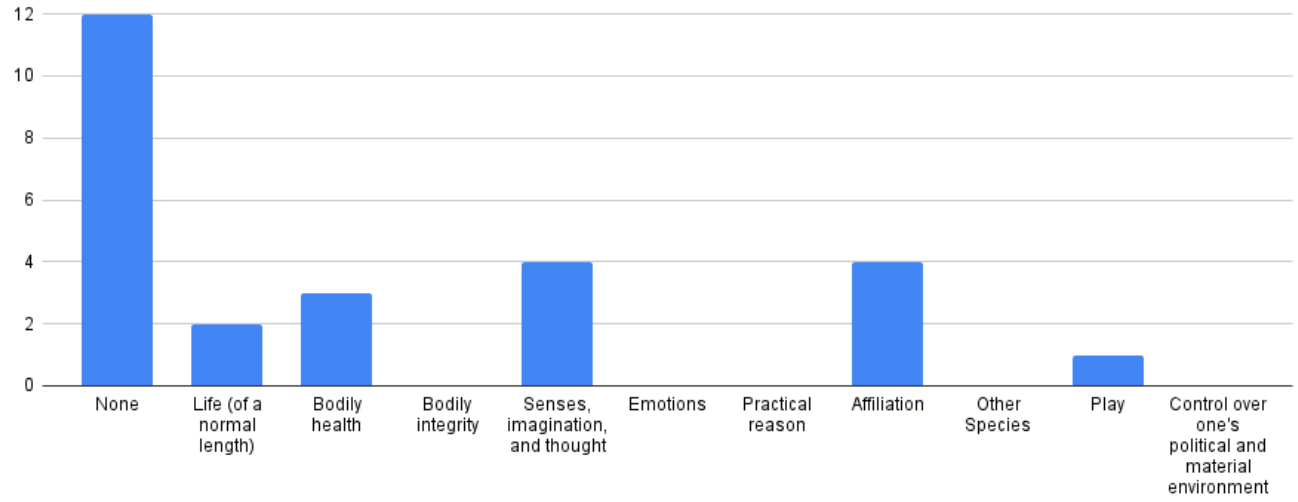

Figure 1: Papers engaging with each of Nussbaum's 10 human capabilities. Papers engaging with multiple capabilities were multiply counted. Papers indirectly engaging with a capability were counted for the purpose of this visualization as fully engaging.

pers presented technologies that either directly or indirectly facilitated *affiliation*, in the sense of encouraging empathy and dignity; **?** and **?**'s approaches were designed with this capability in mind; **?** and **?**'s approaches were designed to achieve other goals *in a way* that advanced this capability. Three of these four papers were among those that specified intended beneficiary communities (**???**) (cf. **?**). Three of the five papers presented technologies that facilitated *bodily health* (**???**), two of which we also categorized as facilitating *life* (**??**). All three were among those that specified intended beneficiary communities.

Of the remaining twelve papers, eleven had concrete motivations not captured by the E4SJ framework: five were motivated by an abstract desire for explainability, three by an abstract desire for trustworthiness, and three by other abstract desires surrounding robot perception, cognition, and behavior modeling. While all of these types of approaches have the *potential* to help create a more equitable society, the lack of articulation of an intended beneficiary community (and thus, subsequently, a lack of specification for how that community was intended to be helped by the technology) evoke a dangerous perspective in which these advances are seemingly cast as beneficial in and of themselves. While it is true that *explainability* and *trustworthiness* are admirable goal in some contexts (e.g., a robot that shares critical systemic knowledge with undocumented communities), these principles can be rendered dangerous when recontextualized 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 for the purpose of surveillance or oppression by corporate or state actors (e.g., police)). Similarly, while *efficiency* can be an admirable goal in the contexts of making robots affordable for low-income communities, serving a greater number of hospitalized children, or enhancing disabled users' mobility, 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. A social justice oriented approach to increasing efficiency in warehouse environments would need to be motivated by a community-provided efficiency concern grounded in one of Nussbaum's 10 human capabilities. For example, such an approach might be grounded in factory workers' fears that robots introduced into the workplace would decrease their efficiency in the sense that their specialized skills could go to waste (**?**). This concern could be justified through its grounding in dignity, autonomy, and *affiliation* that would need to be addressed in particular ways. Indeed, there is good reason to be skeptical of blind emphasis on metrics such as efficiency, effectiveness, and transparency, which are traditionally centered by neoliberal axiologies and theories of value (**??**).

