

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. Kadar, & G. S. Virk (Eds.), *A World with Robots: International Conference on Robot Ethics: ICRE 2015* (pp. 3–18). Cham, Switzerland: Springer International Publishing.

## Chapter 1

# Networks of Social and Moral Norms in Human and Robot Agents

Bertram F. Malle, Matthias Scheutz, and Joseph L. Austerweil

### 1.1 Introduction

The design and construction of intelligent robots has seen steady growth in the past 20 years, and the integration of robots into society is, to many, imminent (Nourbakhsh, 2013; Šabanović, 2010). Ethical questions about such integration have recently gained prominence. For example, academic publications on the topic of robot ethics doubled between 2005 and 2009 and doubled again since then, counting over 200 in early 2016 (Malle, 2015).

One set of ethical questions pertinent to robotics examines how humans should design, deploy, and treat robots (Veruggio et al, 2011); another set of questions examines what moral capacities robots themselves could have (and should have) so as to become viable participants in human society. The latter set of questions is often labeled “machine morality” (Sullins, 2011; Wallach and Allen, 2008) or “machine ethics” (Moor, 2006), and our contribution is to this theme.

Considerations of machine morality are especially important when we assess collaborative relationships between robots and humans. A collaboration can be defined as a set of actions coordinated among two or more agents in pursuit of joint goals. An agent’s pursuit of *joint* goals (rather than merely individual ones) requires several unique capacities, such as social cognition and communication. Even more fundamental, however, collaborations rely on a *norm system* that the partners share—a

---

Bertram F. Malle  
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence RI, USA, e-mail: [bfmalle@brown.edu](mailto:bfmalle@brown.edu)

Matthias Scheutz  
Department of Computer Science, Tufts University, Medford MA, USA, e-mail: [matthias.scheutz@tufts.edu](mailto:matthias.scheutz@tufts.edu)

Joseph L. Austerweil  
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence RI, USA, e-mail: [joseph\\_austerweil@brown.edu](mailto:joseph_austerweil@brown.edu)

system that enables, facilitates, and refines the collaborative interaction (Ullmann-Margalit, 1977).

As a social species, humans have become highly adept at pooling mental and physical resources to achieve goals together that they would never be able to achieve on their own. From big-game hunting to mass migration, from felling a tall tree to playing a symphony—humans work cooperatively to create common goods. But cooperative work comes with risks, because one partner might invest all the work and the other partner might reap all the benefits. Economic scholars have puzzled for a long time why such free-riding is not more common—why people cooperate much more often than they “defect,” as game theorists call it, when defecting would provide the agent with larger utility.

The answer cannot be that humans are “innately” cooperative, because they are perfectly capable of defecting. The answer involves to a significant extent the power of *norms* (Fehr and Fischbacher, 2004). A working definition of a norm is the following (Bicchieri, 2006; Brennan et al, 2013):

*An implicit or explicit instruction to (not) perform a specific or general class of action, whereby a sufficient number of individuals in a community (a) indeed follow this instruction and (b) demand of each other to follow the instruction.*

Why are norms so powerful? First, they increase the predictability of other people’s behavior (Cialdini et al, 1991a). In a norm-guided society, any member can assume that other people will abide by norms, which greatly reduces the uncertainty over what actions they might perform. Second, norms guide a person’s own action selection because norms directly tag possible actions as desirable or undesirable in the given community (Goldstein et al, 2008). Third, norms improve coordination among collaborators. That is because a collaboration involves many requests, agreements, and commitments that bind the individual to a course of action. Public promises, for example, are prototypical commitments to a norm: The declaration “I promise  $X$ ” imposes a norm on oneself to strive toward  $X$ , which involves others’ expectations for the person to strive toward  $X$ , the person’s desire to meet those expectations, and the possible sanctions other people may impose if the person fails to achieve  $X$ .

Norms appear to be indispensable for human social life (Hechter and Opp, 2001; Ullmann-Margalit, 1977). As a result, norms are likely to be indispensable for robots in human societies as well, if we expect people to perceive robots as suitable partners in effective, safe, and trusting collaborations. But what would it mean for a robot to have “norms”—whether moral norms (e.g., “do no harm”) or social norms (e.g., “shake hands when meeting someone”)?

Any robot involved in physical tasks will have to know a number of instrumental rules: *if an object of type  $F$  appears in area<sub>1</sub>, move arm and grab  $F$* . For humans, too, physical tasks require rules—actions that have high utility when certain preconditions hold. By contrast, social and moral norms are rules that are not directly dictated by a personal utility calculation (Andrighetto et al, 2010), and often they are not as action-specific as are instrumental rules (e.g., “Be nice!”). Moreover, social and moral norms have other properties that make them a unique challenge for cog-

nitive and computational examination: There seems to be an enormous number of them but they are activated extremely quickly; they are activated in highly context-specific ways but also come in bundles; they can be in conflict with one another but also can be adjusted; and they are learned fast through a variety of modalities (e.g., observation, inference, instruction).

