Gr Interakt Org (2022) 53:359–366
https://doi.org/10.1007/s11612-022-00645-6

GIO Gruppe. Interaktion. Organisation.

HAUPTBEITRÄGE - THEMENTEIL

Extended norms: locating accountable decision-making in contexts of human-robot interaction

Thomas Arnold[1] · **Matthias Scheutz**[1]

Accepted: 16 August 2022 / Published online: 7 September 2022

Abstract
Machine ethics has sought to establish how autonomous systems could make ethically appropriate decisions in the world. While mere statistical machine learning approaches have focused on learning human preferences from observations and attempted actions, hybrid approaches to machine ethics attempt to provide more explicit guidance for robots based on explicit norm representations. Neither approach, however, might be sufficient for real contexts of human-robot interaction, where reasoning and exchange of information may need to be distributed across automated processes and human improvisation, requiring real-time coordination within a dynamic environment (sharing information, trusting in other agents, and arriving at revised plans together). This paper builds on discussions of "extended minds" in philosophy to examine norms as "extended" systems supported by external cues and an agent's own applications of norms in concrete contexts. Instead of locating norms solely as discrete representations within the AI system, we argue that explicit normative guidance must be extended across human-machine collaborative activity as what does and does not constitute a normative context, and within a norm, might require negotiation of incompletely specified or derive principles that not be self-contained, but become accessible as a result of the agent's actions and interactions and thus representable by agents in social space.

Keywords Norms · Human-robot interaction · Ethics · Explainability · Artificial intelligence

1 Summary

This article in the magazine Gruppe. Interaction. Organization. (GIO) presents a framework for understanding and upholding norms within human-robot interaction. We argue that designing robots to operate with a system of "extended" norms offers more ways to maintain accountability and reliability for interactive, autonomous systems.

2 Introduction

The promise and challenge of machine ethics has long been a matter of mapping what "ethics" is onto the technical realities of what robots or systems could plausibly execute (Wallach and Allen 2008). While science fiction has offered scenarios of machines reaching decisions by principles or instructions (haplessly, Asimov's stories of such laws remind us), the recent trends of machine learning toward massive training data and statistical patterns have rendered ethics more a matter of auditing data for bias than challenging the use of data as such. Regardless, the performance of a system is still largely evaluated through a behavioral criterion, not that of explicit justification on the part of the system in real-time (or through thorough routine ethical testing, which may be required for some autonomous systems, see Arnold and Scheutz (2018)). And as behavioral results meet the complexities of social contexts (e.g., the roads on which self-driving car are meant to drive and transport), machine ethics has increasingly seemed to be an overly ambitious goal relative to basic criteria like safety and functionality. Even as some may call for more exactly organized training data (Tolmeijer et al. 2020), it is harder to see how many more *types* of training a system would need to manage the varied types of tasks autonomous operation could require (e.g., a self-driving car communicating it was stuck and that a car behind it should go around).

✉ Thomas Arnold
thomas.arnold@tufts.edu

Matthias Scheutz
matthias.scheutz@tufts.edu

[1] Tufts University Human-Robot Interaction Laboratory, Department of Computer Science, School of Engineering, Tufts University, 420 Joyce Cummings Center, 177 College Avenue, Medford, MA 02155, USA

