# Understanding the Spirit of a Norm: Challenges for Norm-Learning Agents

Social and moral norms are a fabric for holding human societies together and helping them to function. As such they will also become a means of evaluating the performance of future human-machine systems. While machine ethics has offered various approaches to endowing machines with normative competence, from the more logic-based to the more data-based, none of the proposal so far have consider the challenge of capturing the "spirit of a norm" which often eludes rigid interpretation and complicates doing the right thing. We present some paradigmatic scenarios across contexts to illustrate why the spirit of a norm can be critical to make explicit and why it exposes the inadequacies of mere data-driven "value alignment" techniques such as reinforcement learning *RL* for interactive, real-time human-robot interaction. Instead, we argue that norm learning, in particular, learning to capture the spirit of a norm, requires combining common-sense inference-based and data-driving approaches.

*Key words:* social and moral norms, spirit of a norm, value alignment

## Introduction

Social interactions are hard to imagine without some form of norms. While norms can describe patterns of behavior, they also can possess moral facets of blame, praise, and common values. What Bicchieri calls a "grammar of social interactions", norms are both a form of behavior and a set of judgments and expectations of how things are supposed to be done. While not all norms are moral, even violating social norms can elicit consternation. This applies all the more to autonomous systems or robots that interact with people. If such machines are seen as somewhat intelligent, they will both navigate common patterns of behavior and face expectations to follow some kinds of norms Malle et al. (2019). They could even be expected to sanction norm violations, so as to support the coordination of work and movement in a given social space. These could range from a train station or a factory floor to an elder care facility or a small apartment. Various kinds of social roles and authority could obtain, including who counts as user and on what terms it is permissible to exert control over its operation.

Norms are still more complex than has so far been articulated in much of machine ethics Wallach et al. (2020); Moor (2006). This is not just due to the contradictions among set rules, a well-worn point since Asimov. It is that norms have implicit limits that competent agents grasp. Stopping and waiting at a red light does not mean waiting when the traffic light is broken and never switches to green, nor does it mean refusing to move when motioned to do so by an officer directing traffic. Offering one's seat to the elderly on public transport does not mean to just offer and remain seated. "Staying behind a line" for a subway car does not mean standing between that line and the track.

Norms also carry tacit assumptions about the priorities, purposes, and context-dependent conditions of other norms with which they might conflict. Two adults telling each other crude jokes would be expected to stop upon walking into an elementary school—but if it were summer and they were doing maintenance on that building, perhaps not. When a grocery store's electricity goes out or fire alarm goes off, continuing to wait in line is not compliant but oblivious.

*AI Magazine.* 00(000): 1–16; doi:10.1002/aaai.0001

This might seem like just another kind of common sense challenge for computational approaches, whether explicit and rule-based or implicit and statistically-based (cp. to Moor (2009)). As has been noted in the machine ethics literature, the "frame problem" suggests that AI systems will run into formidable challenges, even with massive amounts of training, about what consequences are sufficient to consider to make an ethically informed decision Briggs (2012).

But in the case of norms the problem goes beyond when to follow a rule or conform to a pattern of behavior. Norms and contexts can shape each other through reasons for acting. Children not being exposed to certain language is part of a norm against swearing, so when they are not in their usual place (i.e., not in a school building) the norm of how to talk around children recedes. And it would be part of responding to accusations to explain why a norm no longer applied in a context, just as it would why a norm has changed a context where it is usually absent (a bartender has had to bring their child into a rowdy bar for a brief time).

Robots, as they move, execute tasks, and communicate, will not always be expected to understand and adjust to these subtleties. Still, even with basic movement, they will meet practices and expectations of passing someone, letting someone through, or alerting a person if one is moving behind them. How would they know not just these basic expectations but also their relation to other norms (e.g. those triggered by an emergency), as well as reasons a norm might not be followed?

In this paper we face these questions through what one could call the "spirit of norms"—a way to capture these often unstated criteria for following a norm amid changing circumstances. Our aim is to explain what this means for machine ethics, why it is such a crucial aspect of socially interactive behavior on the part of robots, and what design challenges it presents going forward. Just as norms occupy a middle ground of social regulation, we argue, so too does the design of robustly interactive systems: they can neither ignore norms in their operation nor exhaustively learn all possible applications of norms. Because the training of purely data-driven systems cannot capture the conceptual basis for the spirit of a norm and related justifications, we suggest how hybrid approaches can contribute to inclusive, accountable design. The chief challenge of there being a "spirit" to norms, we hold, is that of learning a norm adaptively, responsibly, and transparently. This will not just be a complex technical challenge but an imperative for AI systems that would enhance a social context.

## Background and motivation

Embodied systems move into norms as they move into the physical presence of people. The popular robot videos from Boston Dynamics elicit delight, fear, and excitement in part because they have us reflect on what counts as dancing, or running, or violence (when the robot is knocked down). Appropriately enough, given that one kind of such robot is called a "dog," one can be absorbed by how animal-like, rather than anthropomorphic the system is (just as one can remember how robotic "dogs" are employed on the battlefield) Carpenter (2016). Likewise, as robots execute movements in a task environment, they can elicit norm judgments (often positive in terms of efficiency and precision) from their behavior. But when these systems act in concert with others in shared space, not to mention communicate in real-time, the stakes are raised conceptually and ethically. Machine ethics has tried to give philosophical shape to what internal guidance a system could have, whether top-down moral rules or bottom-up via training data from modeled ideal behavior Wallach et al. (2020). Often this depends on whether behavior alone is being sought or, more communicatively, how much a system ought to be able to represent what it is doing and why. The broad goal of "value alignment" will not be sufficient if people need to know why something was done and what would have been done if circumstances had been different Kasenberg et al. (2018). Whether one bases "moral competence" of an autonomous system on norms, utilities, rules, or virtues, the particular application domain for a system will shape those demands for explanation and justification.

One danger identified in previous efforts in machine ethics is that accountability might be deflected from human designers, especially if a system is promoted as an authenticated *moral* agent via the algorithms and training behind it. Still, lack of attention to social

norms in a system's decision-making does not mean the system will be exempt from blame for its actions. Attributions of agency and patiency, Bryson (2018) as human-robot interaction research reminds us, are difficult to corral, instinctual as so many are.

