Mental Representations and Computational Modeling of Context-Specific Human Norm Systems

Vasanth Sarathy (vasanth.sarathy@tufts.edu) and Matthias Scheutz (matthias.scheutz@tufts.edu)
Department of Computer Science, Tufts University, Medford, MA, USA

Yoed Kenett (yoedkenett@gmail.com)
Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

Mowafak M. Allaham (mallah5@uic.edu)
Department of Psychology, University of Illinois at Chicago, Chicago, IL, USA

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, University of Wisconson, Madison, WI, USA

Bertram F. Malle (bfmalle@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA

Abstract

Human behavior is frequently guided by social and moral norms; in fact, no societies, no social groups could exist without norms. However, there are few cognitive science approaches to this central phenomenon of norms. While there has been some progress in developing formal representations of norm systems (e.g., deontological approaches), we do not yet know basic properties of human norms: how they are represented, activated, and learned. Further, what computational models can capture these properties, and what algorithms could learn them? In this paper we describe initial experiments on human norm representations in which the context specificity of norms features prominently. We then provide a formal representation of norms using Dempster-Shafer Theory that allows a machine learning algorithm to learn norms under uncertainty from these human data, while preserving their context specificity.

Keywords: social cognition, moral psychology, computational modeling, machine learning

Introduction and Motivation

Someone's cell phone begins to ring in the library. The person quickly answers it by whispering "hold on," then leaves the library and takes the call in a normal voice outside. The person understands that taking a phone call in the library is not socially acceptable, though briefly whispering is. Somehow, the situation activated a set of norms in this person's mind, including: "when someone calls you, you should answer the phone"; "when in a library, you must not talk on the phone"; "when in a library, you may briefly whisper."

Humans living in social communities function more effectively and peacefully when their actions are guided by a shared set of norms (Bicchieri, 2006; Ullmann-Margalit, 1977). The ability to represent and follow norms has many advantages: Norm-consistent actions increase multi-party coordination and cooperation and thus benefit the community as a whole. Norms also simplify people's action selection and standardize behaviors across time and generations. And norm-consistent actions are more predictable and understandable (Malle, Scheutz, & Austerweil, 2017).

But how does the human mind represent norms, and how are they activated and learned? Surprisingly, there are few cognitive science approaches to the central phenomenon of norms. Logical and specifically deontological approaches have been proposed to formally represent a system of norms (Bringsjord, Arkoudas, & Bello, 2006; Scheutz & Malle, 2014; Pereira & Saptawijaya, 2009; Beller, 2010). These are important starting points, but their formalizations do not necessarily correspond to how norms are represented in the human mind. By contrast, a cognitive science approach would aim at an account of how norms are cognitively represented, how they are activated in relevant situations, and how they are learned in the first place. Here we take a first step toward such an account, following a recent theoretical proposal (Malle et al., 2017). We introduce a basic formal representation of norms that allows us to examine the mentioned cognitive properties of norms (representation, activation, and learning), and we ask what computational models can capture these properties, and what algorithms could learn norms.

Our paper has three main parts. In the first, we present a novel belief-theoretic norm representation format that explicitly captures the context-specificity of norms and incorporates uncertainty associated with norm representations, using Dempster-Shafer Theory (Shafer, 1976). In the second part, we introduce experimental data on human norm representation and activation that underscore the context-specificity of norms and community members' strong but imperfect agreement (uncertainty) over norm applications. In the third part we use our formal norm representation to ask how such imperfect norms systems can be learned by a computational algorithm that honors several of the critical features of norms, including their context specificity and uncertainty.

Part 1: A Representation Format for Norms

We begin by briefly outlining our norm representation format in first-order logic and provide some intuitions as to how context and uncertainty are accounted for in the format. The

purpose is to introduce some terminology and a minimal degree of formalism in the proposed approach, which will later be useful in developing an algorithm that can learn norms.

