Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI

Thomas Arnold
Human-Robot Interaction Laboratory
Tufts University
Medford, MA 02155, USA
thomasarnold@alumni.stanford.edu

Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA 02155, USA
matthias.scheutz@tufts.edu

ABSTRACT

HRI research has yielded intriguing empirical results connected to ethics and how we react in social contexts with robots, even though much of this work has focused on task-based, one-on-one interaction. In this paper, we point to the need to investigate a wider range of ethically relevant dynamics that interaction with robots carries with it – individually and in groups, with a single robot or more. We specifically examine three areas: 1) the primacy and implicit dynamics of bodily perception, 2) the competing interests at work in a single robot-human interaction, and 3) the social intricacy of multiple agents – robots and human beings – communicating and making decisions. While these areas are not exhaustive by any means, we find they yield concrete directions for how HRI can contribute to a widening, intensifying set of ethical debates with critical empirical insight, starting to stake out more of the ethical landscape in HRI.

Keywords

Ethics in HRI

1. INTRODUCTION AND MOTIVATION

Ethics in HRI is expanding with no less complexity than the field of HRI itself. As robots are becoming more autonomous and are being endowed with ever increasing decision-making capabilities, important societal questions about the legal nature and ethical status of robot actions and robot behavior are rising to the surface, not only in mere philosophical debates about agency, but also in technical discussions about autonomy and legal conversations about liability, including the very field of HRI itself.

As multiple disciplines and domains for robotic application have prompted researchers to propose a number of distinct ethical approaches, researchers in robotics and HRI themselves are starting to recognize the need for a systematic overview of their practice, as well as themselves as practitioners, in the form of a code of ethics [30].

HRI '17 March 6–10, 2017, Vienna, Austria

DOI: 10.1145/1235

From the standpoint of public policy and societal deliberation we can investigate how robots entering the social sphere may affect labor markets, healthcare, cybersecurity, and education (to name just a few). While some single out particular roles for ethical perusal, for example nannying or healthcare [34], from the perspective of designing autonomous robots in general HRI has drawn upon landmark attempts to square computational approaches with the best of ethical theory. Researching "machine morality" or robot ethics has illuminated many challenges and opportunities to how autonomous systems could and should decide and act in the world [42].

Yet, another direction for HRI insight lies in how human beings' own reactions to and performance with robots can shape ethical evaluation of their social collaboration. Through realistic and morally charged scenarios, HRI research has examined how humans interpret and evaluate robot behavior, even to the point of applying different moral standards to human vs. robot actions [25]. HRI ethics has also opened out onto dimensions of teamwork and group dynamics, where the impact of the robot is not just through a discrete physical task but in dynamic relationships with people in a certain setting [45]. A common thread in all these aspects of HRI and ethics is the fact that human interests, debates, dilemmas, and conflicts permeate the question "what a robot should do," and that determining good answers is not a technical puzzle but an ongoing, adaptive project. Peering into the ethical landscape of HRI requires a keen eye for overarching themes, from unidirectional emotional bonding to locating responsibility [22, **?**, 9].

Not surprisingly, it has been real life applications like self-driving cars, not just the prodding of science fiction, that ultimately have brought home certain ethical quandaries, including how ethical considerations can come regrettably late in the design and implementation process. Teslas have already been operating on the road, without any kind of ethical provision that would facilitate decision-making in critical situations to minimize harm. It is heavily debated whether their introduction on the road was prudent, safe, and sufficiently transparent to the driving public.

Similarly, domestic robots are being advertised for purchase while leaving in their wake concerns about their purpose and function for their users. For example, is the recently released Barbie chatbot version really intended for meaningless dialogues, or is there an intent to steer the target population – small kids – in particular directions [41]?

It is clear that ethical reflection about the design and use of robots and their possible interactions with humans

must get out in front enough to address these questions as promptly and sagely as possible, lending anticipatory, not belated, insights for public deliberation. These anticipations should not be the nightmares of civilization's takeover via wild projections of human vices and desires onto AI and robotics, but a more distributed, less dramatic, but absolutely critical set of dynamics that social robotics will plausibly engender.

HRI research has yielded intriguing empirical results connected to ethics and how humans tend to react in social contexts with robots. Still, much of this work has focused on short-term, one-on-one interaction, whether it be what happens when a robot is reprimanded or deemed accountable [17], or when a robot objects to a command [4], a robot cheats [35], or how honest we might be with a robot nearby [15]. As critical as it is to follow up on these issues, the ethics of HRI spans spatially and temporally far beyond many of those experimental settings, raising questions in dire need of answers: What are the longer-term effects that ongoing interactions with robots – individually and in groups, with a single robot or more – will have on our everyday lives [8]? How will we recognize and evaluate which interactions are most worthwhile, and toward what kind of life and society they should contribute? And how do we ensure that robots will be genuine helpers contributing to social coherence, rather than social disintegration (with people, for example, more interested in interacting with their individual companion robot than with other humans)?