If systems are ever to uphold norms transparently, accountably, and appropriately, It is critical to address the practical tensions that norms carry within them. The spirit of a norm, we maintain, can exceed a straightforward formal specification of conditional action ("do $x$ when $y$ obtains"), since it can carry implicit judgments toward novel circumstances (where $y$ has not yet obtained in this particular way). The norm-sensitivity of a system, then, involves more than an adherence to conditionals within an expected range for its interactants. It offers accessibility for more subtle and delicate norm guidance on the fly.

The larger social science literature on norms has stressed the varied directions that research could pursue, from how norms are internalized as individuals to their social role in providing information and instituting external obligations Legros and Cislaghi (2020). To recall Bicchieri's definition, norms are a "grammar of social interactions," that constitute expectations of what other agents will do, with the added judgment and sanctioning practice of what they *should* do Bicchieri (2005). In this paper we use this basic definition as a guide, recognizing that social norms can differ in function and degree from "moral" norms if those are taken as "unconditional imperatives" Bicchieri et al. (2018).

That said, in the case of human-robot interaction (HRI), among other fields, the line between social norms that define what is acceptable and moral norms that mark off what is worth censuring or blame is difficult to draw sharply without losing empirical accuracy. For the person who needs to trust in a system's performance, the violation of a social norm can point to various moral implications Kuipers (2018). The requirements for norm competence have recently been sketched by Malle et al. Malle et al. (2019), for lack of a computational model that directly focuses on their representation. They point out that norms have a "prevalence" component that registers the external regularity of people following a normative pattern of action and a "demand" component that represents the way norms shape decisions and incur enforcement by a community Malle et al. (2021). Social norms have both a prevalence and demand component when it comes to robotic action in shared space, though the evaluative terms associated with social norms may be less charged than norms typically designated as moral norms.

Machine ethics has tackled norms in direct and indirect fashion, including machine learning techniques to identify norms as upshots from training Fernandes et al. (2020); Shen et al. (2022); Nahian et al. (2020). Though there has been increasing emphasis on verifiability and explicitness around principles Umbrello and Yampolskiy (2022); Kim et al. (2021), it remains a practical challenge for an AI system represents a norm as a patent justification or hypothesis for action. A prospective, contestable use of norms makes it difficult to train a machine on a "norm bank" Choi (2021) in order to claim an isolable inference is being made from a norm. Work in social robotics and planning has rightly identified the difficulty of representing explicit norms in the reasoning of a system Carlucci et al. (2015).

Coggins and Steinert have recently challenged the very idea of norm-compliant robots as a design objective Coggins and Steinert (2023). Norms can lock in outmoded or oppressive societal practices, as well as marginalize certain groups who lack the social leverage to change them, so robots who remain bound by current norms may functionally resist needed social change. While this critique deserves a more thorough response, for the purposes of this paper we would note that practices of change themselves can and in some cases must rely upon norms. The context of design, use, and implementation of certain norms over others will of course be paramount for pursuing more just forms of norm-compliance, but interactive systems will be hard pressed to ignore them altogether.

With these considerations in mind, this paper seeks to explore what reasoning about norms and applying a given norm to concrete actions would mean. We put forward the "spirit of a norm" as a mediating category between the prevalence and demand features of norms: an integrative component that seeks to preserve and encourage communal commitment to what it is about a norm that is most important to uphold.

The spirit of a norm is not only a logistical challenge to formalizing rules but also a socially pragmatic challenge to the

reliance on data and training (simulated or in the wild). It refers both to what stabilizes norm-compliance and what keeps it responsive changing relationships, facts, and needs in a social context.

The recent travails of COVID regulations and compliance have underscored how the tension between rigidity and resilience can unfold: people can use different rationalizations both before and after acting to justify deviance from a regulation or norm Harris (2020). Robots operating in a hospital and tasked with keeping such regulations or related protocols might eventually face verbal interaction involving such reasons—how will they accountably decide to respond?

In previous work we have argued that exceptions to certain ways of fulfilling norms can actually constitute part of those norms, and indeed that competence with norms often means knowing common exceptions (a waiter clearing a plate from the table seeing that a diner is trying to eat more off it) Sarathy et al. (2019). While in some cases there might be repairs or adjustments to navigate a norm-informed interaction ("I'm sorry, I thought you were done with your meal"), the "spirit" of a norm represents an implicit set of presuppositions that norms entail in uncertain or unexpected circumstances. These may lead to more than slight adjustments, but in fact demand more drastic changes. While recognized in brief form in earlier work Arnold and Scheutz (2022), the idea has not been elaborated enough to suggest a distinct research objective for machine ethics. In the following section we explore three hypothetical cases where the spirit of a norm, as a feature of the norm itself, comes to the fore.

## Three scenarios

We here introduce three specific interactive scenarios to show the persistent challenge of norms as both explicit and adaptive to novelty. The scenarios illustrate why norms are so difficult to disentangle from social interaction writ large, as well as why operating in shared space is interactive along many dimensions. They also present an ongoing challenge to design norm competence that is reliable, yet restricted, adapting to the spirit of a norm while not inviting illusory projections of awareness and sensitivity. Managing relational expectations while upholding important norms in common spaces suggests that neither deduction from top-down rules nor bottom-up induction over data will alone be sufficient as methods.

### *Grocery store*

Recently grocery stores have been introducing robots in their aisles as a promotional foray into customer-robot dynamics. While it is not yet clear what the standard tasks or functions of a grocery store robot are, it provides an apt illustration of where various norms would come into play as a robot roamed the store and mingled with shoppers. Due to the implementation of "Marty" at some grocery stores during the pandemic, there have already been questions about social distancing and this robot's possible interference in the aisles Turmelle (2020). There are many other set rules, naturally, that could be stretched or broken in face of other priorities—escaping fire might justify not picking up or even paying for an item one takes from a shelf.

We can identify at least three main norm-related conflicts to manage. First, there might be competing norms, where the question is which one to prioritize. Should a robot cease its cleaning operation during a fire alarm or medical emergency, perhaps to maintain safety by not obstructing a walkway? Second, there are also boundaries to where a norm applies—for instance, the area beyond a cash register (where people have already paid) and the space in front of it. The robot might not clean beyond a certain area, even when called to do so. But there is a third complication that one might call implicit purposes for norms, that regulate to what extent and in what form a norm is upheld. These may be surrounding standards by which norm's limits are determined, for example fairness, equality, and dignity. For instance, if a usual norm is suspended (allowing people to cut a line) in one case, it will be seen by other customers who might infer that they also should get the benefits of such norm loosening.