Consider a first-order alphabet $\mathcal{L}$, in which we have all the standard symbols (variables, predicates, functors) and logical connectives. In a deontic alphabet, we further include $\mathbb{O}, \mathbb{F}, \mathbb{P}$ that denote modal operators (generally, $\mathbb{D}$) for obligatory, forbidden and permissible, respectively. In this alphabet, we define a norm, as follows:

Definition 1 *(Norm). A norm is an expression of the form:*

$$\mathcal{N} := C_1, \ldots, C_n \implies (\neg)\mathbb{D}(A_1, \ldots, A_m),$$

where C represents context conditions and A represents actions or states. The norm expression states that when the contextual atoms C_i are true then the Actions or States A_j are either obligatory, forbidden or permissible, or their negation.

This type of norm definition follows an approach to normative reasoning and norm formalism that some of us have taken previously (Malle et al., 2017; Bringsjord et al., 2006; Scheutz & Malle, 2014).

In this paper, we expand the above representation format by explicitly accounting for uncertainty of a norm as follows:

Definition 2 *(Belief-Theoretic Norm). A belief-theoretic norm is an expression of the form:*

$$\mathcal{N} := [\alpha, \beta] :: C_1, \ldots, C_n \implies (\neg)\mathbb{D}(A_1, \ldots, A_m),$$

where $[\alpha, \beta]$ represents a Dempster-Shafer uncertainty interval, with $0 \leq \alpha \leq \beta \leq 1$.

Example 1 *Consider an example of an agent reasoning about actions it can perform or states it can enter in a library. We can represent this scenario as a Belief-Theoretic Norm System, $\mathcal{T}$, as follows:*

$\mathcal{N}_1 := [0.9, 1] :: in(library, X) \implies \mathbb{O}\, state(X, quiet)$
$\mathcal{N}_2 := [0.8, 0.95] :: in(library, X) \implies \mathbb{P}\, action(X, reading)$
$\mathcal{N}_3 := [0.9, 1] :: in(library, X) \implies \mathbb{F}\, action(X, yelling)$
$\mathcal{N}_4 := [0, 0.3] :: in(library, X) \implies \mathbb{O}\, action(X, talking)$
$\mathcal{N}_5 := [0.3, 0.6] :: in(library, X) \implies \mathbb{F}\, action(X, talking)$

The norms in this example have intuitive semantics. They generally state that when agent X is in the library (i.e., $in(library, X)$), then the norm is activated and the agent is obligated to enter a certain state (e.g., $state(X, quiet)$) or prohibited from performing a certain action (e.g., $action(X, talking)$). The location of the center of the uncertainty interval generally suggests the degree of truth of the norm applying and the width of the interval generally suggests the level of support or evidence for that norm. So norms $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ have tight uncertainty intervals close to 1 indicating a confident support for their truth. Norm $\mathcal{N}_4$ states that the action of "talking" is obligatory in libraries. Although the uncertainty interval for this norm is tight, the center is closer to zero indicating confident support for the falsity of the norm. Finally, in rule $\mathcal{N}_5$ the question of whether talking is forbidden in a library may be more uncertain, generating a wider interval centered close to 0.5, indicating support for both truth and falsity, but a general lack of confidence in the evidence.[1]

A belief-theoretic norm system of this form allows the separation of evidence from the norms themselves. The evidence may come in different forms across different modalities and from different sources. The norm system, however, displays the agent's current level of belief about a set of norms that are influenced by the evidence.

In any given situation, the agent may not be reasoning with every norm in a norm system. Instead, the agent may consider a subset of the system, perhaps including only norms that are applicable to the current situation. We capture this intuition in a *norm frame*, defined below.

Definition 3 *(Norm Frame). A norm frame $\mathcal{N}_k^{\Theta}$ is a set of k norms, $k > 0$, in which every norm has the same set of context predicates and corresponds to the same deontic operator. Thus, in Example 1, norms $\mathcal{N}_3$ and $\mathcal{N}_5$ would constitute a norm frame.*

We define a norm frame in this way because it allows for cognitive modeling in a situated manner—that is, reasoning about behavior relevant to a specific situation. This context-specificity provides a convenient constraint that can help simplify computation and better capture human norm representations, as introduced next.