As Meyer et al. have recently pointed out, there are various hybrid forms of autonomous systems and human-machine ensembles that the term "robot" tends not to register (Meyer et al. 2022). Systems that feature artificial intelligence need not be embodied, mobile, and interactive within shared physical space. Running throughout such systems, however, are various levels of autonomous decision-making. And on a computational level, trying to make decisions according to various ethical theories or norms can seem infeasible—how can one adjudicate between the preferences and norms between two people in public, much less between groups or interests whose stated ethics may be shared but interpreted as requiring conflicting actions? Nonetheless, work in human-robot interaction has consistently showed how much agent attribution robotic systems can incur through their presence in shared space, including blame judgments that can attend their carrying out of tasks (Malle and Scheutz 2020). If robots—construing that term widely—are to enter a world of human agents making decisions with different levels of coordination and communication, they will perforce be enmeshed with norms and the consequences of violating them. What are norms? Malle et al. have recently formulated a useful definition: "A norm is an instruction, in a given community, to (not) perform an action in a specific context, provided that a sufficient number of community members demand of each other to follow the instruction and do in fact follow it." Norms thus prescribe action and represent a communal standard by which actions are measured and sanctioned. The attributions of agency and capabilities toward robots (and various forms of autonomous systems, including human-machine teams) can induce the accompanying practices of blame and judgment for actions outside various norms. That includes not sanctioning or expressing resistance to violations on the part of other agents (Voiklis et al. 2016). Norms are more subtle than hard and fast rules, and they span morally charged actions (helping a person who was fallen in the street) to more basic ways to fit in socially (keeping one's place in a checkout line). In addition, there can be multiple ways to uphold (or violate) a norm, together with a broad set of implicit conditions and priorities in light of other norms (my standing quietly in line may not be as important if the person behind me has fainted). And while part of norm compliance entails an ability to speak to one's intention, especially in the wake of bad execution, it is possible for an agent to fulfill a norm without any explicit appeal to norms at all. One can stand in line at the supermarket unintentionally, but in just as competent a manner as someone deliberately trying to keep the line intact. The former only coinciding with norm obedience, not *actually* following the norm. Machine ethics, therefore, must span effective actions, intentions, and accountable explanations of how decisions are made. Mapping these demands with the help of the broader literature on social norms is a suggestive research horizon (Legros and Cislaghi 2020). Given recent trends in artificial intelligence research across academy and industry, machine ethics has been pushed less toward norms and more toward "preferences" (e.g., expressed via utility functions), learned through ample training data (e.g., via variants of reinforcement learning) (Russell et al. 2015). The large AI language model like GPT-3, which performs various tasks of natural-language processing (NLP) through training on sequences of words, may yet lack a great deal of common sense or basic knowledge of the concepts it articulates. Nonetheless its training may enable such a system to generate incredibly plausible stretches of text when given related prompts. What if similar training applied to ethics? Perhaps there just needs to be a massive exploration of exhibited human preference in the world based on behavior, rather than ironclad rules that have ambiguous application. Machine ethics seems caught between the complexity of the world that an AI system needs to grasp and the coherent norms that still guide how people maneuver their way through it.

In this paper we wish to take a different tack than defending or criticizing efforts in machine ethics as a whole. Instead, we will offer a broader view of what demands on machine ethics might look like. Namely, we argue that a more viable direction for design and implementation lies in combining autonomous processing and decision-making with an integrated environment that provides complementary guidance. To help break down where machine ethics can draw lines between robot and its supports, we propose three layers of guidance to which robot behavior can adopt and execute: constraints, cues, and concepts. These can make sense of where autonomous decision-making intersects with more socially guided courses of action through immediate instruction and updated information. In addition, they can allow distinctions between mere participation in a shared environment and actual interaction between people and robots.

3 Constraints, cues, and concepts

The idea of "extended" norm is inspired by the notion of "extended" mind in philosophy, as developed by Clark, Chalmers, and others (Clark and Chalmers 1998). The extended mind view stresses how environmentally supported an agent's cognition is. Cognition does not just emerge alongside technologies like memory aids and documentation, but through ongoing, manifold interactions in a social environment. The brain is not, in other words, an isolated entity wherein thought represents the world to itself—thinking depends on scaffolding from socially buoyed traffic with the world. The range of that environmental scaffolding is constitutive, not just assistive, of the process of

thinking, a view with rich implications for robots (Clark 2001).

This purpose of this paper is not to defend that view of cognition so much as apply it to an agent's operation with norms. Humans, and thus a fortiori AI systems, can act in norm-conforming ways without representing within their cognitive system all of the ethical concepts that represent their possible behavior. Instead, there could be complementary support and scaffolding for normative behavior that a well-designed environment has for the system. In what follows we sketch three layers by which to begin mapping out an extended-norm landscape.