### *Public crosswalk*

While a store may operate with broadly understood roles like those of customer, cashier, and attendant, other public spaces

pose a more open-ended landscape of social positions and obligations. What it means for a robot to operate in a street became all too clear in a recent case from Pittsburgh, where a delivery robot blocked a person using a wheelchair from making it across a crosswalk Martines (2019). The robot remained in the middle of a curb cut, a pause for a light on the verge of turning to stop. The function of delivering food was being carried out, it was just that other people in that public space were trying to move and live according to all their various purposes (some of them, perhaps, delivering goods as well). The norm of not crossing until clearly permitted by a light, or keeping a reasonable distance from other people, had an implicit "spirit" attached to it—not interfering with a person who has no other path but through one's own location. There is a myriad of even starker cases of norm conflict and norm resolution to imagine for just this setting.

*Hospital hallway*

With the onset of COVID and the demands for sterile clinical spaces and effective forms of social distancing, robots have emerged even more strongly in the minds of those envisioning how medical facilities can operate without the typical contact and proximity of human caregivers. The TUG robot Bloss (2011) delivers supplies while maneuvering through hospital hallways, and companion robots have long been tested as interactive assurance for worried patients (especially children) ScienceDaily (2021). While social distancing might mean fewer people with whom an autonomous system would need to interact, a robot's material sterility offers a chance to channel, if not provide independently, a social presence. The deathbed farewells over *FaceTime* have now become, tragically, a familiar task for clinicians to provide, but one can easily, if not happily, imagine other forms of technologically-generated interaction for those in care. It was presumably the traumatic and confusing times of being alone while suffering that led one woman to cry out to Alexa for help hours before she passed away.

Setting aside charged interaction in clinical spaces, even the negotiation of a hallway by a delivery robot meets with normative expectations. Being addressed for information by visitors or patients, having to move without blocking or scaring people, moving at a speed that others can adjust to and rely upon: these are norms that are largely unstated but, when violated, show some of their purpose. The spirit of a greeting norm—reciprocating an address, especially if there is a question directed to the addressee—is not to overturn norms of movement and space-taking: stopping the middle of a busy hallway is not a good fulfillment of a greeting norm. Being a non-threatening presence around children does not mean playing a game with them while others are making their way to an appointment.

## The Middle Ground of Norms

Designing robotic systems to fulfill norms involves at least a thorny predicament, if not always an outright dilemma. A thorough understanding of when and where a norm applies will not be possible on the basis of prior examples alone, if that means only the data of sanctioning or non-sanctioning behavior used for reinforcement learning (RL) approaches. The conjunction of one type of behavior with blame or praise (or similar positive or negative feedback) need not impart any reason as to why a norm was enforced in one case and not another. If an unattended plate of cookies with the sign "Take One" reflected a strict norm, any number of interpretations could be still technically correct. It could mean that other colleagues can take one but the person inviting them will take back the dish and whatever remains on it, technically not taking more than one cookie but ending up with many of them. These kind of exploits are part of why RL systems can be successful in finding ways of maximizing utility in a given environment (especially ones, as with video games, where an entire space can be explored and exploited) Shao et al. (2019). On the rule side, simply adding qualifying riders to the norm's specification may still not capture what we call its spirit. There are features of the environment that do not depend on fuller description to determine a norm's extent, but on a different apprehension of surrounding circumstances or conditions. A norm of distancing (especially apt in these pandemic times) might be suspended in a case of flood or other immediate and pressing threats to physical safety. Similarly, norms that apply in roughly marked contexts (inside or outside a store, or between store

and street) are not always improved by making ever more elaborate spatial identifications (defining the threshold between in and out by inches rather than leaving that transitional area vague might be even less functional, not more so).

Some commentators on machine ethics have taken these ambiguities as part of why any morally explicit forms of internal robotic guidance are misguided van Wynsberghe and Robbins (2019). Not only could robots handle these situations badly, but they might also be deceptive in presenting "moral" reasoning they in fact lack. claimed. The very means by which a robot could make an ethically informed decision might be the opening a bad actor could exploit to turn it toward an evil or perverse incentive Vanderelst and Winfield (2018). Likewise, norms might change more rapidly than what is encoded and implemented Coggins and Steinert (2023). If norm-based reasoning is so vulnerable to having the wrong norms encoded or employed, why try?

The chief weakness of these critiques is that error, deception, and evil intent can as easily leverage a norm-free system as it can a norm-guided system. In the case of algorithmic systems designed with little or no consideration of possible norm violations, lacking norms is all the more convenient for bypassing values-based design altogether. Indeed, it risks complicity with recklessly applied opaque systems to say that explicit norm formulation in decision-making is impossible to get right anyway. Moreover, once social spaces are entered interactively, norms are at work regardless of how explicitly the systems can treat them. The question in all cases is how the norms are best upheld. Closing off avenues of norm recognition may endanger them overall more than iterative refinements to formulate them better. The concern of overselling something as "moral" only underscores, as many efforts in AI "explainability" have, where explicit judgments can be located and criticized, lest opacity be accepted as a necessity for any AI system's judgments. It is to related questions of social reception that we now turn.

## Letter vs. Spirit Interpretation

The function of the spirit of a norm can be seen when the "letter" of a norm or request is followed instead. In a fiery discussion, when a moderator reminds a frequent interruptor to "please stop talking, we want to hear this person finish," the letter interpretation might be to stop talking for the rest of the discussion. When the interruptor is invited to comment and just shakes his head with his lips tightly closed, the reaction will not be that he takes "stop talking" seriously, but that he is churlish. Following the letter, but not the spirit of a norm will thus not result so much in sanctioning typical for violations (whether reprimand, fine, or imprisonment) but a broader range of implicit or explicit disapproval: rolling one's eyes, shaking of a head, quite possibly a loss of trust in in the performer. The spirit of a norm is a social challenge to learn, as observers and enforcers have many reactions and cues to signal the norm is being lost as it being applied.

In line with this facet of norms, kids seem to become more focused on the spirit rather than the letter interpretation of a principle as they grow older Bregant et al. (2019). This indicates that as their cognitive system evolves, they are able to understand and focus on the intention behind a principle, rather than its literal interpretation. This in some respects resembles efforts for robots to understand indirect requests, which are not always direct questions Briggs et al. (2017).