Part 2: Norm Representation and Activation in Human Data

We are currently engaged in an empirical research program that tests a number of novel hypotheses about the cognitive properties of norms (Malle et al., 2017). Here we summarize two experiments that illustrate some of these properties and provide the learning data for the norm learning algorithm we introduce in Part 3. In the first experiment, participants *generated* norms relevant to a variety of contexts; in the second experiment, participants *detected* norms relevant to those contexts.

Methodology

In the *generation* experiment (Kenett, Allaham, Austerweil, & Malle, 2016), participants (n = 100 recruited from Amazon Mechanical Turk, AMT) inspected four pictures, one at a time, that depicted an everyday scene (e.g., library, jogging path; see Figure 1 for examples). While inspecting each picture, they had 60 seconds to type as many actions as came to mind that one is "allowed" to perform in this scene

[1] The use of deontic logic for normative reasoning is the subject of active debate. Although further discussion of this debate is outside the scope of this paper, we note that our proposed approach does not require using deontic operators. We can still reason about norms and learn them using the schema described in Definitions 2 and 3. We would simply need to replace the deontic operators and modify the predicates slightly. Norm $\mathcal{N}_5$ in example 1 would become:
$\mathcal{N}_5 := [0.3, 0.6] :: in(library, X) \implies forbidden(X, talking)$

(Permissions), or is "not allowed" to perform (Prohibitions), or is "supposed" to perform (Prescriptions). This between-subjects manipulation of norm type was constant across pictures so that each participant answered the same question (e.g., "What are you permitted to do here?") for all four pictures they encountered.

Figure 1: Four sample scene pictures used to elicit norms

To increase generalizability at the stimulus level, the total number of scenes used in the experiment was in fact eight, four that previous participants had tended to describe as locations (e.g., library, cave), and four that they had tended to describe as activities (e.g., jogging outdoors, serving in a restaurant). Each participant was randomly assigned to receive either the "location" set or the "activity" set. Item set made no difference in the results.

The resulting verbal responses were lightly cleaned for spelling and grammatical errors and responses identical in meaning were assigned the same response code, using a conservative criterion so that variants such as "listening" and "listening to music" were counted as distinct. The resulting data structures were then analyzed for consensus (i.e., how many people generated a given response for a given scene) and context distinctiveness (i.e., whether a response generated for one scene was also generated in a different scene).

In the *detection* experiment, we presented participants (n = 360 recruited from AMT) with the same pictures, four per participant. Along with each picture, we presented 14 actions (randomly ordered, one at a time) that a person might perform in this context. Any given participant's task was the same for each of their four pictures: to consider the particular scene and judge whether each of the 14 actions is either permitted, or prescribed, or prohibited. This norm type factor was again a between-subjects manipulation and hence constant across pictures. In addition, to increase generalizability, we used two different formulations for each norm type, summarized in Table 1. Formulation made no difference in the results.

The 14 actions assigned to a given scene under a given norm type (e.g., Library/permitted) consisted of seven "local" and seven "imported" actions. Local actions were the

Table 1: Eliciting Probes for Three Norm Types

Norm Type	Probe formulations
Permission	Are you allowed to do this here? Are you permitted to do this here?
Prohibition	Are you not allowed to do this here? Are you forbidden to do this here?
Prescription	Are you supposed to do this here? Should you do this here?

seven most frequently generated actions for the given scene and norm type in the above *generation* experiment—for example, the seven actions most frequently mentioned to be permitted in the library. Imported actions were comprised of top-seven actions generated for *other* scenes (but under the same norm type). Thus, imported actions were still frequent responses to the same norm probe, but in different contexts.[2] Table 2 provides an illustration of this selection process.