If our goal is to build trustworthy and morally competent robot collaborators (Malle and Scheutz, 2014), robots must have a computationally implemented norm system. This is because humans will demand that a robot collaborator grasps the norms of its community, and humans will withdraw their trust and cooperation if they realize that the robot does not abide by the same norms as they do.

However, we currently do not know how to incorporate sophisticated norm processing into robotic architectures. We therefore take here initial steps toward a cognitive-computational model of norms by delineating core properties of the human norm system. We then contrast two plausible models of a computational norm system and derive implications for how robotic architectures would implement such a norm system. Ultimately, we will need to examine (1) how a cognitive system can represent and store norms, (2) how and when it activates and retrieves them, (3) how it resolves conflicts among them; (4) how it can use them in decision-making and action execution, and (5) how it can acquire them. In this chapter we begin to address the first two points.

## 1.2 Formal Definition of Norms

To begin, we introduce a general formulation of norms as consisting of three elements: a *context precondition*, a *deontic operator* (“obligatory”, “forbidden”, or “permitted”), and an argument that can be either an *action* or a *state*.

Specifically, let  $C$  be a context expression in a given formal language  $\mathcal{L}$ , and let  $\mathbf{O}$ ,  $\mathbf{F}$ , and  $\mathbf{P}$  denote the modal operators, respectively, for “obligatory”, “forbidden”, and “permissible” (e.g.,  $\mathbf{O}\phi$  means “it is obligatory that  $\phi$ ”). Then we can provide a general schema for capturing simple norms as follows:

$$\mathcal{N} = C \rightarrow (\neg)\{\mathbf{O}, \mathbf{P}, \mathbf{F}\}\{\alpha, \sigma\} \quad (1.1)$$

The deontic operators can be analyzed cognitively as follows. To represent an action or state as *obligatory* [*forbidden*], at least three conditions must be met (Bicchieri, 2006):<sup>1</sup>

---

<sup>1</sup> This is a definition of a norm *representation*, and it thus allows for an agent to endorse an illusory norm—when all three conditions are met but community members do not in fact follow the instruction or do not in fact expect others to follow the instruction. This definition also does not directly incorporate the motivational force that norms undoubtedly have (Brennan et al, 2013), because we are not convinced that this force needs to be contained in the norm representation itself.

- (i) The agent represents an instruction to [not] perform a specific action or general class of action.
- (ii) The agent believes that a sufficient<sup>2</sup> number of individuals in the reference community in fact [do not] follow the instruction.
- (iii) The agent believes that a sufficient number of individuals in the reference community demand of each other to [not] follow the instruction.

Conditions (ii) and (iii) are important. During the learning of a new norm and during continued application of a familiar norm, the agent must be able to update beliefs about what community members do and what they demand of one another. If the agent notices that few community members follow the instruction in question (e.g., staying within highway speed limits), then the instruction is weakened and the agent may no longer treat it as binding. And if the agent notices that few community members demand of others to follow the instruction (whether or not many of them still follow it), the instruction becomes optional and also loses its character as a norm.

These features distinguish *norms* from *goals* and *habits*, because the latter can hold even when individuals completely disregard other community members' actions or expectations. Consider the action of parking one's car nose-in rather than nose-out in parking lots with spots marked like this: // // . If a majority of people perform this action but nobody actually *expects* others to do it, the action is a widely prevalent habit, not governed by a norm. And if a particular agent performs the action but is unaware that others expect him to (and they in fact do), then this agent acts to achieve some goal but does not abide by a social norm.

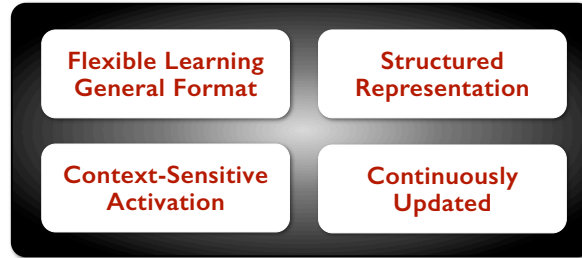
### 1.3 Properties of the Human Norm System

We propose that human norm systems have four major properties (Figure 1.1). We first introduce each of these properties and then significantly expand on the properties of representation and activation.

