Requirements for an Artificial Agent with Norm Competence

Bertram F. Malle
Cognitive, Linguistic, and Psychological Sciences
Brown University
Providence, RI 02912, USA
bfmalle@brown.edu

Paul Bello
U.S. Naval Research Laboratory
Washington, D.C. 20375, USA
paul.bello@nrl.navy.mil

Matthias Scheutz
Department of Computer Science
Tufts University
Medford, MA, USA
matthias.scheutz@tufts.edu

ABSTRACT

Human behavior is frequently guided by social and moral norms, and no human community can exist without norms. Robots that enter human societies must therefore behave in norm-conforming ways as well. However, currently there is no solid cognitive or computational model available of how human norms are represented, activated, and learned. We provide a conceptual and psychological analysis of key properties of human norms and identify the demands these properties put on any artificial agent that incorporates norms—demands on the format of norm representations, their structured organization, and their learning algorithms.

KEYWORDS

norms; cognition; learning; artificial agents; robot ethics

ACM Reference format:

Bertram F. Malle, Paul Bello, and Matthias Scheutz. 2019. Requirements for an Artificial Agent with Norm Competence. In *Proceedings of 2nd Artificial Intelligence and Ethics conference (AIES'19).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3314252

1 Introduction

No human community can exist without norms [17, 44], and many past human communities have gone extinct with suboptimal systems of norms [45]. It stands to reason that communities that include both humans and machines as partners will also not succeed without norms. If this is true then we need to understand and formalize what norms are in the human mind—how people represent, learn, activate, update, and deploy norms to guide their behavior—so that we can effectively design artificial agents with appropriate capacities to represent and obey norms.

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the United States Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only
AIES '19, January 27–28, 2019, Honolulu, HI, USA
© 2019 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314252

If an artificial agent is to acquire human norms, its formalisms and algorithms must be informed by the properties of human norms—how humans represent norms, learn them, and use them to guide behavior. We introduce here core properties of human norms and define the demands these properties put on any artificial agent that incorporates norms. These demands range from the format of norm representations to their structured organization, from learning algorithms to communication skills.

In sociology and experimental economics, the importance of norms has long been recognized [31, 37]. These literatures try to explain human cooperation despite the individual's rational self-interest, and norms are an external force that constrains human action. But it is not known how such external forces can operate cognitively and computationally. A person complying with norms must have something in their mind that allows their action to conform to the norm, and that something may be called a *norm representation*. A few empirical studies have examined the automatic activation of such norm representations by situation cues—for example, garbage on the floor triggers the "don't litter" norm [14], or the sight of a library triggers the "be quiet" norm [1]. But no cognitive model has been offered that specifies at least some of the key properties of norm representations. This is what we attempt to do here.

2 Properties of Human Norms

2.1 Working Definition of Norm

We define a norm as follows [9, 11, 27]:

> A norm is an instruction to (not) perform action *A* in context *C*, provided that a sufficient number of individuals in the community (i) indeed follow this instruction and (ii) demand of each other to follow the instruction.

Elaboration. This definition captures both the "external" aspect of norms (they are obeyed and enforced by communities) and the "internal" aspect (that they guide actions). The term *action* covers a broad class—including physical (observable) or mental acts, omissions, as well as acts of bringing about a certain outcome. The separation into

conditions (i) and (ii) follows a long tradition of considering both "descriptive" and "injunctive" elements of norms [9, 14, 16].

We can call condition (i) the *prevalence component* of a norm—that members of a community do in fact follow a norm (with a certain degree of consistency); and we can call condition (ii) the *normative demand component* of a norm—the degree to which community members demand of one another that each follow the norm. Because of this normative demand, violations of norms often lead to sanctions (e.g., criticism, isolation, reform, punishment). However, the existence of sanctions need not be part of the definition of a norm, as some have argued [8]. It is entirely conceivable that for some norm in some communities, no sanctions have been necessary to uphold a norm. That would obviously not make it any less of a norm.

For something to be a *norm* requires that people in a relevant community meet conditions (i) and (ii). For someone to have a *norm representation* requires that the agent knows conditions (i) and (ii); and for someone to show *norm compliance*, the agent both knows conditions (i) and (ii) and tries to follow the instruction *because* of conditions (i) and (ii).