3.1 Constraints

Constraints mark the type of physical barriers and features that limit, without any information gained from a system, the movement or operation on the part of an agent or system. Straitjackets and walls restrict movement without any additional adjustment to the object bound within them. Ramps or sluices direct without discriminating what type of object goes along them (a sorter on an assembly line might be a two-phase constraint, since it could sort by shape before directing some objects in one direction rather than another). While constraints are chiefly framed negatively (what they prevent from being done), their practical function can be more directive (an angled wall that turns a robot right or left). The physical details of a constraint can vary—a laser "virtual wall" for a Roomba serves as a constraint to the Roomba's guidance system, not a solid wall. Still, the category marks out what, given an object's physical features and sensory apparatus, directs and restricts actions with the least amount of deliberation. While previous work in human-computer interaction has noted the importance of contextual constraints for agents (Kandefer and Shapiro 2008), their function for norm-oriented decision-making has yet to be treated as directly as we wish to attempt here.

3.2 Cues

Cues are basic signals that indicate states of affairs or actions that are common objectives in a space. They can have directive force, a velvet rope marking a line for example. What distinguishes them from constraints is that they require some degree of interpretation and comprehension on the part of the agent sensing them in order to work as intended. A yellow caution light does not dictate a behavior necessarily, but it does heighten and focus attention around its position. Auditory signals can be cues around danger or a transition (a ring when there is an entrance or exit to a store, or buzzing for an errant screen action). Their guidance is simple but not always perfectly clear. A plastic yellow stand at the end of a just-mopped aisle may be sufficient to cue shoppers, without needing to state "Wet, proceed with caution." Likewise, a green light blinking at a register may be enough to have a customer step forward for their items to be scanned.

The line between constraints and cues can be difficult to draw if one thinks of how instinctive and direct a reaction to a cue can be. The interpreted nature of a cue can slip into the background. Still, the importance of drawing the distinction lies in mapping where cues, as suggestive signals, do not have the unambiguous direction and force of constraints. Mitigating circumstances, such as how common a car alarm can be, can shape what the usual behavior at the behest of a cue is (a car alarm known by neighbors to be sensitive may, over time, be ignored). So, too, the operation of cues can be less intrusive because of that position in the overall environment. A blinking light will likely not prevent one from leaving the building in a fire, but a wall or barrier might.

3.3 Concepts

Concepts, the third layer, require more interpretation and understanding to grasp what work they can do. Not surprisingly, concepts make guiding appeals with language and other interpreted symbols, gestures, and expressive movements. They comprise representations in their most explicit form. A blinking light might prompt a shopper to come to the register, but the sign "12 items or fewer" depends upon more explicit recognition of number and objecthood. Moreover, control via concepts is open to a wider set of compliant actions because of its interpretive spread. I can bring 13 items and say two of them should only count as one. I can appeal to the cashier to agree that what I have in my cart is manageable enough. I might notice that the customer ahead of me had 14 items and was let through. In each of these cases, I can appeal to the concepts as the keys to the control measure.

Obviously, the guidance of represented concepts in an environment has strong bases in natural language, which means spoken or represented words by other agents in an environment will be expected as possible facets of norms and their adherence. Natural language instructions from interactants may require discernment for appropriate responses (a child asking for an item in a store in order to play will be a different request than an employee asking for an item they dropped to be left alone so that it can be disposed of). Nonetheless, the guidance of concepts will depend on some degree of linguistic competence. The line between cues and concepts need not be sharp, since some basic symbols have a representable content that is shared as common knowledge (the difference between a green arrow on a stoplight and a plain green light being a conceptualized difference involving the need to yield, which is implied but

not represented). But by the category of concepts we mean to demarcate forms of control and guidance that contain explicit, accessible appeals that can be cited and interpreted in representing intentions, plans, and goals to other agents.