## Norms and Context in Mutual Formation

As previously discussed, norms are not just prevalent behavior, but a demand on behavior to come. Thus, norm competence extends beyond than available actions within presently existing conditions. It includes counterfactual variations that justify or explain why the actual performance was decided upon, what possible actions would have been taken in different circumstances. Our three scenarios show how context can govern how far those norms apply. Outside of a public road a mobile robot will not be bound by a walk sign, likewise for a robot moving between a sterilized space to a public area of a hospital. At the same time, the current context should not be thought of as defining norms, without any formative influence from norms themselves. Upon reflection, we realize ordinary norms can sometimes change how a context is to be identified. A person who needs medical assistance in a library may override

the usual expectation of silence from the emergency medical technicians (EMTs) who enter. If that person is transported through different areas on the way out (e.g., office, bookstore), the emergency situation shapes them to be more like one another. So recognizing what is done is such a situation (including not bothering the EMTs with other requests) traverses and connects typically distinct contexts into a unified territory where the norm still applies (one might alternatively say these contexts can constitute a broader norm context).

This point is critical for thinking about how norms and contexts are learned, what prescriptions, permissions, and prohibitions do and do not carry over. The spirit of a norm designates what reasons and priorities justify those traversals, and how to identify what those contexts share that invite them. In the EMT example, the suspension of ordinary activity would not apply to a utility room where an electrician was making a critical repair to the building's circuitry. The spirit of a medical emergency norm means that other urgent, safety-related actions in a given space are not expected to adapt the way a bookstore or registrar's office might.

It is worth acknowledging again that norms run the gamut from social regularities that are instrumental for a clear purpose (a line forming here rather than there) to more morally charged demands whose violation incur sanction (helping someone up who has fallen down in the line). At the same time, the distinction itself seems in need of practical mapping and contextual nuance (breaking in line carries different weight depending on what is being waited for).

## Restricted and Reliable: Striking the Right Balance in Norm Competence

How does a system convey its true norm competence? The upside of acting with the "spirit" of a norm is a greater trust and facility among other people with whom one is interacting, including an attribution of being responsible, considerate, etc. In the case of a robot, it may also invite attributions of understanding, patiency, and sensitivity that the robot does not possess. Even observing robot touch can elicit various attributions (often gender-dependent) about its social competence Arnold and Scheutz (2018). How can a responsibly designed robot convey some form of norm understanding without evoking, if unintentionally, a more sensitive and wise grasp of fairness or suffering than it truly possesses? This risk applies all the more in the case of RL agents, where projections of explicit reasoning may misrepresent how they work.

What the spirit of a norm reflects is that true norm fulfillment comes not through sheer rigidity, but rather an acceptable modulation of behavior amid the demands of other norms and relatively novel circumstances. Being adaptive enough to fulfill a norm means the application of a rule may develop contours through previous encounters of analogous cases Forbus (2019). At the same time, fulfilling a norm also entails an accessibility to challenge, accounting, and sanction by a larger community. Agents often will need to offer explicit reasons that can be tested, corrected, and developed in the course of discussing them with others. Thus, being adaptive—even with massive amounts of training data—is no excuse for being opaque.

Presented in these abstract terms, it seems difficult to pin down how AI systems ought to manage these challenges. Some work in explainability has at least tried to resolve some of these tensions, but these largely concentrate on causal, as opposed to normative force Miller (2018). The difference is paramount for human-machine systems, since accounting for a cause that something happened may be sorely lacking as an account or moral justification for what happened. Likewise, projecting a likely outcome of a certain action may not speak at all to whether the predicted course of action threatens to violate a norm (not to mention what alternative courses of action would be preferable as a result).

Norm competence, then, is about upholding with transparency what keeps a norm from being pointless or irrelevant. The accompanying problem is how to represent enough about a norm to make an agent reliable, transparent, and assistive while not suggesting that the system is drawing upon subtle feelings or superior discernment about how norms work. Were that suggestion of sophistication to be generated beyond the system's actual technical operation, a robotic implementation would run greater risk of disappointment, pain, and a loss of dignity. How can systems present restricted norm competence without overstating the extent of that competence? How does

equipping systems with norm recognition prevent normative manipulation of interactants into trusting them?

*Neutralization and Norm Negation*

One vulnerability in norm awareness might, of course, stem from a malign agent gaining access to and changing a system's specifications from the outside Vanderelst and Winfield (2018). Besides the broad challenge of hacking (which applies across various forms of computer systems), we find it worth mentioning one final problem with capturing the spirit of a norm: norm negation and neutralization. As shown in their classic paper, Sykes and Matza lay out five ways that norms can be neutralized in order to justify an agent's delinquency Sykes and Matza (1957). Appeals, some no doubt in bad faith, can be made to why deviation does not injure other parties, or why it fulfills higher norms or priorities. How will a robot engaged in dialogue respond to these appeals on the part of those who might be violating a norm for bad interests? Here again, this might prompt some to scrap the project of norm compliance as too difficult and unwieldy, perhaps with a "dark side" of manipulation possible Vanderelst and Winfield (2018). This, however, still leaves the question of competent interaction in an environment of norms unanswered. What seems more justifiable is a set of prompts and explicit articulation of how far the robot can go in reasoning about norms (and indeed, whether such explicit reasoning takes place at all). The grocery store robots cannot reason about medical protocols, though a doctor tending to a fainting customer might have to do so. There has to be norm integrity and norm fragility, where larger purposes will not be impeded but not just any rationalization will manipulate the robot. How can systems have restricted norm competence and not invite appeals beyond that limit?