Related concepts. Norms differ from other action guides, such as preferences, goals, and collective habits. The normative force condition marks this difference. A lot of people put milk in their coffee, but they do not demand it of each other, so this action is not a norm but a wide-spread preference. By contrast, getting in line to order coffee is a norm, because that is what people would expect of each other. Norms also differ from values, and the notion of being instructions to act in a particular context marks this difference. Values (such as fairness, freedom, dignity) typically govern a larger class of possible actions/outcomes across a wider range of contexts [39].

2.2 Implied and Suggested Properties

We now discuss six properties of norms, either implied by or further elaborated from the working definition, and develop demands that these properties put on artificial agents' norm representations.

2.2.1. Multiple norm types . A widely recognized property is that norms can be of multiple deontic types: at least prescriptions, prohibitions, and permissions. Any representation of a norm-guided action must signal which of the types governs the particular action.

2.2.2. Context sensitivity. A second critical property of norms is that they are context-specific and must somehow be activated by characteristic features of a given context. In initial research we asked people to state the prescription norms that applied to a variety of everyday scenes (e.g., board room, jogging path). Of the mentioned norms, 95% applied uniquely to one specific context.

2.2.3. Community prevalence. A third property is that genuine norms have sufficient prevalence—that is, most community members comply with a given norm and are aware of that collective compliance. In initial research we found that the top 8 prescription or prohibition norms that people stated for various everyday scenes showed prevalence rates between 94% (most prevalent) and 21% (8th-most prevalent). Beyond the actual agreement rates, people also have a *belief* about the norm's prevalence in the community, which we can call the *prevalence parameter*. Though this belief is not necessarily accurate [33], it will often be based on reasonably representative behavioral data.

2.2.4. Graded normative demand. An infrequently noted property is that (at least) prescriptions and prohibitions come in degrees of community demand [15, 26, 29]. At least in English, terms of prescription can capture low demand ("it is suggested to A") to high demand ("it is required to *A*"), with further gradations in between. Likewise, terms of prohibition can capture low demand ("it is frowned upon to *A*") to high demand ("it is forbidden to *A*"). In preliminary research we found that people show high consensus in ordering these terms along a dimension of normative demand, and Figure 1 shows such orderings for all three norm types. Thus, human norm representations include a graded normative demand parameter.

A norm's degree of normative demand is likely to be closely related to its prevalence, but the two are not reducible to each other. In general, the more strongly people demand of others to conform to norm N_i, the more people will obey it. But when community norms change, strong demand sometimes lingers even though prevalence is declining [24]; and for some norms of only modest normative demand, prevalence may be high (e.g., if the benefits of norm conformity are considerable). Even though the exact relationship between prevalence and normative demand is unknown, as a first approximation we can assume that demand is a linear function of prevalence and other factors such as severity of consequences.

2.2.5 Resolving norm conflict by normative demand. Sometimes norms stand in conflict with one another such that, in the given context, every action violates at least one norm (e.g., in moral dilemmas). Because normative demand comes in degrees, violating some norms will be more costly than violating others, so norm conflict resolution will have to take graded normative demand into account [15, 20].

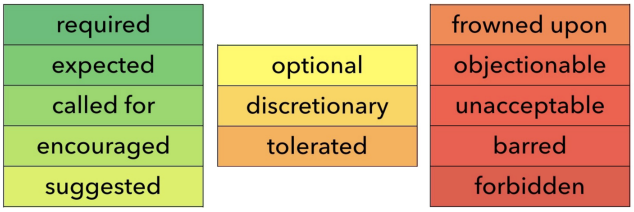

Figure 1: English language terms for graded normative demand in prescriptions (left), prohibitions (right), and permissions (middle).

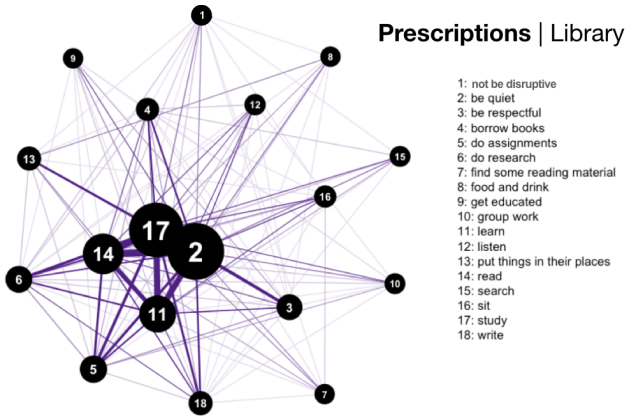

Figure 2. Network representation of prescriptions people generated when exposed to the picture of a library. Larger circles indicate norms mentioned by more people.