Machine ethics, when it has tackled tasks carried out by artificially intelligent systems, has different emphases on concepts as guides in an interactive environment. Deontic logic, which draws inferences about actions based on permissions, obligations, prohibitions and other duty-oriented designations, allows for explicit reasons for why a particular action was decided upon. Strongly deontic work like that of Bringsjord et al. (2006) puts a premium on explicit ethical codes not being violated. Deontic logic, which focuses on the inferential relationships between permissions, obligations, and prohibitions allow deontic logic-based approaches like this to yield explicit, provable reasons that a decision was taken over another. Decision-making on the basis of reasons that are offered to other agents, and that open themselves to correction and challenge by the people hearing them, has drawn interest in human-robot interaction and social robotics, if in rudimentary fashion. Alternatively, in settings of motion planning and human-robot coordination, explicit concepts have largely been left behind in favor of approaches that feature statistical inference of, for example, intended motions from observed trajectories (Dragan et al. 2013). Still, if machine ethics is to register the inferential links between intentions and actions, including counterfactual conditions that show how a norm would be upheld in different circumstances, the conceptual level will continue to pose a challenge to real-world robotic applications.

4 Integration and reorientation

Viewing environments for AI systems as equipped with different layers of control—constraints, cues, and concepts—opens up a different avenue for designing these systems and their roles. Instead of asking how exhaustive a system's ethical principles or training is, one can situate and elaborate a system's control features in tandem with supports. In that respect extended norms more closely resembles behavior-based robotics approaches, which directly utilize environmental features and constraints for generating goal-oriented behavior (instead of performing operations on representations of the environment which might be incomplete or outdated by the time they are needed). In some cases, this will mean developing better means for people to understand and query what a robot does. In other cases, the scale and complexity of the environment may require more stringent constraints and clearer cues to keep the system's capacities from being overwhelmed.

In the case of a supermarket, for instance, the planning and tasks assigned to a robot may call for certain designations to be accountable and identifiable parts of the system's architecture: certain functions of cleaning, or security, or notification of danger may mean explicit representations of what conditions define what to do when. At the same time, during unusual times or in particularly ambiguous spaces (outside the door where solicitors may be asking for a petition to be signed), what people are doing may be more than a system should be asked to navigate. There may be constraints to where a robot goes, just as there are cues and concepts that the robot does not need to have available for its operation and interactions (e.g., processing the license plate number for a double-parked vehicle). In line with work on "overtrust" (Robinette et al. 2016), it would make even more sense for a system not to suggest too much understanding of the social context as a whole. Otherwise, the system's appointed tasks might portray judgments and priorities that are well beyond its capability (e.g., judging whether a parent is acting appropriately when yelling at a child).

Consider how robotic arms may be designed to assist people with an action like feeding. There are broad questions to be answered about how such assistance would be managed by hospitals, clinics, and domestic environments. The technical details of how a robotic system might respond to slight movements and gestures will face added social challenges of how the person being fed wants the arm to move, as opposed to care providers or loved ones (Riek and Robinson 2011). What if the sense of control that a patient needs in order to be emotionally ready to be fed means instructions to the arm that are suboptimal physically (too fast for safe swallowing, say)? The extended norm view recommends looking not just at a comprehensive representation of all these ethical wrinkles in the arm itself, but at the feeding environment and how it is run as well. At what point, providers and users can discuss, does an arm need to heed natural language and at what point should it discourage spoken instructions?

The three layers of extended norms can help diagnose where the vulnerable spots of machine ethics lie. The various levels of control needed to facilitate successful robotic assistance in a social environment do not vitiate machine ethics—it can flesh it out. The role of norms is not bypassed by rejecting machine ethics instead of, for example, safety, but only deferred until intentional opacity and violations crop up as robots move and act among others. Constraints and cues provide a more solid basis for the needed interactivity and transparency that concepts afford, they do not remove their role entirely.

Distinguishing the role of constraints and cues offers opportunities for a more complementary orientation of a system with its environment, instead of it being a free-ranging

agent that needs to represent every specific belief about the environment in order to navigate it (e.g., that walls cannot be penetrated and therefore constrain possible planning pathways). If a system's limited set of representations mesh properly with the supports of constraints and easily interpreted cues, then the demands of representation and explicit planning for the control architecture may be less onerous.