## The Norm-Representation Failure of RL-based Algorithms

A common approach to normative behavior from proponents of "implicit ethical agents" Moor (2009) is to not represent norms in an agent in the first place, but to let the agent learn how to behave in a non-conforming manner, e.g., from observing human behavior Russell et al. (2015). The claim is that there are simply too many norms to be hand-coded or engineered into the agent, and that algorithms trained via *Inverse Reinforcement Learning* (IRL) (or its variants) will eventually, given enough good observations, learn an appropriate value function which they can use to learn a policy (e.g., through RL) that is consistent with human norms Milli et al. (2017). Their "values" are aligned with human values—even though "value" here is equivocating because RL-based values (such as $Q$ values, or expected utilities) may not serve as "values" gauged cognitively and affectively in moral reasoning Greene (2014). Instead, if anything is aligned, then it is dispositions: in situation $S$, both agents are (ideally) disposed to perform (or refrain from performing) action $A$. The difficulty, of course, with learning norms from observed behavior alone is that it is not possible to distinguish an action $A$ performed in $C$ because it is obligatory from one performed because it has the highest expected reward. Conversely, $A$ not being performed in $C$ might mean very different things normatively—it may be permissible not to perform it or prohibited. But it is not possible to determine from its absence whether an action is prohibited. If $A$ is prohibited in $C$, then if the learner never observes $A$ in $C$, it could also be because $A$ in $C$ is suboptimal. Alternatively, if $A$ is exhibited in $C$, the learner cannot infer that $A$ is permitted in $C$ because the performer might have violated a prohibition to perform $A$ in $C$ (and was not sanctioned, hence the violation could not be detected). Yet, there is a critical difference between failure to meet obligations and a failure to perform the best action in a given context: the latter will simply fail to get the agent the best possible reward, while the former might break the law.

In defense of IRL, one could argue that as long as the agent conforms to human norms by way of its behavior, it does not matter whether it could discriminate between obligation and optionality, and permission and prohibition. Yet, not being able to do so has serious shortcomings that will ultimately make such norm learners unfit for human societies. For one, it is not clear whether the above method allows the learner to generalize properly to unobserved contexts and actions as often required by the spirit of a

norm. In human daily life, there are simply too many coincidental aspects of human actions (depending on the context in which an action is performed and the performer's goal, role, etc.) for a learner to get a good enough representative sample to make meaningful generalizations from observed behavior alone.

If the agent never observes pushing another person (because pushing is in general not allowed), and a careless distracted pedestrian crossing the street is about to be run over by a rapidly approaching truck, the agent should push them out of harms way (violating one norm to follow another, "inflict some physical harm to prevent worse harm"). An IRL learner having seen pushing before in different contexts, but not in this one, will not push, it has no incentive to do so (for it to push, it would need to be rewarded for the other person not being harmed, but that aspect will typically not be part of its learned reward function; and, in general, it is not clear how broad and wide the scope then would have to be in order to include the effects of large and longer causal chains of performed or omitted actions). While reward shaping on the learned reward structure might be able to help in some cases, it will likely miss important modifications (e.g., if the observed scene and action is being filmed as part of a movie shot).

The spirit of a norm significantly exacerbates the above shortcoming because it allows for a broader range of observed human behaviors that, while not ideal, are strictly speaking still not violations. Putting kids on a leash to obey the prescription to "not leave your kids unattended" would raise eyebrows, even though it does not violate a law. Remaining stopped at a red light at an intersection with no cross traffic. and not moving out of the way when a rapidly approaching car is about to crash into the stopped vehicle, is following the law, but it is not in the spirit of the law (to regulate traffic and avoid harm).

What the spirit of a norm underscores is that immediate and direct blame is much more difficult to model when a norm is atypically and oddly upheld. The person who does not "get" that spirit is frustrating precisely because they are, technically, following a basic rule or norm. The person whose 12 items in the grocery are all giant cases of bottles or cleaning goods has not violated a "12 items or less" direction, but everyone behind them may stew and roll their eyes. They are not immoral, perhaps, but they are exasperating and inconsiderate. What kind of correction or adjustment to norm-following would make sense in such cases, especially in real-time?

In resolving norms practically, what helps immensely are modular reasons and concepts that can guide action. If a manager determines that the surface of a floor has, because of a unprecedented spill, become unsafe, they might need to communicate "Clean another aisle, this one is not safe". The training that a robot like Marty has so far received about its spill detection behavior may have yielded no relationship between prior abnormalities on the shop floor and its cleaning task, and there would be no time for it to try out its mopping on the current spill. Instead, it needs an explicit enough consideration to adjust its upholding of a norm while still holding on to its general force in other cases.

The power of RL/IRL approaches—now commonly, though dubiously, granted sole ownership of the title "value alignment"—are to exploit what is not specified or expressly coded from "above," so to speak. The heel of that particular Achilles will be acquiring the concepts that spell the difference between competence with a norm and asocial (sometimes social, sometimes anti-social, without a dependable means of telling when or why) compliance.

## Norm Learning: A Convening Challenge

The spirit of a norm is what preserves its integrity amid other norms, assumed conditions, and contexts. As our preceding discussion suggests, resolving such a spirit is not as easy as a higher-order rule or an optimizing utility. This is not only a problem for a system remaining explicit in its reasons for action to keep them accountable and revisable. It also convenes an accompanying challenge for a socially interactive system to learn about norms. How can learning both be dynamic enough to be instructed in changing circumstances while genuinely learning a functional, accessible, identifiable norm? Dynamic learning, we propose, is an intersection of inquiry where various AI approaches could propose how to meet the explicit and implicit needs of acting in a normative space.

It is important to distinguish learning norms from the usual reinforcement learning framework of learning optimal policies or

state-action pairs. In the latter case, of course, the task is learning what the best action is given a particular state (factoring in what future rewards it could lead to through subsequent states and actions). Learning a norm, on the other hand, means finding the prescriptive entity or entities that themselves apply to actions in a context. They are explicit means of evaluation that occur amidst action, in some cases through sanctioning.

Once could argue that recent advances in RL with human feedback (RLHF), as it is used for fine-tuning training large language models (LLMs), might be of help here Ouyang et al. (2022). One could imagine using LLMs to generate various norm contexts and ask humans about different actions in those contexts, using human preferences to learn a reward model for the RL algorithm that is subsequently used to update the agent's norm-following policy. But as improvements to the agent's behavior are likely (as they are with LLMs), they will only be incremental and difficult to imagine getting at the "spirit of the norm." Just consider the question of how many such feedback examples contrasting different actions in different norm contexts would be necessary to capture the spirit of normative requests such as "wait for your turn" or "show her some respect:" 100, 1000, more? While it is easy to generate any number of norm context variations automatically, it is unclear what coverage and thus human comparative experiments are needed for different types of norms, especially those fairly broadly applicable ones. In addition to the question of how many examples will be needed, it is unclear whether a reward model trained from human feedback based on comparisons of different actions in a normative contexts makes sense normatively if the actions are equally bad from a normative point of view and thus hard to compare. And if the reward model is trained with a "moral score" there might be too much divergence among labelers to reveal a clear preference. The general problem here is that in cases of norm violations, and especially those with norm conflicts, there is often no right answer but the answer depends on the circumstances and how people interpret them and argue with respect to principles to justify choices and actions. The RLHF setting is simply not designed to deal with the complexity of human normative judgments.