2.2.6. Structured representation of norms. The norms relevant to a particular context must be organized in some way. One plausible organization is a network structure that has some core (strong or consensual) norms at the center and other (weaker or less consensual) norms in the periphery. In our preliminary research we found that norms generated by a sample of participants for specific contexts clustered together in such networks (see Figure 2 as an illustration). We also found that, when exposed to a particular context, norms mentioned earlier were more prevalent—thus, they were likely at the core of the network.

3. Formal Representation

Most formal approaches to norm representations have used logical formalisms [12, 32]. One way of formalizing context-specific norms is to introduce a deontic operator $\mathbb{D}$ (instantiating a prescription, prohibition, or permission) for an action A, thus $N_i := C_j \rightarrow \mathbb{D}(A_k)$. C_j may be defined extensionally as a set of preconditions in the world under which A_k is prescribed/prohibited/permitted. It is implausible that these preconditions are linked together as a long conjunction; more plausible is an overall likelihood estimation that aggregates the presence of features into something akin to a sufficient statistic. Given a sample of features from a population, it should be more likely that C_j holds than that any other C holds, even if the specific sample may slightly differ from instance to instance.

However, the material implication $C_j \rightarrow \mathbb{D}(A_k)$ has a number of unattractive properties for deontic reasoning (e.g., [13]). An alternative formulation would be $N_i := \mathbb{D}(A_k, C_j)$, in which the deontic operator establishes a relation between actions and contexts. To capture the parameter of *graded normative demand*, however, the classical interpretation of $\mathbb{D}$ must be expanded to take on a value between 0 and 1, where 0 = prohibited, 0.5 = permitted (optional), and 1.0 = prescribed. A recent proposal further integrated the *prevalence* parameter by defining prevalence as the uncertainty over an estimated deontic value, formally a confidence interval around the normative demand estimate [26].

4. Requirements for Norm-Competent Agents

Summarizing the above properties of norm representations and the suggested formal representation we can now identify a number of requirements for an artificial agent to appropriately represent and learn norms. Some of these constraints have also been discussed in the multi-agent systems literature [25], but typically from a more behavioral and collective perspective. Our contribution is to focus on cognitive properties of norms in individual agents.

4.1 Multiple Norm Types and Action Planning

The three deontic categories (prescriptions, prohibitions, or permissions) must be properly mapped onto the agent's planning and action modules [40]. *Prescriptions* generate goals, with a goal priority value dictated by the prescriptions' normative demand values. *Permissions* affirm current goals that are already set or pursued by the system. *Prohibitions* stop the pursuit of goals, if during this pursuit any actions or consequences fall under prohibited actions or outcomes. In principle, one might wish that during planning *any* considered action path be compared to the agent's norm network to avoid violating prohibitions, but this strategy could quickly lead to computational explosion. The number of norms against which the candidate action is compared must be restricted. Limiting the contexts the agent may find itself in would help. In addition, the agent's planning could be primarily guided by context-specific *prescriptions* because, if the right ones are activated, they are safe to pursue, and their numbers will generally be manageable. If the agent is about to pursue a goal, then it could first be compared against this more limited prescription set for the given context, and if there is no match, a question of clarification may be in order (e.g., "This goal is not among my duties…").

4.2. Context Sensitivity and Context Recognition

The property of context sensitivity requires that the agent recognize what context it is in and activate the context-appropriate norms. It is currently unknown exactly how humans recognize contexts and how context activates the relevant norms. But it is clear that many elements go into context: space (e.g., the room one enters), time (e.g., morning vs. evening), event type (e.g., party, debate), who is present in what role (e.g., friends, authorities), agent's own role (e.g., assistant, guest), and more. In order to activate the right set of norms, the norms "preconditions" must be tested. at least two approaches are available [27]. One, the agent may collect a sufficient number of features from the observed environment and from its knowledge base that make being in a particular context C_j likely, which then activates the norm set that is

relevant for C_j. Second, the agent may activate specific norms relevant to specific features in the environment (e.g., a chair, someone else's phone, someone giving a speech) without necessarily making an overall categorical determination which "context" it is in; instead, it would rely on the world to naturally make features co-occur such that the set of feature-activated norms turns out to be the right set for the particular context.