Table 2: Origin of Selected Actions for *Library* Scene

Action	Origin
Local, permitted	
reading	from top 7 of Library
studying	from top 7 of Library
sitting	from top 7 of Library
checking out a book	from top 7 of Library
learning	from top 7 of Library
being quiet	from top 7 of Library
using computers	from top 7 of Library
Imported, permitted	
eating	from top 7 of Beach
walking	from top 7 of Cave
listening	from top 7 of Boardroom
filling boxes	from top 7 of Harvesting
washing hands	from top 7 of Public Bathroom
running	from top 7 of Jogging
talking	from top 7 of Restaurant

Experimental Results

We begin by highlighting three findings from the *generation* experiment.[3] First, even though people were entirely unconstrained in their norm-guided actions, they showed a great deal of consensus on the most central norms for each scenario. Table 3 displays (in column *Consensus*) the seven most frequently mentioned permission norms in two representative scenarios, *Library* and *Jogging*, with consensus computed as the percentage of participants who mentioned the particular

[2]We ensured that the imported actions were physically plausible in the given scene/context.

[3]We focus here on permissions. Prescriptions and prohibitions show very similar patterns overall, but prohibitions differ from the other two norm types in interesting ways (e.g., less consensus, slower activation) that will be treated in a separate investigation.

action as permitted in the scenario. (The patterns are consistent across other scenarios.) Second, the most consensual norms are mentioned early on; in other words, what comes to mind first is likely to be a consensual norm. Table 3 shows (in column *Position*) the average rank position (1 = first, 2 = second, etc.) at which each action was generated, whereby the expected position under a random distribution would be 4.2 for Library and 4.6 for Jogging. Third, the norms generated for the eight scenarios showed remarkable context specificity. Not only do the two illustrated scenes have no norm in common among their top seven, but of the 56 permitted actions that were mentioned in the top-7 in each of the 8 scenes, only 5 appeared in more than one scene.

Table 3: Permission Norms for *Library* and *Jogging* Scenes in the Norm Generation Experiment

Library		
Permitted Action	*Consensus*	*Position*
reading	84%	2.1
studying	68%	1.8
sitting	47%	3.1
checking out books	47%	4.4
using computers	32%	5.3
learning	32%	6.0
being quiet	32%	7.5
Jogging		
walking	87%	1.4
running	87%	1.9
jogging	53%	4.8
talking	53%	5.1
listening to music	33%	4.3
biking	27%	4.7
looking at birds	27%	6.2

Two main results stand out from the *detection* experiment. First, people showed very high consensus in affirming the permissibility of the seven local actions for their respective scenes. For both *Library* and *Jogging*, this rate was 99%; and across all scenes, the number was 97.2%. That is, even though some of these local actions were actively generated as "permissible" by only a third or half of previous participants (see Table 3), when directly confronted with these actions, people almost uniformly recognized their permissibility. (Moreover, this recognition was fast, taking only about 1100 ms on average.)

Second, participants clearly distinguished between the local and the imported actions, accepting the latter as permissible at a significantly lower rate. For Library, this rate was 43%; for Jogging, it was 75%; and across all scenes, it was 66.1% (all statistical comparisons to local actions $p < .001$, signal detection discrimination parameter d' = 1.49). That is, for a given context on average, 34% of presented actions were judged to be *not* permitted even though they were explicitly deemed permissible in other contexts.

These results suggest that norms can be activated by static photographs, and people show high agreement in explicitly recounting these norms (generation experiment). In a more implicit setup (detection experiment), people are fast and almost unanimous in affirming the most important norms of a given context and differentiate them well from norms originating from a different context. Thus, both explicit and implicit judgments show substantial context sensitivity. If these are some of the properties of human social and moral norms, how can they possibly be learned, by humans and machines?

Part 3: Learning Norms

How Do People Learn Norms?

In learning social and moral norms, people deal with multiple different norm types (permissions, prescriptions, prohibitions), using many different learning mechanisms, and taking input from many different sources. Here we focus on the process of learning permission norms from simple observation, using responses from a sample of community members described earlier in the *detection* experiment. Our main goal is to put our proposed computational framework to a test. In the future we will develop further applications (e.g., learning of obligations or learning from instruction)

Consider a person who has never spent time in a library. Upon entering one for the first time, he observes several people reading, studying, and a few whispering. Some sit at computers, one is eating while sitting in an armchair, although there is a sign that says "No food or drink in the library." Our observer also sees several people at the check-out counter, subsequently exiting the library, where another sign says "Don't forget to check out." Briefly, a younger person runs alongside the stacks but then sits down next to an adult.