The point thus remains that inferring what actions fulfill a norm, and in particular what fulfills the spirit of a norm, complicates state-action pairing as the only learned regularity. A norm is much more than a state feature (though it can be constituted in part by it). As we will point out below, norms can apply to disparate states without depending on a definite set of common features, hence relying on common state features to generalize from learned states to new states will not work (as would be required for various machine learning techniques, like reward shaping). Norms thus serve as standards that both draw upon usual behavior but also have a conceptual and counterfactual element of how actual performance has fallen short. Tracking or registering a norm, in other words, may require the use of analogy in order to offer a reason why an unexpected state calls for a certain action to conform to a norm Forbus et al. (2020).

### *Features and Multiple Realizability*

How can an entity like a norm be learned in a way that remains explicit, accountable, revisable, and accessible? One way to think about norm learning is through an analogy with functionalism. Putnam famously proposed, as a way of presenting a functionalist account of mind, that certain states could be "multiply realizable" Putnam (1967). That is, different physical states could achieve the same mental state (e.g., pain in various species of animal). Accordingly, just as functional states of mind would be realized in different physical systems (leaving aside for now the debate of how different they could be), so norms could be seen as a constant function across different instantiations. A store need not sell the same products, have the same layout, or the same checkout procedure to enforce a norm of paying for an item before one exits. But what would be the constant data structure or rule set that a system could apply and infer actions from across different stores? Further, what would be the spirit of that norm that could accommodate nuances and variations that norm followers implicitly understand (e.g., being asked by a clerk to pay outside the store at a checkout counter they have set up there)? Again, such a spirit reflects how norm competence means knowing certain basic limitations, edge cases, and larger frames of reference for how a norm is sustained and defended.

The idea of norms as multiply realizable among different contexts suggests how norm

learning will require finer details as to what delineates a norm than reading off a set of prescriptions and prohibitions attached to specified states. There might not be a distinct prescription and/or prohibition that applies to every context where the norm functions, just as the range of states where the norm applies might not all share a common aspect. The spirit of a norm, as a function and context is being learned, could be seen as a cluster concept, where a concentration of overlaps marks discernible norms. The various instances of the norm being followed may overlap in various features without one or more features being a constant through all of them.

A system could cite a number of these sets of shared aspects as a way of projecting a norm into a new context. The place to purchase an item could roughly amass cases of exchange at some forms of barriers (a counter, a desk, a cash register, perhaps a self-checkout scanner), without asserting that all the cases of sales must have a barrier.

Norm learning would mean coming to recognize, even in novel circumstances, when a norm applies. While generating an explicit deontic norm representation from data may not be possible (for reasons we have discussed), there may also be features learned through RL-based means that elaborate how a norm works. Alongside other norms, its "spirit" could have unanticipated but instructive regularities gleaned from observations. Returning to our analogies, there may be revealing patterns in moving in hospital walkways that people do not show in stores that an pattern-based RL-learner might pick up.

We call this a convening challenge because how tradeoffs between stable concepts and updated data require more concerted articulation than has so far been attempted in dealing with norms. New data does not always license the identification of a new norm, but instead might broaden how a norm is realized. At the same time, the representation of conditions a system includes in its inference of norm-compliant action can be subject to revision. If that revision is only through training (perhaps a motion plan that is too precise for verbal directions), then explicit interaction over that revision may be largely pointless. But if natural language in real-time is able to say what is allowed or not, what a violation is or is not, then a system should be able to revise an accessible representation in its architecture.

The spirit of a norm encapsulates how norms work across contexts and represents a type of cluster concept for mapping states onto permissions, prohibitions, and obligations. While a norm might be learned distinctly in one context, what carries over in strength and relative priority with other norms in other contexts is yet another learning challenge.

This seems especially critical to tackle given the different roles of sanctioning across social and/or moral norms, including how artificial agents relate to sanctioning practices Sarathy et al. (2019); Jackson and Williams (2019). There are a host of social attributions that could lead to robots eliciting confusion or incurring blame as agents in shared space, but for norm learning the primary question might be how observed sanctioning contributes to building a properly scoped norm representation? Would a sanction of an action reclassify previous instances of an action that had been observed or attempted, or would that be held for later determination? Ideally a norm learning approach would learn from instruction in generalizable, shared concepts of what feature of an action in a particular context was wrong and which features were acceptable.

### *Affordances, Appropriate Uses, and the Scope of Realization*

The example of learning the affordances of an object is a distinctively instructive analogy to learning norms in a dynamic environment. What the use of an object presents in one context may carry very different practical implications from its use in another. Watching a tennis coach swing a racket near a student may look similar to a violent swing of a weapon, both in terms of speed and relative proximity to another person. A heavy paperweight on a desk, when picked up and raised above a cowering person's head, may have lost connection with keeping notes from slipping off of a desk. On the one hand, one might say that no amount of rules could capture all possible uses of such objects, meaning that rule-based interpretations of an action could be too constricted. On the other hand, the generalization principle works against data-driven machine learning as well, since those approaches have no recourse to

top-down abstraction to curate a set of experienced objects (by weight, shape, texture, and ease of holding) without the interference of irrelevant aspects (color, texture, noise made from internal loose parts). One could make a related analogy about doorstops, which come in a vast array of shapes but possess some rough constants of weight, height, and traction on a floor that pertain to their chief function. These allow for one to see a new object and justify why it could or could not serve as a doorstop. In terms of projecting and applying a concept to fresh circumstances, it seems more promising to have an explicit concept to organize and consolidate a system's perceptions than to hope that reinforcement will distill a concept of doorstops from such perceptions alone.