Whichever model is correct, both context (feature) recognition and norm activation put enormous demands on the agent in terms of perception, attention, and categorization capacities. In general, sophisticated scene understanding is not possible for today's artificial agents. Significant progress has been made in object classification, but recognizing relations that define contexts (e.g., *when this person in this role performs this action in this space*) is currently out of reach. What is challenging about relational information is that many relations that uniquely pick out contexts need to be recognized over different temporal scales, and all sorts of relations could exist (in principle) between the various relata (i.e. objects, events, etc.) in a scene. Considering all possible relations and testing whether they indicate that one is in a certain context is computationally inefficient.

Memory for the objects and events that typically appear in contexts might provide initial hypotheses for which context one might be in, but verifying and disconfirming these hypotheses requires further action. One candidate to supplement memory-based relational processing is *attention* [21], which offers several advantages. First, top-down attentional guidance will narrow the space of relations to be considered, and attention plausibly realizes so-called "visual routines" [43], which can be run to verify whether or not a relation holds [7, 47].

Artificial agents might also do better when constraints are in place on the contexts to be recognized. First, the agent could be deployed in a limited domain that has a limited number of contexts (e.g., a nursing home robot, a cafeteria bussing robot). The agent could then be equipped with a knowledge base of reasonably reliable indicators of the likely (sub)contexts it could find itself. Searching for such tell-tale indicators (rather than scanning all possible scene features) would greatly speed up processing and increase accuracy of context recognition. Second, hypotheses about which context the agent is in and hypotheses about which norms apply could mutually constrain one another. That is, some initially activated norms (e.g., activated by a few salient objects or persons in the scene) might serve to narrow down just what context an agent is in. As norms are highly context specific, the co-activation of even just a small subset of initially activated norms could make it very likely that one is in context C_j, rather than in any other candidate context. This hypothesis of being in C_j would then initiate selective attention to additional features that tend to uniquely characterize C_j, further (dis)confirming this context hypothesis. By extension, the context hypothesis makes predictions about additional applicable norms that can be tested by observing other agents' behaviors or by calculating their joint likelihood with the subset of initially activated norms.

4.3 Prevalence Parameter

An agent's prevalence parameter would reflect an estimate of the proportion of people in the community who actually obey the norm. Mere observation will typically not suffice to achieve such an estimate (see our discussion of the limits of observation below), so the designer will have to provide a starting value that can be updated, through both observing community members and querying them about their norm perceptions and practices (just like humans in foreign countries query locals about their norms).

4.4. Normative Demand Parameter

Including a graded normative demand parameter will require collating evidence form a number of sources in the relevant community: prevalence of norm compliance (because stronger norms tend to be followed more consistently); intensity of demand expressions (e.g., verbal exhortations, warnings); and intensity of sanctions upon violation (e.g., yelling at a person for committing a violation).

4.5. Norm Conflict Resolution Algorithms

An agent facing conflicts among norms will have to rely on the conflicting norms' demand parameters to minimize violation costs [22]. For example, [20] proposed an algorithm within a Markov Decision Process framework where norms are represented in linear temporal logic as temporal expressions that the agent intends to make true. In doing so, agents attempt to obey as many norms as possible, closely monitor the relative normative demand (importance), and minimize aggregated violation costs (whereby greater costs accrue for stronger norms and for temporally more extended states of violation).

4.6 Structured Organization

Building norm representations that have a structured organization (e.g., network structure) will be challenging because little is known about how human norm networks are organized. A plausible hypothesis is that such networks have a core and a periphery, with core nodes being more prevalent and/or more important. Nodes would represent action instructions (with a normative demand parameter), and edges would stand for the probability of co-activation in the same context. It can be expected that human norm networks have "small-world properties" [42]: Nodes in the network are connected to only a small subset of all nodes and have high degrees of clustering, but the clusters can be traversed with relatively short path lengths (i.e., there are many context-specific subnetworks that are nonetheless connected by linking nodes). Though semantic networks have these properties [41], norm networks are likely to differ from semantic networks. For example, norms and their high degree of context specificity would likely show even more clustering and sparseness;

moreover, the demand and prevalence parameters associated with norms differ considerably from the properties of word meanings.