Conclusions: Envisioning a Social Justice oriented AI-HRI

It is not our intent to imply that the papers we've chosen to (implicitly) critique are of poor quality or unethical. None of the papers published at AI-HRI last year were actively malicious or anti-social justice[1]. Neither is it our intent to imply that they do not stand to address key capabilities from Nisbaum's taxonomy in some way. And moreover, most of the critiques leveled in this paper can be readily applied to the authors' own papers. Rather, we suggest that if we want to

[1]Cp. recent papers (**?**) published in the HRI community that have presented technologies actively intended to empower racist, violent institutions, without taking into account the perspectives of those likely to be targeted by the proposed technologies.

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. Who is our technology actually intended to help? Whose capabilities (and which of their capabilities) are prioritized by our research efforts? And do our technologies actually help advance those capabilities?

Moreover, while only briefly touched on in this paper, we intend to suggest that our community should consider how the success of our attempts to advance key human capabilities *equitably* are mediated by the other Engineering for Social Justice criteria:

> 1. What risks and harms are imposed by our technologies? 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 or 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?

Asking ourselves these questions during our research process is a necessary step not only for ensuring that our engineering education efforts (for those of us teaching at universities) encourage students to engage in equitable and societally beneficial engineering practices, but also for ensuring that the technical advances we present at symposia like AI-HRI are truly advances as far as our society is concerned. A collective effort to re-focus on these types of questions may also lead us to reconsider the way that we read, interpret, enact, write our professional codes of ethics, We also admit that there are legitimate drawbacks to the approach proposed in this paper, and that the success of the E4SJ lens depends on the precise manner in which it is employed. A straightforward step that AI-HRI researchers could take to use the E4SJ criteria when motivating their work is to be specific about the specific community they are trying to help, and the specific capabilities their work is intended to advance for those communities. Even for highly theoretical work, researchers could give examples of communities their research would be expected to benefit and in what ways. However, this approach has some clear problems.

In particular, the E4SJ criteria clearly center *contextual listening* to communities. Mere speculation 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 (**??**)). 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. However, theoretical researchers could at least cite the work of others who *have* done the work of documenting communities' needs, values, and priorities.

Another concern is that researchers could use this type of justification 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 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 (**???**)).

Finally, we conclude by envisioning in the section below what a possible future might look like, in which the AI-HRI community aggressively pursued a collective research program grounded in Social Justice, working to develop technical robotics advances that advanced key human capabilities for societally disadvantaged communities.

Speculative Exercise: AI-HRI 2022 List of Accepted Papers (Titles Only)

1. Facilitating **life**: autonomous robot distribution of blankets to **homeless people** in public parks.
2. Facilitating **bodily health**: socially assistive robots for encouraging exercise therapy participation in **older adults**.
3. Facilitating **bodily integrity**: social robots for helping **sex workers** safely report abuse suffered at the hands of law enforcement.
4. Facilitating **senses, imagination, and thought**: **literacy tutoring robots** for **students from oppressed racial groups** attending underfunded segregated schools.
5. Facilitating **emotions**: conversational agents providing a safe sharing environment for **LGBT+ teenagers**.
6. Facilitating **practical reason**: robot-led goal reflection with **first generation college students**.
7. Facilitating **affiliation**: building cultivating environments for **women in STEM** with sexism-rebuking robots.
8. Facilitating **connections to other species**: forest terrain adaptation algorithms for **robotic wheelchair users**.
9. Facilitating **play**: bilingual robots encourage structured play with **immigrant children**.
10. Facilitating **control over one's environment**: building trustworthy robots to encourage census participation in **undocumented communities**.

Acknowledgements

This work was funded in party by NSF grant IIS-1909847. We would like to thank Qin Zhu, Zhao Han, Commodi aliquid nam eos laboriosam tempora tenetur eaque dolore repellendus ipsum, nihil molestias laboriosam perferendis hic provident, voluptas corporis necessitatibus.