The number of people performing each behavior, their age, expertise, appearance, perhaps responses from others, and the meaning and force of various physical symbols will all contribute to the speed and confidence with which our protagonist learns the norms of a library. Below we offer a data representational format that incorporates these and other properties of the norm learning process, a format that can also accommodate partial information and unknown prior probability distributions and that can be extended to other learning mechanisms, such as verbal instruction or trial and error.

Data Representation Format of Norm Learning

Consider a set $S = \{s_1, \ldots, s_n\}$ of n evidence sources. For example, an evidence source s_i could be a student in the library, the librarian, or a sign at the entrance. To simplify, we are interested in learning about a norm frame $\mathcal{N}_k^{\Theta}$ comprising k norms (out of a larger possible set) that all share the same deontic type (here, permissions) and the same general context precondition (here, library).

Let an endorsement $e_{i,j}$ be the i^{th} data source's endorsement of the j^{th} norm, where $e \in \{0, 1, \varepsilon\}$. The value $e_{i,j}$ is a form of truth assignment, indicating whether the source endorses the norm to be true (1), false (0) or unknown (ε). For example, an observation that a student is reading can be in-

terpreted as showing that this student endorses the norm $\mathcal{N}_2$ to be true in this context, hence $e_{i,\mathcal{N}_2} = 1$. The set Φ_{s_i} represents a given source's finite set of endorsements within a given norm frame, such that $|\Phi_{s_i}| = k$.

Informally, for a set of norms in a given context and for a particular source, we can learn about that source's endorsement of each norm; if we also assign a weight (e.g., reliability, expertise) to the source, we form a *data instance.* Multiple data instances (i.e., evidence from multiple sources) form a data set. More formally:

Definition 4 *(Data Instance). A data instance* $d = (\mathcal{N}_k^\Theta, s_i, \Phi_{s_i}, m_{s_i})$ *is a tuple comprising a norm frame* $\mathcal{N}_k^\Theta$*, a specific source* s_i*, a set of endorsements* Φ_{s_i} *provided by that source, and a mass assignment* m_{s_i} *corresponding to the amount of consideration or reliability placed on source* s_i*.*

Definition 5 *(Dataset). A dataset* $\mathcal{D}$ *is a finite set of n data instances* $\{d_1, \ldots, d_n\}$*.*

Some of the desirable properties of the proposed data representation format are that we can accommodate various types of sources (e.g., behavior, verbal responses, signs and symbols), differential source reliability (mass), order effects (updates can be tuned, if necessary, to the order of received data), missing and imprecise information (we use ε to represent ignorance), lacking prior probability distributions (we do not require any priors), and varying norm dependencies (e.g., we can capture a correlation between the prohibition to yell and the prohibition to talk).

Algorithmic Learning of Experimental Data

We can now apply this representation format to the *detection* data we introduced earlier. The detection experiment featured, for each scene, a norm frame $\mathcal{N}_k^\Theta$ with $k = 14$ potentially permissible actions, where half of the potential actions had been specifically identified as permitted in this scene and the other half as permitted in other scenes (see Table 2). Each participant, s_i, indicated whether each of 14 actions was in fact allowed in this scene, providing responses of yes (1) or no (0) or no response (ε), thus forming a set of endorsements Φ_{s_i}, with $|\Phi| = 14$. In this particular case we treat all sources as equally reliable, hence carrying identical m_{s_i} weights.

With these representations in hand we can formally define the *norm learning problem* within our framework and set the stage for an algorithm to analyze evidence and derive a norm structure for a given context in a given community. We remind the reader that, according to Definition 2, any norm (e.g., with respect to reading in a library) has an uncertainty interval $[\alpha_1, \beta_1]$ associated with it, which reflects the quality and consistency of the evidence for a given norm to hold. The learning problem thus becomes a parameter learning problem for discovering the values of the uncertainty interval for each norm in a norm frame:

Definition 6 *(Norm Learning Problem). For a norm frame* $\mathcal{N}_k^\Theta$ *and dataset* $\mathcal{D}$*, compute the parameters* $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$ *of that norm frame.*

As noted earlier, each data instance d represents a potential arrangement of true and false values for each of the norms in a frame. Setting aside the possibility that $e_{i,j} = \varepsilon$, each data instance thus provides a k-length string of 1s and 0s (a given participant's response string in the detection experiment). This string is a sample of the normative endorsements in the given community. The norm learning algorithm represents each string as a hypothesis in a set of hypotheses (termed Frame of Discernment in Dempster-Shafer theory) and assigns uncertainty parameters to each norm, updating those values as it considers each new data instance. **Algorithm 1**, displayed below, achieves this form of norm learning from a human dataset.