5. Implications for Norm Learning

We now turn to the greatest challenge of designing norm-competent artificial agents: how they could acquire norms. We begin with the case of human norm learning and then develop implications for norm learning in artificial agents.

5.1 Human Norm Learning

Most generally, norm learning refers to the process of extracting $\mathbb{D}(A_k \mid C_i)$ relations from evidence in the world, estimating prevalence and normative demand. What is this evidence in the world that reveals norms? We categorize this evidence into four types and briefly discuss each type.

5.1.1 Explicit instructions. The most direct evidence for a norm is its declaration, in symbols (e.g., signs) or verbal utterances. However, signs are rare (consider how few laws are publicly displayed), and verbal norm instructions may be even rarer [46]. When provided, however, such explicit instructions can both provide information about prevalence (e.g., "Everybody here is making a donation…") and also scale community demand by choosing the graded linguistic expressions of normative demand, as shown in Figure 1.

5.1.2 Behavior patterns. Observing other community members' behavior is a powerful second type of evidence for the presence of a norm [28]. For example, by looking at others in a cafeteria we deduce whether we are expected to bus our own dishes or someone else does it. Already from ages 2 to 3 on, children readily infer norms from other people's behavior [34]. Mere behavioral observation, however, provides only limited information. First, whereas trends of behavior reveal the prevalence of prescriptions, they are sparse with respect to prohibitions, because when people comply with prohibitions there is typically no behavior to be observed. Second, prevalent behaviors performed by a number of people can also be *desired* by those individuals, rather than reflecting a norm. For example, on a warm summer day at the beach many people eat ice-cream, and many people stand in line at the ice-cream stand; only the latter behavior is norm-guided. To differentiate norm-guided from desired behavior additional information is needed. On the behavior side itself, a group of people all performing a certain behavior in highly similar ways increases the likelihood of a norm operating. Patiently waiting one's turn in line is quite orderly and uniform, whereas eating the ice cream afterwards shows more variability. Beyond the behavior itself, critical evidence to distinguish norms from desires lies in consequences of the observed behavior, the third type of evidence.

5.1.3 Behavior Consequences. Relevant consequences include at least two kinds: costs for the agent and benefits for other people. If an agent's foregone alternative behaviors would be individually more attractive, then the observed behavior is *costly* for the agent and suggests the presence of a norm [18]. Everybody would prefer to order ice-cream right when they arrive at the stand, so waiting in line is costly and likely norm-guided. In addition, if an agent's observed behavior causes *benefits* to others, then this provides evidence for a norm, as with tipping, table manners, holding doors open, etc. Conversely, rare behaviors with visible negative impact on others (taking another person's ice-cream instead of purchasing one) suggest a violated norm of prohibition.

5.1.4 Social (dis)approval. Community members' expressions of approval or disapproval of a performed behavior constitute the fourth type of evidence. Disapproval, such as chiding someone who cuts in front of the line, clearly reveals a violated prohibition, both to the person who violated the norm and to an observer. Disapproval often comes in degrees through varying facial, verbal, and bodily signals, so normative demand can be inferred; indeed, a violation's perceived degree of "deviance" (i.e., a proxy for normative demand) is a strong predictor of likelihood of expressed social disapproval [10].

Expressions of approval for a performed behavior may suggest a prescriptive norm that has been met, but such approvals are fairly rare and increase primarily when the behavior exceeds, rather than just meets, the relevant norm. Nobody gets praised for standing in line or treating others with respect, precisely because the norm has made compliance literally "normal." Approval for omissions could indicate prohibitions that were upheld, but such praise is even rarer ("good job for not cheating on the test").

5.2 Machine Norm Learning

Artificial agents, just as humans, should be able to learn from instruction, observation of behavior and consequences, and from social (dis)approval. Learning norms from instruction is challenging for many reasons, not the least is that the preconditions (context) and the action will always be underspecified. Nonetheless, some success has been reported in robots learning recipes (analogous to cooking norms) from written data [30] and learning new action norms from spoken commands [38]. Broadly, programming a robot with a set of a priori norms (e.g., [4]) is a form of teaching by instruction as well, though still challenging because a robot that knows *If C then A* needs to identify the instances in which *C* holds and select the specific form of *A*.