5 Facing challenges

Any task that requires perceptual accuracy, as well as conflicting priorities in shared space, will lead an agent to engage with norms. This does not mean an artificial agent must have a developed ethical theory with which to resolve all possible conflict, but coordinated behavior among agents in that space demands some degree of reliable behavior. It is because social robots enter arenas of coordinated behavior that they will be, consciously or not, subject to judgments enforcing, sanctioning, and evaluating anti-social actions (if only to ask what is agent is intending to do, why they violated a certain norm, and how they would have acted if circumstances were different) (Malle and Scheutz 2014). The intentional fabric of social action is woven with norms.

Recent critiques of machine ethics as a means of controlling social robots (e.g., Van Wynsberghe and Robbins 2019; Vanderelst and Winfield 2018) have stressed how ethical principles might be misapplied or even turn into wrong actions because of delicate social circumstances. Machine ethics is seen by such commentators as too disconnected from larger design approaches involving those who will be using the technology (Van Wynsberghe 2013). The normative emphasis of these kinds of critiques is to pan back from a robot's decision and survey how much is left out by technical crudity. If what is "in" the robot can be manipulated or isolated enough from the wealth of detail that justify judgment calls, what is left is no longer ethics.

Extended norms, however, can integrate some degree of machine ethics and still reinforce the approaches of value-centered (Van Wynsberghe 2020) or value sensitive design (Friedman et al. 2013), which emphasize the involvement of different stakeholders and practitioners as a system is designed (e.g., assistive robots for nurses in care settings). Constraints, cues, and concepts offer different modes of access for those whose work and life will be enmeshed with AI systems, just as they afford more opportunities for insight into how a social setting works best. Feedback about the system's performance cannot just be more accurate articulations of implicit norms and their guidance ("When patients have requested privacy, do not enter [this location]") but different means to bring about the normatively sound practices across a facility or space ("A better privacy cue is needed in this hallway"). By allowing exploration of different control measures, the different kinds of expertise and insight can circulate back into the coding of control architectures. Though architectures would be less burdened computationally by the extended norm scaffolding, they could actually receive more pointed, constructive criticism of what crucial representations they still needed to offer. This criticism, in turn, could generate a more productive assessment of risk, safety, priorities, and policies that are being upheld.

To build further upon this point, an extended-norms approach offers broader means of accountability than is often discussed in work devoted to fairness and equity in AI. The recourse to banning a technology, or emphasizing how data is gathered, still does not make headway on how interactive environment have already come into algorithmic influence. Nor do bans touch on what larger environmental structures would need to change to fulfill norms, not just avoid their violation. While it may seem less fraught to keep robotic action cast as either safe or unsafe, there are intimate measures of respect, dignity, and sensitivity to people's specific needs that do not entirely coincide with concepts like safety or even fairness.

Machine ethics needs to put more stress on transparency and interpretability in real-world interaction than "solving" ethical dilemmas in the abstract. Because people will also be making norm-relative judgments and intentions as they act, the isolated AI system's planning in accordance with a norm needs to be accessible and, given the right circumstances, revisable based on how people will react to that system. One could say that extended norms might reinforce how extended intentions are shared intentions of an organization or group, aided by constraints and cues that lighten the deliberative load of agents across typical circumstances.

5.1 Participation and interaction: lines of explicit and implicit consent

The concerted roles of constraints, cues, and concepts can help mark a more realistic range of robotic decisions and reasoning than a single solution to dramatic moral dilemmas. In turn, it may be important to distinguish "interaction" in a shared environment from "participation." Participation could be distinguished as a less direct confrontation and communication between human and robot agents in space, when agents do not respond to one another while maneuvering and responding to built-in constraints and cues (e.g., a supermarket spill-detection robot automatically moving out of the way of a customer trying to reach an item on the shelf behind it). In this way a designed environment could be judged not on direct dialogue or collaborative planning from human-robot teams, but more so how well agents are accommodated and supported alongside one another.