To begin mapping that terrain of questions, we will examine the areas of ethical significance in HRI: 1) the primacy and implicit dynamics of bodily perception, 2) the competing interests at work in a single robot-human interaction, and 3) the social intricacy of multiple agents – robots and human beings – communicating and making decisions. These areas are not exhaustive by any means and do not occlude the many established ways in which HRI has tackled ethics so far. Still, we find that they yield concrete directions for how HRI can contribute to a widening, intensifying set of ethical debates with critical empirical insight. By providing concrete suggestions for empirical studies to investigate the themes underwriting the three ares we hope that HRI research will begin to broaden and start to address issues that will sooner than later have a big impact on societal structures.

2. THE PRIMAL, FUNDAMENTAL CHANNELS OF EMBODIED PRESENCE

The functionality of robots often and understandably centers on their perceptual, manipular, and action capabilities, maybe augmented by natural language recognition/production, and other cognitive functions such as planning and problem solving. For social robots, however, their very physical appearance and presence comes to the fore, and the particular manner of executing behaviors starts to matter: not only does robot appearance cause humans to make automatic inferences about robot properties (e.g., whether the robot is likeable, smart, etc.), but robot motion is similarly viewed as a cue to the inner life of the machine (e.g., whether it has a goal and started moving in a particular direction for that reason, or whether it is looking for something the observer does not know). When moving in delicate environments, for example, robot physical presence becomes a dimension of safety and effectiveness – how can the robot accomplish a task while not risking others in its ambit? As the implications of robotic presence are observed, especially with an array of different roles imagined for them, a different foundation of somatic factors is coming into view.

Recent developments in soft robotics as well as compliant control enable new types of physical interactions that were not possible before. General proximal presence and particularly touch – which might be called *interactive proprioception* – point to the manifold of physical interactions of human and robots that might be largely submerged from conscious articulation, yet require explicit accounting and investigation. For the role of touch is critically involved in forming attachments and social bonding in primates [11] and physical interactions with robots, especially soft robots, might have similar effects, regardless of their designers' intentions [**?**].

One obvious avenue through which the market for robots is playing up the importance of physicality with robots is seen in the looming emergence of sex robots. With first prototypes already on the market, the HRI community is off to a late start investigating the implications of sex robots for human societies, including issues around physicality, privacy, and intimacy [33]. Though it is certainly important to weigh the abuse and exploitation sex robots might exacerbate [29], cordoning off "sex robots" as a distinct type of interaction is a difficult task. It is no less bodily than many other interactions of intimate care and co-presence assistive robots, for example, are being imagined to conduct. Bathing, or helping someone after a fall, or sitting while one sleeps, or watching a movie on a couch, each have rich implications for what a human being is expecting, relying upon, and projecting for the robot with whom they are sharing space. The trust and vulnerability can be conscious or not, and even simple gestures and acts of touch usher in research from psychology.

Recent preliminary work by Li suggests that physical and tactile presence may elicit more from us than we might imagine, and for reasons we have yet to explain. Robots were found more persuasive and engaging when physically present than onscreen, though virtual agents and physical robots were regarded similarly when both onscreen– not physical form, but presence, seemed to make the difference [20]. Subjects touching robots were found to be aroused in ways they may not have expected or even perceived themselves [21]. HRI research needs to address how a robot's presence and touch – with all the different forms, textures, and sounds that can come with it – affect how human beings regard and evaluate a robot's function. How do the many tactile features of robots, for example, serve as anchors and cues not just for "likeability" or positivity (categories largely used in HRI work thus far) but more targeted attributions and investments (smart about finances or wise about relationships)?

Though the popular press continues to center on sex robots when it comes to tactile relationships, the intimacy of physical movement and contact is charged on more than sexual lines. Its cultural and emotional resonances, which a robot succeeds or fails to hit upon, bear heavily on what kind of familiarity the robot can generate, what expectations a subject might be lured into forming. Psychologists have shown how seemingly insignificant aspects of one's environment can affect feelings and behavior – a warm coffee cup

can facilitate feelings of trust, a trash can with eyes painted on it serve to limit littering [3, 44].

It is not difficult to transfer those lessons to the ways robots' shapes, movements, and function can command powerful reactions. The intense and poignant impact of IED (improvised explosive devices) detecting robot "dogs" on soldiers has an immediate intuitive sense – seeing these robots destroyed or disfigured is a difficult isomorphism to suppress, especially in the grueling, intense environment of a battlefield. But a whole host of less gripping, yet still somatically robust, interactions will present HRI researchers with opportunities for ethical foresight. Subtle configurations of care – what a robot does for a person, how well it performs social rituals and tasks, and what it forces the person to do for it – can have immediate effects and engender longer-term patterns of behavior [37].