Learning norms from observation may be enabled by Inverse Reinforcement Learning (IRL, e.g., [5]), which has been proposed to ensure that agents "align with human values" [35]. There, the agent observes other agents' behavior as well as the rewards and punishments those agents receive, and it derives a value function that encodes what the proper behaviors are. IRL algorithms can grasp behavior patterns in specific contexts and may be able to code for degrees of norm demand. But without further enrichment, this approach cannot distinguish between actions that benefit the individual agent and actions

that benefit the community [6], or between norms that hold for some people but not for others (e.g., observing students and teachers in the classroom, the agent would infer that the teacher violates norms).

Reinforcement learning (RL) approaches to norms [2, 23] are responsive to rewards and punishment, which could come in the form of social (dis)approval. Thus, if trustworthy community representatives give the agent feedback, it could learn appropriate actions for specific contexts. However, we would not want learning agents to "experiment" in social environments and learn from trial and error what appropriate actions it should take. Such training could occur in virtual and game-like worlds, but creating those worlds may be as difficult as building norms into the agent from the start.

Recent work identified norms from sanctioning behavior in multi-agent simulations [36]. This approach takes advantage of the diagnostic evidence of social (dis)approval, but in natural environments it faces the problem that praise is infrequent (hence learning prescriptions becomes difficult), and it would learn prohibitions only when it commits or observes a violation (which is a costly form of learning). In addition, disapproval is less frequently expressed in societies with high norm compliance—which is exactly where one would want to "raise" an artificial learning agent).

Most generally, none of the foregoing approaches learn norm *representations*; they learn only how to *behave* in accordance with observed human patterns or human sanctions. As a result, they cannot express what they have learned, only that the learned action is the "best" in this context. But the standard of what is best may not be a norm but could be a desire or habit, or even a physically convenient movement (e.g., going upright is not a norm but anatomically convenient for humans). Moreover, such agents cannot represent norm conflicts, because without having representations of norms they cannot diagnose norms as standing in conflict. Yet, representing norm conflicts, and explaining how they should be resolved, will be critical in human-robot interaction, when action recommendations differ or when the human is surprised by the robot's behavior. Trying to address these limitations, some authors have developed algorithms that learn explicit norm representations from observed behavior and are therefore able to recognize norm conflicts and attempt to resolve those conflicts [19, 20].

None of the current algorithms, however, can exploit the rich information contained in (foregone and actual) consequences of observed behaviors. To that end, agents would have to infer an observed person's goals, assess costs and benefits for the person and for other individuals, and compare actual to counterfactual actions (e.g., the agent's foregone benefits as an indicator of prescription norms). Thus, agents would need to have social-cognitive capacities to ground their norm competence.

Given the constraints of learning from observation (e.g., no observed data on prohibitions, no sanctioning data on prescriptions) and the current limitations of algorithms for such learning, observation alone will not generate norm-competent agents. There would be too few data points to grasp the complex context specificity of human norms, too little knowledge about important distinctions in the behavior stream, and no sense of which observed behaviors represent norms, rather than desires or habits.

Currently, and for the foreseeable future, the safest and most effective way of designing norm-competent agents would therefore be a "hybrid" approach [3] in which *a priori* legal, moral, and social norms combine with abilities to learn new norms and update existing norms. Science would have to identify the relevant *a priori* norms (e.g., for a robot in a particular role in a particular community) and implement them in ways that replicate key properties of human norms, such as context specificity, graded normative demand, and network organization. Successful artificial agents would then update this starting package by learning from instruction, observing behaviors, consequences, and social (dis)approval, and requesting advice when necessary. Continuous teaching by instruction and observation is attractive in part because robots deployed in social contexts will be surrounded by teachers. Not all community members are equally good teachers (and as we know form the cases of Tay and Chappie, some will actively try to corrupt the agent). But these problems arise for human children learning norms as much as for robots learning norms. We must trust human communities to make up for the failings of some teachers and find ways to correct individual agents' missteps.