We have argued elsewhere that the concept of interaction practically implicates robots in issues of consent (Sarathy et al. 2019). Whether a robot in a store accosts a customer, or a delivery robot merely blocks the path of a visitor walking down the same hospital hallway, robots will raise the question of how much a person should have to give consent before a robot's action enters their space. This is especially relevant for contexts of transport, where different functions demand different levels of communication and coordination. While a hospital delivery robot does not have to interact with every person it passes, its actions may affect people in the hallway at some future time. Norm compliance may not be noticed (some people may not ever realize the hallway had a robot in it), because its demands go beyond that of, say, mutual articulation of a norm through dialogue or directed gesture (e.g., greetings).

5.2 Counterfactuals: implicit and explicit

One implication of an extended norms approach is that explanations and justifications from an interactive AI system will have more limited reference to explicit concepts than perhaps initially envisioned. If there are sensory failures that resulted in ignoring cues, or in being damaged instead of guided by constraints, then an explanation of action on the basis of the concept level will not tell the whole story. There will, accordingly, need to be explanatory practices on the part of those working with such systems in context that locate implicit failures from the narrower explanations available to the agent. In other words, a sensory difficulty could be traced back to a time when a system reports no registering of what should have been a perceptible cue. The problem is not the norm represented by the system but its sensory apparatus and possibly the placement of the cue itself.

These implied counterfactuals ("The system would have stopped or proceeded correctly had it noticed the cue" or "The robot would not have blocked traffic if its radar had picked up the virtual wall") should in no way replace all explicit counterfactuals. A socially interactive robot at times should promptly and aptly represent its basis for action (along the lines of "To travel from A to B faster would have meant traveling at an unsafe speed" or "A spill in the aisle would have been made worse by traveling down the aisle"). Data-based systems trained in simulations, which have no explicit, discrete reasons to cite for why one action was taken rather than another, need to meet more specific causal demands on explaining robot behavior's in context (beyond "the system learned through training data").

5.3 The "spirit" of the norm

Norms function through a mix of automatism, interpretation, and exemplary cases. Their enforcement by compliant agents often appeals to this complexity. Revisions of the norm often come from their relation to competing norms that may emerge, in specific circumstances or in general stretches, as being more critical to uphold. As briefly hinted at above, there are situations in a social setting where the usual and acceptable adherence to a norm is out of place. A certain *spirit of a norm*, while not explicit as a set of exceptions or qualifications, are part of a tacit understanding of how far a norm is to go before violating more important ones (e.g., a robot in a bank blocking the cashier's window should move out of the way for a customer but not a bank robber, that would be against the spirit of the politeness norm to make room for people). The case of the sign "12 items or fewer," for instance, would involve certain edge-cases and common sense judgments around what adhering to that guideline is meant to accomplish. It would be odd and unsettling for someone to sanction with thirteen apples, say, or laboriously count out one's items once a cashier has said that it's fine to proceed.

One might cite this spirit by way of showing how rules are too brittle for an AI system to employ in real-world social contexts, how they are bound to take things too far in rigid compliance. This is a reason, in fact, that appeals to training through demonstration have some compelling force. Why not just observe what people do in line and, from the built-up data points from demonstration, learn some of the implicit cues (a wave from a cashier) that make the system's performance fairly familiar?

The problem is that the spirit of a norm, and the navigation of special circumstances, is one where explicit representations are crucial for mediating norm-compliance and adaptive behavior. "You go ahead, I'm still deciding" is a way not just to cue someone to go in front in a line, but a justification that other people can understand and use to decide whether they also could go forward. If the person looks decided, it is probably no longer appropriate.

For this and many related reasons, the idea of extended norms in a context seems like a constructive way to handle the intricacies of the spirit of a norm. There may be constraints that relieve the system of needing to pick up on certain of those nuances, yet there may be critical concepts of safety and acknowledgement that a robot has to represent as it maneuvers through space. The difficulty of norms does not mean circumventing them through imponderably many observations. Instead, it means remaining accountable along communicative and physical lines as people find their own way alongside a system. And, again, extended norms may prevent an exaggerated picture on the part of the interactive system about what it does and does not understand.