The primacy of bodily interaction, with many channels of tactile or proprioceptive feedback the robot may give and receive, may not rise to the level of an urgent ban or utopian promise of a particular role for a robot. As Turkle has recently pointed out with respect to conversation, it can be more subtle and gradual transfigurations in use of technology that gain more permanent purchase on culture [39]. How will forms of human-robot interaction in sensitive areas – especially with children and elderly people – alter expectations of personal space and safety? What forms of information can the robot gather, and what forms can it solicit, through its whole repertoire of sensory outreach? A verbal response of sympathy has certain rules and conventions according to subject matter. But how does that change when a reassuring, or offputting, backpat is at issue? The many concerns about abuse and lack of consent that eldercare facilities face merely presage the complex negotiations of what touch, personal space, and consent could and will mean for robots in that domain. Some of these considerations may seem to invite technical advances of precision. How do we get the robot to deliver the tea more smoothly, for example? But the ethics involved remain thorny. Consider the following scenario as one point of departure.

SCENARIO 1.

A domestic robot operates in the living room, where family members spend most of their time. The robot maneuvers around people who are on the floor, but has various points of the room where it stays still. The robot has a repertoire of gestures when engaged with a person, including a shoulder pat for assurance and a high five for fun. The robot also has accompanying sounds it can play, including one of the company's jingle that designers had play during high fives or when someone touched the robot and looked happy. The robot and child play a hand-clapping game, but as it gets faster the child makes a mistake. The child pushes the robot in frustration and the parent yanks the child away angrily. The robot recognizes the child's distress and puts a game up on its chest screen for distraction, then motions for the child to sit in its warm lap to watch.

This scenario hints at the ways tactile, mobile interactions could have physical and relational ramifications, both immediate and possibly longer-term. What relationships between child, parent, robot and living space will develop, and how can that be evaluated? In what types of related situations would a robot's action violate norms of intimacy, family bonds, consent, or dignity, or at least wear away at common means of establishing good relationships between people? What is the line between a manufacturer designing multiple physical means for engagement, and manipulating users' vulnerabilities? HRI research has long recognized different aspects of fluency and disfluency in how a robot is perceived, whether through language, form, and gesture [32], but must probe into even more constellations of features that could be shaping an interaction: subtle movements, postures, orientations, temperature, surface texture, etc., including the psychological effects (whether positive or exploitative, upbuilding or dignity-depriving) of a robot changing some of its features in response to a person's affect.

For the above scenario in particular, we see the research challenge of getting beyond the "one-shot lab visit", although even a short interaction could elicit visceral reactions (such arousal as in the above-mentioned study) and thus hint at a potentially larger aspect in need of further investigation. To whatever degree actual homes are viable exercises of testing robots "in the wild," or researchers can manage the careful and responsible use of public space, recognizing the richness of physical contexts will be key to tracking the means by which such physical interaction is defined [27, 31]. Through a repertoire of physical cues one could test how to distract a child, or alternatively test how best to facilitate the child's contact with other people in the room. The ways in which a robot's presence can anchor a social context will require sustained fine-grained variations of setups and a multi-modal means of observation, with attention to the different practitioners who inhabit the interaction's intended context [7]. Finally, discerning whole-body interaction will also require capturing longer-term dynamics of relationships. Drawing conclusions in HRI about the dynamics of trust, intimacy, and isolation between and among people and robots clearly calls for more longitudinal strategies in the research.

3. COMPETING INTERESTS OF INTERACTION – WHAT INTERESTS INTERSECT IN ONE-ON-ONE INTERACTION?

Given the intimate nature of many service contexts envisioned for robots (whether medical, educational, or domestic settings), HRI work has identified many communicative challenges a successful social robot should meet, including how fully a user's responses should be anticipated. Across different tasks and interaction contexts, the robot's general goal is to understand the intention of the person and to communicate in a clear, accountable way about how it will act. This coordination usually applies to a task or problem that the person and robot can face together. In the past several years, there has been more thorough acknowledgement of how social and institutional backdrops complicate what "accountable" communication actually comprises. Is the robot working at the behest of the individual interactant, or, say, a medical provider (or loved ones) who bought the robot, or the larger community? Whose interests are being served and accounted for in how the robot represents and performs its actions? And what kinds of information, including ways of gaining it, does a robot have to share with an interactant?

From the other direction, what projections or attributions will a person have toward a robot that contort ordinary ethical associations with interaction (e.g., politeness, trust, co-

operation)? How will people represent their own interests to a robot, and when will they pursue their interests warily and in more adversarial fashion? Can they, or should they, keep a secret [18]? Do they consider the robot a witness, with some degree of moral authority? Or is a robot viewed as a suspect tool of more distant authorities "behind the code?"