### *Towards Learning the Spirit of a Norm*

The above discussion points to the challenges with standard "value alignment" approaches that attempt to learn action policies that map states to actions in a way that maximizes the agent's reward based on a reward function it learned from observing human behavior (e.g., through IRL). Instead of this two-phase process—first learning the human's reward function spelled out in terms of some features that might not be sufficient or relevant—we propose a four-step learning process (where learning of the four parts could be intertwined) that has the advantage of being explicit, accountable, revisable, and accessible and allows a learner to engage in dialogues about the principles it learned and how it used them:

1. learn how environmental states $S$ are related to norm contexts $NC$
2. learn explicit norm representations $N$ of norms that apply in $NC$ (in some formal language with clearly defined semantics)
3. learn which consistent subset of all applicable norms $N$ in $NC$ to follow (this subset might be subject to change)
4. based on the chosen consistent subset of applicable norms, learn which actions maximize goal accomplishments

The first learning problem is aimed at learning a *functional concept* from instantiations of the concept: what constitutes a particular norm context. For example, an "asking for directions" (AfD) context might consist of an inquirer intending to know how to get to a particular place, and a responder providing the directions if they know them. This interaction can take place in a great variety of physical and virtual conditions (at the entrance of a mall, a parking lot, hotel lobby, subway station, a sidewalk in a city) but could also be through voice or text on a phone (including social media. It seems very likely that there will not be any particular environmental features that all of these contexts have in common other than the above mentioned abstractions: two agents, one with an intent to know (that manifests itself in different ways), another with the potential to answer the query. Hence, any algorithm that attempts to characterize AfD contexts in terms of physical characteristics of the surrounding space, the types of clothing the agents wear, or their age, etc., will almost certainly fail or only be able to capture parts of the norm concepts. For norm contexts are "cluster concepts" (just like most other human concepts) in that there might be no single defining property that fully characterizes all of them (some might be defined by disjunctions, others by broad concepts plus exceptions, etc.). The learning algorithm must thus learn to abstract "all the way" to the most general relevant features, which could be represented in different ways (e.g., logical form, natural language embedded in a vector space, etc.). This abstraction learning could be accomplished by starting with rich contextual features when experiencing an AfD context for the first time, and with each subsequent different encounter, the algorithm could relax the context description until the most general description has been obtained, including mental states of other agents (such as intentions or belief states); alternatively, the agent might ask explicitly about what constitutes such a context and use a natural language definition to perform "one-shot" context learning, e.g., Scheutz et al. (2017).

The second algorithm then must learn what norms apply in AfD contexts and how to represent them. In the simplest case, norms are obligations to act in particular ways, prohibitions from performing certain actions, or a mixture of both (more complex norms involve temporal aspects that require more complex formal representations, e.g., see Arnold et al. (2021)). For example, if one knows the answer to the directions query and has no good reason to withold it, one ought to provide the

directions. Reasons not to provide them could include: conflicts with the responder's privacy (as when a stranger asks for directions to the house of the direction provider, or their car or their partner's workplace); if the direction provider is in a rush; the provider is not allow to speak aloud (e.g., in a theater); a legal reason not to aid the inquirer (e.g., helping a robber on the run escape from a mall). A simple norm representation might look like this:

$$\begin{aligned} & in(AfD - context) \quad \wedge \\ & ITK(a, directionto(loc)) \quad \wedge \\ & knows(b, directionto(loc)) \quad \rightarrow \quad provide(b, a, directions(loc)) \end{aligned}$$

with a refined expression adding the "default" exception clause "normal(AfD)-context" to the antecedents (with non-normal AfD context including the above described exceptions). But other representations are possible (again, we could use natural language embeddings).

The above obligation to provide directions includes aspects of what information to provide, which has to be learned as well: providing truthful information, indicating if one is not sure about part of the directions, confirming that the directions were understood by inquirer, and the like. Moreover, additional norms might have to be acquired, such as the norm to write down directions for a hearing impaired person, to maybe walk the person to the target location if it is close enough and the person is vision-impaired, a child, or a foreigner not able to understand the directions. Additional modifications might also have to be learned– say, that the direction giver is not obligated to perform any actions that might be in some way risky (getting into the car with the inquirer to drive to the target location, lending the inquirer the phone with the direction shown on Google Maps) or when attempts to communicate directions can be abandoned (for instance, the inquirer does not listen and keeps asking over and over again, showing they are not interested in understanding them).

The third algorithm then must learn how the norms in the AfD trade off with norms that are important from outside contexts, e.g., general norms such as "don't lie," "don't hurt a person," "be polite," "be respectful," etc. For example, the obligation for the responder to be polite or to even provide the information does no longer apply if the inquirer does not show the appropriate demeanor ("hey jerk, where's the next pharmacy around here" does not demand a response). There might be an emergency in which directing someone to an exit might only exacerbate a crowding problem there, so that waiting before giving any instructions could be the best option.

Finally, once all of the above aspects are learned (at least to a sufficient level of proficiency), then within the space of allowable actions the learner could presumably try to fine-tune what optimal behavior (using an optimization method like RL that takes normative constraints into account). Again, however, the incorporation of such a method is not the first priority—rather, it is the acquisition of the right norm constraints.

While we know how to solve the fourth learning problem, the first three are open research challenges that need to be addressed (for some effort on resolving the second one, see Kasenberg and Scheutz (2018)). They need to grapple both with scarcity of data (for understanding the scope of the spirit of a norm) and the wealth of irrelevant information that needs to be abstracted over (in order to get at the core of what the norm attempts to regulate). Most importantly, the learning methods and representations need to allow for the openness of norms, that they apply in cases the agent may not have encountered. Extending to the "open world" is a constitutive aspect of norms in that they cover the unknown, i.e., novel contexts in which they ought to apply. Capturing the spirit of a norm thus is the ultimate goal for agents operating in the open world, because enumerating all possible contexts in which a norm applies is not possible. The difficult and demanding nature of this learning may invite practical assessment of how best to manage and design systems with limited resources, but pursuing open world agency without norm constraints is to untether autonomous action from responsible control.

## Conclusion

That norms carry such subtlety and intricacy that they have a "spirit" could seem reason enough to steer clear of trying to encode them, even if only partially. But even with modest interactive ambitions for an autonomous system, such a spirit for various norms will be hard to avoid. Normative competence, and norm learning, will need to feature varying

levels of complexity while remaining accessible to those in a system's practical reach. The life of norms depends on practices that are up for new formulations, not a letter preserved without ambiguity of interpretation and reimagining. At the same time, those practices hinge on understood references and inferences, not just a behavioral record. How much should robots be designed to try to replicate such judgments, especially around moral (not just social) norms? Should norms be outside these systems' formal vocabulary? We have shown that the burdens of interaction in social space, however slight, make norms an imperative feature for systems to recognize, communicate about, and make transparent efforts to uphold. We thus proposed a four-part learning problem as a possible solution to addressing the challenge of understanding and acquiring the spirit of a norm and encourage the AI community to tackle it, either by providing alternative approaches to solutions or showing what is still missing. Among the many threats that algorithmic systems pose across society, there is already high urgency to this interactive problem. Whether it be with chat bots inflicting psychological harm through ethically deficient expressions or autonomous cars not making accountable, recognizable decisions on the road, the fabric of norms cannot be ignored.