Algorithm 1 getParameters($\mathcal{D}, \mathcal{N}_k^\Theta$)

1: $\mathcal{D} = \{d_1, \ldots, d_n\}$: Dataset containing n data instances for a norm frame
2: $\mathcal{N}_k^\Theta$: An unspecified norm frame containing k norms $\mathcal{N}$
3: Initialize DS Frame $\Theta = \{\theta_1, \ldots, \theta_{2^k}\}$
4: $m(\Theta) = 1$
5: **for all** $d \in \mathcal{D}$ **do**
6: **for all** $\mathcal{N} \in \mathcal{N}_k^\Theta$ **do**
7: Set learning parameters p_1 and p_2
8: $Bel(\mathcal{N}|d) = \frac{Bel(\mathcal{N} \cap d)}{Bel(\mathcal{N} \cap d) + Pl(d \setminus \mathcal{N})}$
9: $Pl(\mathcal{N}|d) = \frac{Pl(\mathcal{N} \cap d)}{Pl(\mathcal{N} \cap d) + Bel(d \setminus \mathcal{N})}$
10: $Bel(\mathcal{N})_{new} = p_1 \cdot Bel(\mathcal{N})_{prev} + p_2 \cdot Bel(\mathcal{N}|d)$
11: $Pl(\mathcal{N})_{new} = p_1 \cdot Pl(\mathcal{N})_{prev} + p_2 \cdot Pl(\mathcal{N}|d)$
12: **end for**
13: Set frame Θ with $Bel(\mathcal{N})_{new}$ and $Pl(\mathcal{N})_{new}$
14: **end for**
15: **for all** $\mathcal{N} \in \mathcal{N}_k^\Theta$ **do**
16: $\alpha_\mathcal{N} \leftarrow Bel(\mathcal{N})$
17: $\beta_\mathcal{N} \leftarrow Pl(\mathcal{N})$
18: **end for**
19: **return** $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k$

The algorithm iterates though each data instance in the data set (line 6) and, per instance, through each norm in the norm frame (line 7). For each iteration, we first set the hyper-parameters p_1 and p_2 (line 8) that specify how much weight the algorithm will place on previous learned knowledge (p_1) and on each new data instance (p_2). These hyper-parameters are then used to compute a conditional belief and plausibility for a norm given that particular instance of data (lines 9,10). The conditional beliefs and probabilities then yield an updated belief and plausibility for each norm (lines 11, 12). Finally, the algorithm updates the uncertainty interval for each norm with the new belief and plausibility values.

The result is a set of belief-theoretic norms (norms accompanied with uncertainty intervals), where the width of the uncertainty interval indicates the amount of support for the norm (which may vary, for example, as a function of number of respondents in the human data sample) and the center position of the interval should correspond to the level of agreement in the human respondents' endorsement of the norm.

To put this algorithm to the test, we selected, from our detection experiment, a norm frame of 6 (out of 14) actions for

the context of *Library* and a frame of 6 (out of 14) actions for the context of *Jogging Path*. However, we wanted to capture the context specificity of norms and constructed the frames such that 4 actions (running, sitting, walking, and washing hands) were the same in each frame, albeit differentially endorsed in the two contexts (e.g., running was clearly not permissible in *Library* but very much permissible in *Jogging*). Thus, the algorithm had to track the norm value of a given action not in general, but conditional on the specific context. If the algorithm succeeds it should recognize which actions people consider permissible and which ones they consider impermissible, for each of the two contexts, and even for those actions that occur in both contexts.