6. Conclusion

We have drawn a map of human norm representations that can guide the development of comparable norm representations in artificial agents. Creating such norm-competent agents is a significant challenge, but it is vital for a society in which human and artificial agents co-exist, both guided by the social and moral norms of their shared community

ACKNOWLEDGMENTS

This project was supported by a grant from the Office of Naval Research (ONR), No. N00014-16-1-2278. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

REFERENCES

[1] Aarts, H. and Dijksterhuis, A. 2003. The silence of the library: Environment, situational norm, and social behavior. *Journal of Personality and Social Psychology*. 84, 1 (Jan. 2003), 18–28. DOI:https://doi.org/10.1037/0022-3514.84.1.18.

[2] Abel, D. et al. 2016. Reinforcement learning as a framework for ethical decision making. *AAAI Workshop: AI, Ethics, and Society, volume WS-16-02 of 13th AAAI Workshops* (2016).

[3] Allen, C. et al. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*. 7, 3 (Sep. 2005), 149–155. DOI:https://doi.org/10.1007/s10676-006-0004-4.

[4] Anderson, M. et al. 2018. A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior

paradigm. *Proceedings of the IEEE*. (2018), 1–15. DOI:https://doi.org/10.1109/JPROC.2018.2840045.

[5] Arai, S. and Suzuki, K. 2014. Encouragement of right social norms by inverse reinforcement learning. *Journal of Information Processing*. 22, 2 (2014), 299–306. DOI:https://doi.org/10.2197/ipsjjip.22.299.

[6] Arnold, T. et al. 2017. Value alignment or misalignment – What will keep systems accountable? *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society*. The AAAI Press. 81–88.

[7] Bello, P. et al. 2018. An attention-driven computational model of human causal reasoning. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (Austin, TX, 2018), 1353–1358.

[8] Bendor, J. and Swistak, P. 2001. The evolution of norms. *American Journal of Sociology*. 106, 6 (2001), 1493–1545.

[9] Bicchieri, C. 2006. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

[10] Brauer, M. and Chaurand, N. 2010. Descriptive norms, prescriptive norms, and social control: An intercultural comparison of people's reactions to uncivil behaviors. *European Journal of Social Psychology*. 40, 3 (Apr. 2010), 490–499. DOI:https://doi.org/10.1002/ejsp.640.

[11] Brennan, G. et al. 2013. *Explaining norms*. Oxford University Press.

[12] Bringsjord, S. et al. 2006. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems, IEEE*. 21, 4 (2006), 38–44.

[13] Chisholm, R.M. 1963. Contrary-to-duty imperatives and deontic logic. *Analysis*. 24, 2 (1963), 33–36. DOI:https://doi.org/10.2307/3327064.

[14] Cialdini, R.B. et al. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*. 58, 6 (1990), 1015–1026. DOI:https://doi.org/10.1037/0022-3514.58.6.1015.

[15] Gasparini, L. et al. 2018. Severity-sensitive norm-governed multi-agent planning. *Autonomous Agents and Multi-Agent Systems*. 32, 1 (Jan. 2018), 26–58. DOI:https://doi.org/10.1007/s10458-017-9372-x.

[16] Gibbs, J.P. 1965. Norms: The problem of definition and classification. *American Journal of Sociology*. 70, 5 (1965), 586–594. DOI:https://doi.org/10.1086/223933.

[17] Hechter, M. and Opp, K.-D. 2001. *Social Norms*. Russell Sage Foundation.

[18] Henrich, J. 2009. The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*. 30, 4 (Jul. 2009), 244–260. DOI:https://doi.org/10.1016/j.evolhumbehav.2009.03.005.

[19] Kasenberg, D. and Scheutz, M. 2017. Interpretable apprenticeship learning with temporal logic specifications. *Proceedings of the 56th IEEE Conference on Decision and Control (CDC 2017)*. IEEE Press. 4914–4921.

[20] Kasenberg, D. and Scheutz, M. 2018. Norm conflict resolution in stochastic domains. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[21] Kim, J. et al. 2018. Not-So-CLEVR: learning same–different relations strains feed-forward neural networks. *Interface Focus*. 8, 4 (2018). DOI:https://doi.org/10.1098/rsfs.2018.0011.