As robots progress in both their ability to process feedback from a person – intention, emotional state, physical condition, and so on – possible therapeutic and social roles have appeared more vividly in the public's imagination. The movie *Robot and Frank*, for example, shows a robot dutifully serving an elderly man as companion and vigilant attendant in terms of diet and health, but having no clue as to what societal norms it might violate as Frank's willing companion. While the movie puts those violations in criminal (and cinematically entertaining) forms, there is a more fundamental, expansive question of which this is only a variation: which interests of a human-robot interaction define its one-on-one space, and which ones take precedence?

When the question "What should be done?" is aimed at common-sense tasks and conversational patterns, this question seems to be a technical question of getting from input to generally agreed-upon and straightforward output (e.g., fetching the right object). But companionship through a social robot, again, entangles itself with moral norms about confidentiality, trust, and transparency. What kinds of information should a person expect a robot companion to answer about itself, and what should a robot be expected to do with the information the person shares with it? Making interests explicit – whether personal health, or public safety, or remedying loneliness – can abstractly identify a robot's chief purpose. Some HRI research has already delved into how a robot could effectively refuse or protest a command based on represented interests [4].

The more realistic ethical problem to work out, however, is how multiple interests will be prioritized. The way therapists and caregivers develop professional boundaries provides some insight about how interactions can honestly convey and defend those priorities, yet for a robot there is not the same personal/professional interest to settle. Indeed, robotic abilities (of unseen surveillance, of extraordinary perception, of networked communication) and voids (of emotional and physical pain) may adversely affect a user's expectations of what different interests a robot could represent.

In certain therapeutic contexts the coordination of multiple interests can come together in innovative and inspiring ways. Demiris's work with occupational therapists and their patients shows how an autonomous system can incorporate therapists' interest to thwart a patient's immediate interest (having a vehicle automatically keep them from hitting an obstacle) in the service of the patient's longer-term, developmental interest (improving in their ability to steer the vehicle themselves) [10] . The transparency is not essentially deceptive – the child patient could understand how the system was designed and still work with it just as, if not more, effectively. But in cases where a user or patient (and their status can be important to specify) contends with interests in tension, or is in conflict with others around them, it is much less clear what "the right thing to say and do" would be on the part of the robot. There may be a therapeutic interest in getting the person to decide to do one thing or another, followed by an explicit response from the robot explaining why it will agree to perform accordingly. But it may be that the robot more properly has to pull back entirely from the one-on-one nature of the exchange, instead of acting in the stead of people or institutions that may need to represent interests directly. The prospect of more social police robots, where the question of authority takes on intricate, life-and-death forms, brings this vividly into view [16].

Important to note about interests, especially for HRI, is their temporal character. Just within a single interaction, at one point in time, interests can mesh and conflict. But interests, especially longer-term ones, may be seen to develop through interaction over time. While there may be no hard and fast line between those two type of interest intersections, in this paper we would stress that even a single interaction has interests in play.

The following scenario presents the challenge of negotiating different interests in a concrete social context.

SCENARIO 2.

A robot attendant works in a senior center common area, most often providing companionship and performing basic tasks of bringing food, books, blankets, etc. to residents. There is a particular resident who has started talking to the robot at length about their life and their feelings, and the staff has noticed an improvement in the resident's mood since these conversations began. The resident relies on the robot to remember previous conversations, including meaningful personal experiences they recounted. Senior center staff depend on the robot to notice and report certain obvious symptoms of physical distress or erratic behavior, and the robot is designed to enforce the center's rules in the common area. Recently this resident has brought high-sugar foods to eat in this common area, food that medical providers and the resident's family have asked not be available to them. Other residents have taken note of this food as well as the resident's frequent conversations, and staff have been asked by other residents to get more time with the robot. One evening, the resident asks the robot to stay beyond usual common area hours, to discuss a "private" matter they want the robot to keep confidential. They say it concerns their physical health and their treatment, and they feel only the robot, not the family or doctors, will listen and perhaps give good advice.

This scenario points to the difficulty of restricting the conception of human-robot dialogue to a task-specific or problem-specific framework. The one-on-one conversation proposed at the scenario's end is enough to expose a knot of ethical questions about what is best for all those with something at stake in the interaction. Even without forecasting extended interactions between the resident and robot, one can evaluate how this interaction could defend and threaten different interests. The idea of a solution seems less applicable than a negotiation, where possibly no one interest will be perfectly upheld. Personal information from the resident, who perhaps will complain, confess, conspire, or some combination thereof, will tax any simple role of companionship. This will not just be a case of individual vs. society, though it may be that too – it will involve how the individual's different interests themselves (including relationships they may, in a more secure and reflective mood, want to strengthen) will be met. Nor will HRI measures like "rapport" be goals

in and of themselves – depending on the interests involved, an interaction may need to be more adversarial than cooperative. It may be better to have the robot state the limits of its confidentiality and perhaps disappoint the resident than to avoid necessary conflict because of an imagined bonding that is impossible. This latter point applies as much to those who design and deploy the social robot as to those who interact with it.