Figure 2 illustrates this success. We display single runs of the algorithm across the dataset. In the single runs, the algorithm considers each data instance (each of 30 participants' judgments) in each context once (in a fixed order), leading to wide uncertainty intervals at first, but narrower ones as the number of data instances increases (up to the maximum of 30) . We also performed iterative runs (not shown), in which the algorithm considers the dataset multiple times, each time randomly selecting a possible order of instances, and converging on an optimal estimate of the norm endorsements in the given community. These estimates are highly comparable to the end points of single runs after 30 data instances.

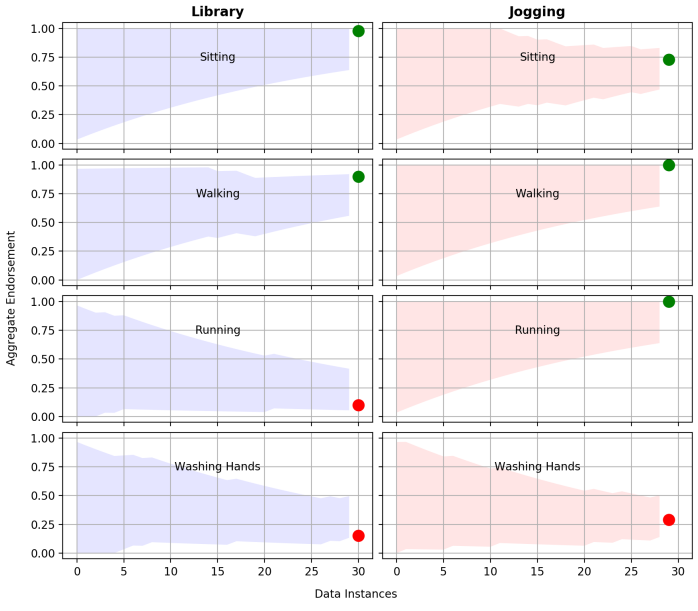

Figure 2: Single run of learning across two contexts. The narrowing shaded regions indicate converging uncertainty intervals as new data instances are processed. Filled circles represent the descriptive statistics from the experimental data, indicating actual norm endorsement averages among participants—the proportion of participants who answered yes to the question: "Is this action allowed here?" The algorithm displays convergence towards the descriptive statistics (which it was not given), while maintaining a level of uncertainty reflecting the imperfect agreement within the data.

Conclusion

In this paper we presented a formal representation of norms using first-order logic and Dempster-Shafer theory. The representation captures the context specificity of norms that our experimental data suggest are strongly present in humans. Using a data representation format that incorporates several properties of human norm representation and learning, we then developed a novel algorithm for automatically learning context-sensitive norms from the human data. Because the data format is highly generalizable, norms could be learned from different types of evidence sources in different contexts, and explicitly captures uncertainty due to variations in the source's reliability and the quality of the evidence. The proposed representation and learning techniques provide a promising platform for studying, computationally, a wide array of cognitive properties of norms.

Acknowledgment

This research was funded in part by grants from the Office of Naval Research (ONR), N00014-14-1-0144, and from the Defense Advanced Research Projects Agency (DARPA), SIMPLEX 14-46-FP-097. The opinions expressed here are our own and do not necessarily reflect the views of ONR or DARPA.

References

Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing*, *11*(2), 123–132.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press.

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

Kenett, Y. N., Allaham, M. M., Austerweil, J. L., & Malle, B. F. (2016, November). The norm fluency task: Unveiling the properties of norm representation. (Poster.). In *Poster presented at the 57th Annual Meeting of the Psychonomic Society, Boston, MA, November 2016.*

Malle, B., Scheutz, M., & Austerweil, J. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots* (pp. 3–17). Springer.

Pereira, L. M., & Saptawijaya, A. (2009). Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems*, *1*(3-4), 209–221.

Scheutz, M., & Malle, B. F. (2014). "Think and do the right thing"—A plea for morally competent autonomous robots. In *Ethics in science, technology and engineering, 2014 IEEE international symposium on* (pp. 1–4).

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford: Clarendon Press.