## Acknowledgments

This work was in part funded by AFOSR grant #FA9550-23-1-0425. The authors have no conflicts of interest to report.

## References

T. Arnold and M. Scheutz. Observing robot touch in context: How does touch and attitude affect perceptions of a robot?s social qualities? In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 352–360. IEEE, 2018.

T. Arnold and M. Scheutz. Extended norms: locating accountable decision-making in contexts of human-robot interaction. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, pages 1–8, 2022.

T. Arnold, D. Kasenberg, and M. Scheutz. Explaining in time: Meeting interactive standards of explanation for robotic systems. *ACM Trans. Hum.-Robot Interact.*, 2021.

C. Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

C. Bicchieri, R. Muldoon, and A. Sontuoso. Social Norms. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition, 2018.

R. Bloss. Mobile hospital robots cure numerous logistic needs. *Industrial Robot: An International Journal*, 2011.

J. Bregant, I. Wellbery, and A. Shaw. Crime but not punishment? children are more lenient toward rule-breaking when the ?spirit of the law? is unbroken. *Journal of experimental child psychology*, 178: 266–282, 2019.

G. Briggs. Machine ethics, the frame problem, and theory of mind. In *Proceedings of the AISB/IACAP world congress*, 2012.

G. Briggs, T. Williams, and M. Scheutz. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94, 2017.

J. J. Bryson. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1):15–26, 2018.

F. M. Carlucci, L. Nardi, L. Iocchi, and D. Nardi. Explicit representation of social norms for social robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4191–4196. IEEE, 2015.

J. Carpenter. *Culture and human-robot interaction in militarized spaces: A war story*. Routledge, 2016.

C. Q. Choi. Machines learn good from commonsense norm bank, Nov 2021. URL https://spectrum.ieee.org/ai-ethics-machines-learn-good.

T. N. Coggins and S. Steinert. The seven troubles with norm-compliant robots. *Ethics and Information Technology*, 25(2):29, 2023.

P. M. Fernandes, F. C. Santos, and M. Lopes. Norms for beneficial ai: A computational analysis of the societal value alignment problem. *AI Communications*, 33(3-6):155–171, 2020.

K. D. Forbus. *Qualitative representations: How people reason and learn about the continuous world*. MIT Press, 2019.

K. D. Forbus, E. T. Hinrichs, E. M. Crouse, and J. Blass. Analogies versus rules in cogntiive architecture. *Proceedings of Advances in Cognitive Systems*, 2020.

J. Greene. *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin, 2014.

L. C. Harris. Breaking lockdown during lockdown: a neutralization theory evaluation of misbehavior during the covid 19 pandemic. *Deviant Behavior*, pages 1–15, 2020.

R. B. Jackson and T. Williams. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 401–410. IEEE, 2019.

D. Kasenberg and M. Scheutz. Norm conflict resolution in stochastic domains. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. URL https://hrilab.tufts.edu/publications/kasenbergscheutz18aaai.pdf.

D. Kasenberg, T. Arnold, and M. Scheutz. Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 184–190. ACM, 2018.

T. W. Kim, J. Hooker, and T. Donaldson. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890, 2021.

B. Kuipers. How can we trust a robot? *Communications of the ACM*, 61(3):86–95, 2018.

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

B. F. Malle, P. Bello, and M. Scheutz. Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 21–27, 2019.

B. F. Malle, J. L. Austerweil, V. B. Chi, Y. Kenett, E. D. Beck, S. Thapa, and M. Allaham. Cognitive properties of norm representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

J. Martines. Pitt suspends delivery robots after wheelchair user reports safety hazard, Oct 2019. URL https://triblive.com/local/pittsburgh-allegheny/pitt-suspends-delivery-robots-after-wheelchair-user-reports-safety-hazard/.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell. Should robots be obedient? *arXiv preprint arXiv:1705.09990*, 2017.

J. Moor. Four kinds of ethical robots. *Philosophy Now*, 72:12–14, 2009.

J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4): 18–21, 2006.

M. S. A. Nahian, S. Frazier, M. Riedl, and B. Harrison. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 124–130, 2020.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

H. Putnam. Psychological predicates. *Art, mind, and religion*, 1:37–48, 1967.

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

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

M. Scheutz, E. Krause, B. Oosterveld, T. Frasca, and R. Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th International Conference on Autoomous Agents and Multiagent Systems*, 2017.

ScienceDaily. A visit from a social robot improves hospitalized children's outlook, Oct 2021. URL https://www.sciencedaily.com/releases/2021/10/211009093146.htm.

K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao. A survey of deep reinforcement learning in video games. *arXiv preprint arXiv:1912.10944*, 2019.

T. Shen, X. Geng, and D. Jiang. Social norms-grounded machine ethics in complex narrative situation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1333–1343, 2022.

G. M. Sykes and D. Matza. Techniques of neutralization: A theory of delinquency. *American sociological review*, 22(6):664–670, 1957.

L. Turmelle. Is marty impeding social distancing? stop & shop responds, May 2020. URL https://www.nhregister.com/news/coronavirus/article/Don-t-worry-shoppers-they-wipe-Marty-Stop-15253438.php.

S. Umbrello and R. V. Yampolskiy. Designing ai for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles. *International Journal of Social Robotics*, 14(2):313–322, 2022.

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

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

W. Wallach, C. Allen, and I. Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In *Machine Ethics and Robot Ethics*, pages 249–266. Routledge, 2020.

## AUTHOR BIOGRAPHIES

**Thomas Arnold** is Visiting Scholar of Technology Ethics in the Computer Science Department at Tufts University, and a Research Associate at the Tufts Human-Robot Interaction Laboratory.

**Matthias Scheutz** is the Karol Applied Technology Professor in the School of Engineering at Tufts University. His current research focuses on complex ethical cognitive robots with natural language interaction,

problem-solving, and instruction-based learning capabilities in open worlds.