The overall research upshot for HRI is the imperative of testing how interests extend the scope of interaction beyond a single identified task or purpose, as well as how they redefine what "successful" communication and work may need to look like. Even when limited to a one-on-one interaction, HRI research can look more acutely into how interactions proceed when challenges or appeals to different interests (both those of a subject and broader social ones) are introduced, and how people's reflection on interest may affect their attributions and expectations toward robots. This can help a more robust and grounded HRI ethics weigh in on why, and in what contexts, a robot needs to act on personal disclosures, and when to occupy a stance of confidentiality. This may inform design choices about modes of reception, for example retentive listening (for one-on-one recounting) vs. mere receptive listening (in case the person wants no memory of their words on record). While it will be challenging to craft personal interactions that do not make subjects themselves vulnerable, it will be increasingly critical to track how different interests, not just measures like rapport, agency, or likability, reveal the dynamics that human-robot interaction contains.

One major challenge in this context will be the development of experimental paradigms that are informative and rigorous while not endangering vulnerable populations. For it would be an ethical violation to experiment with the very group of people robots are supposed to support. Rather, experimental designs need to be targeted at key aspects of one-to-one interaction that can be probed with other subject populations within and outside carefully crafted social contexts (e.g., setting up social situations where subjects feel compelled to voluntarily disclose information to the robot that they do not want to share with others).

4. SYSTEMS AND GROUPS

The ethical reflection on a human-robot interaction tends to broaden the intuitive boundaries of that interaction – there are always wider environments and yet more principles and consequences to fold into one's account. As the social complexity of robotic work in various fields has emerged, the idea of teamwork has naturally drawn research interest. The organizational and computational tasks of coordinating a robot's role with multiple human colleagues (if not also multiple robots alongside them) demand several priorities to be made explicit. How does robot participation maximize efficiency? What are the most effective ways for robots to give and receive information in such a multi-agent context? How can robots enable the team and its human teammates to perform better?

While the literature on human-robot teams is breaking interesting ground across a variety of application domains, some of the morally-ambiguous dynamics of group decision-making have only begun to surface [1]. On one level, there are the wrinkles of resolving different perspectives and contributions in the course of deciding on a plan of action. There is a process/product distinction that ethical evaluation must recognize: what is the best decision for the team vs. what is the best method in general for reaching good decisions as a team going forward. And how does the robot communicate productively while navigating the authoritative rules of whose ultimate view holds sway? Within those questions lie matters of social status, authority, knowledge, trust, camaraderie, and integrity, some of which clearly would not apply to decision support systems, but they do apply to robots perceived as *agents*.

Increasingly robust settings have been found for social robots to provide organizational help, for instance a Nao robot assisting with scheduling at a busy medical practice [14]. There is a general recognition by the people interacting with the robot in such a case about what kind of information the Nao can deliver (though its performance may still surprise and impress). In other cases on the horizon, however, a robot may face unexpected negotiations amid conflict. What happens, for example, when multiple incompatible commands or requests are given to the robot by different people? There may be stock answers available, such as "This is who I am authorized to follow on this issue;" some research has looked at verbal feedback and improved performance [36]. But the exchange of reasons, enhanced when robots (and there may well be many on a team) exclusively possess information needed for deliberation, runs against the grain of pure obedience. Nor will reason-giving operate the same way in every context – there may be judgment calls to make when more evidence is practically beside the point. Relational communication and decision-making form part of this horizon not because of incorrect interpretation or incongruent purposes among the interactants, but because robots might be observing, mediating, and participating in the midst of human relationships. The status of robots, instead of a metaphysical context-free question, becomes one of social performance within a context, with their actions having indirect impact on behavior between people, not just attitudes toward the robots themselves [5]. Systematic indices of performance, therefore, must not focus on what people do with robots to the exclusion what they subsequently do with each other. We can consider the following scenario to draw out the issues further.

SCENARIO 3.

A rescue team – with a few members of the National Guard along with civilian volunteers, with the permission of National Guard command center – is patrolling neighborhoods after a flood. Due to the severity of the disaster the team will be working together for days if not weeks. A robot is brought along, with natural language capability, networked communication, and versatile mobility in water or on land. At various points team members will yell to the others that they have come upon a stranded pet, person, or structure under threat (with likely people inside). While one member of the National Guard is the designated leader, even the leader needs the rest of the team to update her in order to make a plan. The robot can often receive evidence or information that contradicts that of team members, including which area needs attention first. The team also faces deliberations about how best, given the state of the area, to triage for the needs that the team perceives around it.