[22] Kollingbaum, M.J. et al. 2008. Managing conflict resolution in norm-regulated environments. *Engineering Societies in the Agents World VIII*. A. Artikis et al., eds. Springer Berlin Heidelberg. 55–71.

[23] Li, J. et al. 2015. Reinforcement learning of normative monitoring intensities. *Proceedings of the International Workshop on Coordination, Organisation, Institutions and Norms in Multi-Agent Systems* (2015).

[24] Mack, A. ed. 2018. Changing social norms. *Social Research: An International Quarterly*. 85, 1 (2018), 1–271.

[25] Mahmoud, M.A. et al. 2014. A review of norms and normative multiagent systems. *The Scientific World Journal*. 2014, (2014), 1–23. DOI:https://doi.org/10.1155/2014/684587.

[26] Malle, B.F. 2018. From binary deontics to deontic continua: The nature of human (and robot) norm systems.

[27] Malle, B.F. et al. 2017. Networks of social and moral norms in human and robot agents. *A World with Robots: International Conference on Robot Ethics: ICRE 2015*. M.I. Aldinhas Ferreira et al., eds. Springer International Publishing. 3–17.

[28] Milgram, S. et al. 1969. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*. 13, 2 (Oct. 1969), 79–82. DOI:https://doi.org/10.1037/h0028070.

[29] Nickles, M. 2007. Towards a logic of graded normativity and norm adherence. *Normative Multi-agent Systems: Dagstuhl Seminar Proceedings* (Dagstuhl, Germany, 2007).

[30] Nyga, D. and Beetz, M. 2012. Everything robots always wanted to know about housework (but were afraid to ask). *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE Press. 243–250.

[31] Parsons, T. 1951. *The social system*. Free Press.

[32] Pereira, L.M. and Saptawijaya, A. 2007. Modelling morality with prospective logic. *Progress in Artificial Intelligence*. J. Neves et al., eds. Springer Berlin Heidelberg. 99–111.

[33] Prentice, D.A. and Miller, D.T. 1996. Pluralistic ignorance and the perpetuation of social norms by unwitting actors. *Advances in Experimental Social Psychology*. M.P. Zanna, ed. Academic Press. 161–209.

[34] Rakoczy, H. et al. 2008. The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*. 44, 3 (May 2008), 875–881. DOI:https://doi.org/10.1037/0012-1649.44.3.875.

[35] Russell, S. et al. 2016. Research priorities for robust and beneficial artificial intelligence. *arXiv preprint arXiv:1602.03506*. (2016).

[36] Savarimuthu, B.T.R. et al. 2013. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law*. 21, 1 (2013), 1–46.

[37] Schelling, T.C. 1960. *The strategy of conflict*. Harvard University Press.

[38] Scheutz, M. et al. 2017. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 1378–1386.

[39] Serramia, M. et al. 2018. Moral values in norm decision making. *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)* (Richland, SC, 2018), 1294–1302.

[40] Shams, Z. et al. 2017. Practical reasoning with norms for autonomous software agents. *Engineering Applications of Artificial Intelligence*. 65, (Oct. 2017), 388–399. DOI:https://doi.org/10.1016/j.engappai.2017.07.021.

[41] Steyvers, M. and Tenenbaum, J.B. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*. 29, 1 (Jan. 2005), 41–78. DOI:https://doi.org/10.1207/s15516709cog2901_3.

[42] Telesford, Q.K. et al. 2011. The ubiquity of small-world networks. *Brain Connectivity*. 1, 5 (Dec. 2011), 367–375. DOI:https://doi.org/10.1089/brain.2011.0038.

[43] Ullman, S. 1984. Visual routines. *Cognition*. 18, 1–3 (Dec. 1984), 97–159.

[44] Ullmann-Margalit, E. 1977. *The emergence of norms*. Clarendon Press.

[45] Wilson, D.S. 2002. *Darwin's cathedral: Evolution, religion, and the nature of society*. University of Chicago Press.

[46] Wright, J.C. and Bartsch, K. 2008. Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly*. 54, 1 (Mar. 2008), 56–85. DOI:https://doi.org/10.1353/mpq.2008.0010.

[47] Yuan, L. et al. 2016. Are categorical spatial relations encoded by shifting visual attention between objects? *PLOS ONE*. 11, 10 (2016), 1–22. DOI:https://doi.org/10.1371/journal.pone.0163141.