People should be given better sense of where a system's explanations will begin and end.

6 Conclusion

Some hyped AI applications have started to reveal insufficiently considered depths to social contexts, like the failures of self-driving cars leading to more consideration of "smart" roads (Toh et al. 2020). This points toward the need to consider more integrated, hybrid approaches to decision-making, planning, and justifications. We have explored how, instead of abandoning ethics as something represented with a system's architecture, the intricacy and inescapable presence of norms suggests a better path of design: complementing internal representations with an extension of norm-informed constraints, cues, and concepts. These will not only flesh out what kinds of norms are being upheld and revised in a social contexts, but they will also offer inroads to insight, experience, and expertise from those affected by the AI system's successes and failures. By separating these layers of a norm-shaped context, one can offer more adaptable, yet no less accountable, means for designing and implementing robots. This also invites better descriptions of what kind of human-robot interaction are anticipated, including multi-agent group interactions, as opposed to mere shared participation in a designed setting. In what contexts do explicit and direct interactions demand some recognition of concepts, and when might cues or constraints alone be sufficient? Ultimately, approaching control through varying degrees of representations will make for a more grounded, inclusive mode of design and deliberation for robotics in particular, moving AI ethics out of "ethics-washing" and into the difficult situations it has always promised to help find our way to resolve.

References

Arnold, T., & Scheutz, M. (2018). The ?big red button? is too late: an alternative model for the ethical evaluation of ai systems. *Ethics and Information Technology*, *20*(1), 59–69.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, *21*(4), 38–44.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language*, *16*(2), 121–145.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7–19.

Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 301–308). IEEE.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: opening up the laboratory* (pp. 55–95). Springer.

Kandefer, M., & Shapiro, S. C. (2008). A categorization of contextual constraints. In *AAAI Fall Symposium: Biologically Inspired Cognitive Architectures* (pp. 88–93).

Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: an overview of reviews. *Perspectives on Psychological Science*, *15*(1), 62–80.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology* (p. 8). IEEE Press.

Malle, B. F., & Scheutz, M. (2020). Moral competence in social robots. In *Machine ethics and robot ethics* (pp. 225–230). Routledge.

Meyer, S., Mandl, S., Gesmann-Nuissl, D., & Strobel, A. (2022). Responsibility in hybrid societies: concepts and terms. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00184-2.

Riek, L. D., & Robinson, P. (2011). Challenges and opportunities in building socially intelligent machines [social sciences]. *IEEE Signal Processing Magazine*, *28*(3), 146–149.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 101–108). IEEE.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, *36*(4), 105–114.

Sarathy, V., Arnold, T., & Scheutz, M. (2019). When exceptions are the norm: exploring the role of consent in hri. *ACM Transactions on Human-Robot Interaction (THRI)*, *8*(3), 1–21.

Toh, C. K., Sanguesa, J. A., Cano, J. C., & Martinez, F. J. (2020). Advances in smart roads for future smart cities. *Proceedings of the Royal Society A*, *476*(2233), 20190439.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)*, *53*(6), 1–38.

Van Wynsberghe, A. (2013). Designing robots for care: care centered value-sensitive design. *Science and engineering ethics*, *19*(2), 407–433.

Van Wynsberghe, A. (2020). Designing robots for care: care centered value-sensitive design. In *Machine ethics and robot ethics* (pp. 185–211). Routledge.

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and engineering ethics*, *25*(3), 719–735.

Vanderelst, D., & Winfield, A. (2018). The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 317–322).

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 775–780). IEEE.

Wallach, W., & Allen, C. (2008). *Moral machines: teaching robots right from wrong*. Oxford University Press.

Thomas Arnold is a research associate at the Tufts University Human-Robot Interaction Laboratory.

Matthias Scheutz is a professor of cognitive and computer science, director of the Human-Robot Interaction Laboratory and director of the human-robot interaction degree programs at Tufts University.