This scenario encompasses some of the issues raised above

about sizing up different interests, while illustrating additional complexities in terms of decision-making and collaboration. Unlike most discussions about the ethics of self-driving cars, treating this scenario means factoring in actual ethical reasoning and debate as part of a robot's participation (or least environment in which they participate). How does a robot negotiate the many consequences for following one direction over another, especially given information it alone might have? How can the robot best enable strong teamwork and better group deliberations, or at least not impede them? At what point does a robot assert its authoritative information into a debate, even though it has no formal social status? And will such insertions result in the robot's obtaining authoritative status among the team members? Unforeseen and severe circumstances will bring these questions all the more to the fore given life-and-death stakes under time pressure. The interlocking of social forms with approaches to ethical reasoning could get quite convoluted without focused empirical work and accurately simulated scenarios.

HRI research has certainly broached some issues of communication and cooperation, tackling important questions of culture and multiple ways of measuring performance [2, 43, 12]. Still, many of the measures in this work boil down to basic attributions: trusting/not trusting, ingroup/outgroup, in favor/against, etc. But part of a true evaluative trajectory must address both process and product, and do so over enough time to capture how the human-robot work can reliably develop over many crucial situations. A one-shot achievement of teamwork is hard to imagine, much less stipulate, and the ethical stakes of the group work will include both task-specific and relational dimensions – including, crucially, relationships between people working with or among robots. The challenge for empirical testing, then, is to configure group settings and complex tasks wherein varied interactions, human-robot and human-human, can be observed and tracked along practical lines. Tracking in these setting could be more than simple attributions from team members, but objective events in how the team deliberates, decides, and discovers ways to work. One way to do this may be complex puzzle situations for groups to tackle (similar perhaps to the interactive game *Escape the Room*), where certain relational goods and team goals ("cooperate," "get to know one another") accompany a main task (e.g., "get across the ravine together"). Without some kind of more organizationally robust measures, the important ethical determination of how robots and humans can work, and in what domains, will have much less of an empirical foundation. There will be much less substance available, in sum, to flesh out what a "morally competent" robot even means [24].

5. RESEARCH CHALLENGES AND MOVING BEYOND REPLACEMENT

The horizons for HRI ethics we have described are not easily reached from where we stand today, though the sophistication of social robots they assume will also not arrive promptly tomorrow. Meeting the next level of ethical challenges will require imaginative, rigorous efforts to enable solid empirical footing for society's next steps with technology. It is worth identifying some common characteristics for research that delves into the three areas we have described, as a way not only to contribute to HRI ethics, but also to support its distinctive contributions to society's wider discussions of technology.

To begin with, studying fully embodied interaction, interest-laden communication, and group performance suggests rethinking the time frame for an interaction and its effects. Physical presence with a robot, especially the social setting that may evoke multiple interests, may be briefer or longer than a standard task-based period (e.g. 30 minutes) in order to register the types of dynamics we have discussed. In addition, however long laboratory or in-the-wild interactions are, the effects of interaction may need to be followed for longer periods than a typical lab visit. Unquestionably the longitudinal approach poses practical and logistical obstacles, just as it does with other forms of scientific research. But HRI work has broken ground in this area, for example on the habituation effect [19]. More to the point, as public debates about robots heat up, anecdotes or personal speculation will not buttress reliable conclusions about what robots are doing, nor where they should be doing it. HRI can expand the public's ethical lens to interactive elements that are more subtle, yet more empirically based, than the headlines.

Another key feature of HRI research for the issues we raise is re-examining the implicit premises of a one-on-one human-robot interaction. While a good deal of HRI research already extends beyond dyadic interactions, changing the number of people or robots does not always provide enough contrast to that model. As we have discussed with respect to interests, a one-on-one interaction may be a crossroads for broader conflicts rather than a self-contained exercise and task; making some of these conflicts and consideration more explicit can test how contrived our previous models of interaction have been. In addition, if human beings are interacting on so many physical and affective levels with robots, other forms of technology (smartphones, computer, TV, game consoles) seem all the more called-for as controls for direct interactants, or at least part of the setting in which to test human-robot interaction. A recent article about the iPal robot has been met with some alarm in terms of childcare, but what if the operative contrast is not with a babysitter but a tablet [46]?

Even as they expose new depths to one-on-one interaction, these areas should allow robotic work to be viewed as more than a one-to-one replacement for a human being's role. Navigating the complex currents of embodied interaction, cross-cutting interests, and interpersonal conflict may mean reshaping how robots have been envisioned to share space and work. Auxillary or peripheral functions for social robotic work, from physical presence and movement to higher-level communication, can open up application domains to richer models of robots complementing people's efforts in service to people's needs. By contrast, testing those different positions can show where independence and contention, perhaps to correct a group's belief, can best contribute in collaborative efforts. Robots do not have to be teammates to work with a team, especially given the ethical and empirical question of how the whole range of physical presence with a robot can affect others. Keeping account of such interactive layers will help determine, in turn, when a robot can effectively share, solicit, or exchange reasons amid different sensory environments, especially within conflicted and ethically charged situations. .

All in all, these research dimensions help to demonstrate,

for the sake of larger ethical discussions, HRI's distinct contributions. While some public narratives, fed by Terminator photo-topped articles, might grab attention through familiar futurist projections, this distracts from the concrete contexts where practitioners and engineers are plying their efforts. Instead of taking common stories from movies and literature and slapping them onto present-day work, the sober challenge is to recognize the ordinary struggles and stakes that accompany proximal applications. As HRI research will be able to better demonstrate 1) how many modes of bodily presence exert a mix of influences cognitively, emotionally, physiologically, and perceptually, 2) how a direct dialogue is a node in a wider network of interests, and 3) how communal and socially spontaneous decision-making can be, we can start to more comprehensively assess what application domains are permissible or preferable for social robots. Instead of repeating the foreboding headline "The ___ robots are here" and assuming a loss of a human occupation or role – whether as teacher, tutor, therapist, lover, or companion – we can more carefully sort through physical, communicative, and assistive roles in relation to human dignity and needs. As AI ethics struggles to connect to people's experience in face of inaccessible algorithms and practically disembodied systems working at a distance, HRI will have even more reason to bear witness to the embodied, culturally embedded, and communicatively dense conditions in which social robots operate.

This is both an opportunity and challenge at the same time, for social robots will, by way of their physical bodies, have the advantage of being able to communicate and interact in ways that look natural and familiar to us (assuming these interactions are done right), while disembodied AI systems have to overcome the perceptual hiatus of lacking bodily communication. Yet, while for us the bodily presence of robots is immediately readable, it can also be misread. The job for HRI ethics is exactly to stake out the territory of influence where physical presence and physical interactions, one-on-one or in groups, in organized teams or loose clusters of people, can affect humans psychologically, both in the short and the long term.

6. CONCLUSIONS

HRI ethics will continue to convene inquiries that draw on law, policy, economics, psychology, medicine, education, and popular culture. Sizing up the effects of HRI and its prospects for increased application demand that ethical reflection stay conversant and vigilant across disciplines. Ethics must also be institutionally permeative, integral from the beginning of design processes and in the mix of academic and non-academic treatments.

In this paper we have explicated three challenging dimensions that will pose increasingly significant sets of questions for social robotics. They are practically compelling all the more for not being intuitively attention-grabbing: concrete subtleties of embodiment, competing interests, and decision-making in groups may seem too ordinary in the face of science-fiction's modes of heroic or apocalyptic transcendence. But the varied social contexts into which robots may enter suggest that overly individualized scenarios, with thin articulations of embodiment and social interests, are not sufficient guides for the ethical issues that will develop. Products that grab headlines according to familiar roles (i.e., an adorable domestic companion) can still usher in physical and interpersonal subtleties whose effects reverberate beyond the marketed use (much as smartphones did not show car drivers using them compulsively). Against the grain of some of the disembodied ways that AI is presented in public discussion, HRI research is poised to amplify its ethical voice, witnessing to the multi-layered physicality and expressiveness – part of its deep-seated "social valence" – to show how we can situate robots in the world [6]. Through research reaching further into those depths, HRI work can continue to illuminate and enact humanity, for the sake society's best interests and most genuine needs.

7. ACKNOWLEDGMENTS

This work was in part funded by grant N00014-14-1-0144 from the US Office of Naval Research and grant IIS 1316809 from the US National Science Foundation.

8. REFERENCES

[1] R. C. Arkin, P. Ulam, and A. R. Wagner. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3):571–589, 2012.

[2] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[3] M. Bateson, D. Nettle, and G. Roberts. Cues of being watched enhance cooperation in a real-world setting. *Biology letters*, 2(3):412–414, 2006.

[4] G. Briggs and M. Scheutz. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, 6.

[5] K. Caine, S. Šabanović, and M. Carter. The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pages 343–350. IEEE, 2012.

[6] R. Calo. Robotics and the lessons of cyberlaw. *Cal. L. Rev.*, 103:513, 2015.

[7] W.-L. Chang and S. Šabanović. Studying socially assistive robots in their organizational context: Studies with paro in a nursing home. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 227–228. ACM, 2015.

[8] E. Datteri. Predicting the long-term effects of human-robot interaction: A reflection on responsibility in medical robotics. *Science and engineering ethics*, 19(1):139–160, 2013.

[9] M. M. de Graaf. An ethical evaluation of human–robot relationships. *International journal of social robotics*, 8(4):589–598, 2016.

[10] Y. Demiris. Knowing when to assist: Developmental issues in lifelong assistive robotics. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3357–3360. IEEE, 2009.

[11] R. I. Dunbar. The social role of touch in humans and primates: behavioural function and neurobiological

mechanisms. *Neuroscience & Biobehavioral Reviews*, 34(2):260–268, 2010.

[12] V. Evers, H. Maldonado, T. Brodecki, and P. Hinds. Relational vs. group self-construal: untangling the role of national culture in hri. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pages 255–262. IEEE, 2008.

[13] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology*, 4:859, 2013.

[14] M. C. Gombolay. Apprenticeship scheduling for human-robot teams. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] G. Hoffman, J. Forlizzi, S. Ayal, A. Steinfeld, J. Antanitis, G. Hochman, E. Hochendoner, and J. Finkenaur. Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 181–188. ACM, 2015.

[16] E. E. Joh. Policing police robots. *UCLA L. Rev. Discourse (2016), Forthcoming*, 2016.

[17] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson. Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 33–40. ACM, 2012.

[18] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, H. E. Gary, and J. H. Ruckert. Will people keep the secret of a humanoid robot?: Psychological intimacy in hri. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 173–180. ACM, 2015.

[19] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn. Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 564–569. IEEE, 2007.

[20] J. Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015.

[21] J. Li, W. Ju, and B. Reeves. Touching a mechanical body: Tactile contact with intimate parts of a humanoid robot is physiologically arousing. 2016.

[22] P. Lichocki, A. Billard, and P. H. Kahn. The ethical landscape of robotics. *IEEE Robotics & Automation Magazine*, 18(1):39–50, 2011.

[23] P. Lin, K. Abney, and G. A. Bekey. *Robot ethics: the ethical and social implications of robotics*. MIT press, 2011.

[24] B. Malle and M. Scheutz. When will people regard robots as morally competent social partners? In *Proceedings of Ro-Man*. 2015.

[25] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 117–124. ACM, 2015.

[26] J. Millar. Technology as moral proxy: Autonomy and paternalism by design. In *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on*, pages 1–7. IEEE, 2014.

[27] N. Mirnig, E. Strasser, A. Weiss, and M. Tscheligi. Studies in public places as a means to positively influence people?s attitude towards robots. In *International Conference on Social Robotics*, pages 209–218. Springer, 2012.

[28] I. R. Nourbakhsh. *Robot futures*. MIT Press, 2013.

[29] K. Richardson. The asymmetrical'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society*, 45(3):290–293, 2016.

[30] L. D. Riek and D. Howard. A code of ethics for the human-robot interaction profession. *Proceedings of We Robot*, 2014.

[31] S. Sabanovic, M. P. Michalowski, and R. Simmons. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 596–601. IEEE, 2006.

[32] M. Salem, K. Rohlfing, S. Kopp, and F. Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 RO-MAN*, pages 247–252. IEEE, 2011.

[33] M. Scheutz and T. Arnold. Are we ready for sex robots? In *The Eleventh ACM/IEEE International Conference on Human Robot Interation*, pages 351–358. IEEE Press, 2016.

[34] N. Sharkey and A. Sharkey. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–190, 2010.

[35] E. Short, J. Hart, M. Vu, and B. Scassellati. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226. IEEE, 2010.

[36] A. St Clair and M. Mataric. How robot verbal feedback can improve team performance in human-robot task collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 213–220. ACM, 2015.

[37] L. Suchman. Reconfiguring human-robot relations. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 652–654. IEEE, 2006.

[38] S. Turkle. *Alone together: Why we expect more from technology and less from each other*. Basic books, 2012.

[39] S. Turkle. *Reclaiming conversation: The power of talk in a digital age*. Penguin Press HC, 2015.

[40] A. van Wynsberghe. *Healthcare Robots: Ethics, Design and Implementation*. Routledge, 2016.

[41] J. Vlahos. Barbie wants to get to know your child. *The New York Times. Retrieved from http://www.nytimes.com/2015/09/20/magazine/barbie-wants-to-get-to-know-your-child.html?_r=2*,

2015.

[42] W. Wallach, C. Allen, and I. Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, 22(4):565–582, 2008.

[43] L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson, and P. Hinds. When in rome: the role of culture & context in adherence to robot recommendations. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 359–366. IEEE Press, 2010.

[44] L. E. Williams and J. A. Bargh. Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901):606–607, 2008.

[45] J. R. Wilson, T. Arnold, and M. Scheutz. Relational enhancement: A framework for evaluating and designing human-robot relationships. In *Proceedings of the AAAI Workshop on AI, Ethics, and Society*, 2016.

[46] J. Wong. 'this is awful': robot can keep children occupied for hours without supervision, 2016.