#### 1.3.1 Property 1: Flexible Learning

The first property is that norm systems are learned through a variety of means (e.g., conditioning, imitation, observation, inference, and verbal instruction) but are stored in a generalized format sketched above (equation 1): as a representation of actions or states (post-conditions), given contextual preconditions. A more detailed treatment of how learning could be implemented computationally requires a better understand-

<sup>2</sup> The threshold of sufficiency will typically be a majority but may vary by norm type and community.



**Fig. 1.1** Four cognitive properties of the human norm system

ing of how norms are represented in the first place, and this is what we will attempt shortly.

### ***1.3.2 Property 2: Structured Representations***

The second property is that norm systems are encoded using structured representations, systematically organized in at least three ways: *vertically* (as hierarchical layers of abstraction, ranging from action rules to general values), *horizontally* (as bundles of covarying norms tied together by the contexts in which they apply), and *temporally* (as “scripts” [Schank and Abelson (1977)] that prescribe normative action sequences in a particular context, such as visiting a restaurant, greeting a friend, or boarding an airplane).

These organizing principles reflect actual features of the world. Because preconditions covary in real-world contexts, activated norms will also covary as bundles within contexts (horizontal organization). Likewise, because the human action planning and execution system is organized hierarchically and temporally, norms that guide such action will incorporate this organization as well.

The structured organization of norms is likely to have far superior processing characteristics over the simplest alternative—(long) lists of singlet norms. That is because norms can be thought of as nodes in a memory network, and we know that structured organization of memory representations has significant advantages in memory accuracy, efficiency, and speed of retrieval (Bower, 1970).

### ***1.3.3 Property 3: Context-Sensitive, Bundled Activation***

As a third property, we suggest that specific contexts rapidly activate norms as connected bundles. There is evidence that norms are indeed activated in highly context-specific ways (Harvey and Enzle, 1981; Aarts and Dijksterhuis, 2003; Cialdini et al, 1991b) and that norm violations are detected very quickly (Van Berkum et al, 2009).

These characteristics are responses to a world in which a large number of norms exists but only a small subset is relevant in any given context. As a result, the norm system must be both comprehensive in its representational capacity and selective in its activation patterns. These demands pose numerous challenges for the computational implementation of a norm network, so we will dedicate much of our subsequent analysis to these challenges (section 1.4).

### ***1.3.4 Property 4: Continuous Updating***

The fourth property of the human norms system is that the context-sensitive norm networks are continuously updated—for example, when a new norm is learned or a new context feature is added as a precondition to a previously learned norm. This makes the norm system highly flexible when people encounter mixed contexts, mixed roles, or enter unfamiliar communities. It also allows for rapid societal change—whether due to natural events (e.g., climate change), technological innovation (e.g., the internet), or collective preferences (e.g., gay marriage).

Cognitively speaking, when a context feature is added as an additional precondition for a given norm, the likelihoods of co-activation (bundling) among norms will change because these likelihoods are a direct function of the number of preconditions shared between norms. How quickly the likelihoods change will depend on general principles of the norm network. For example, updating will be frequent if co-activation of two norms instantly forms a direct connection between them. Likewise, updating will be frequent if equivalence between contexts is loose (i.e., features that define contexts are correlated both within and between contexts, rather than figuring as necessary and sufficient conditions).

We now turn to the central portion of our paper: an analysis of how norms, defined as context-specific instructions, can be activated in bundles tailored to their particular contexts.

## **1.4 Challenges of Context-Sensitive, Bundled Norm Activation**

We have argued that norms are activated in *specific contexts* and as *connected bundles*. How can we account for these characteristics? We first outline the logical format of these bundles and then consider potential computational models of how they are represented and activated.

### 1.4.1 Logical Format

Expressed in the logical format of Equation 1.1, each norm has a set of preconditions  $C$  that correspond to contexts in which the norm applies (e.g.,  $C \rightarrow \mathbf{F}\phi$ ) or in which the norm is specifically suspended ( $C \rightarrow \neg \mathbf{F}\phi$ ). When a given situation  $\Sigma$  meets the contextual preconditions  $C$  of a given norm, the norm will be quickly activated. The critical open question here is what “meeting the contextual preconditions” means.

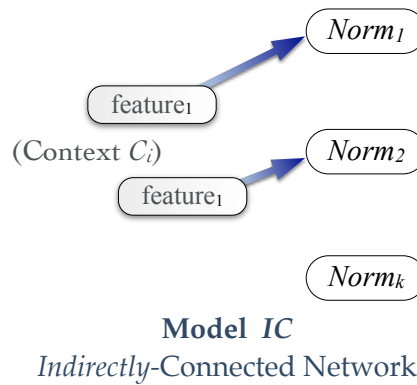
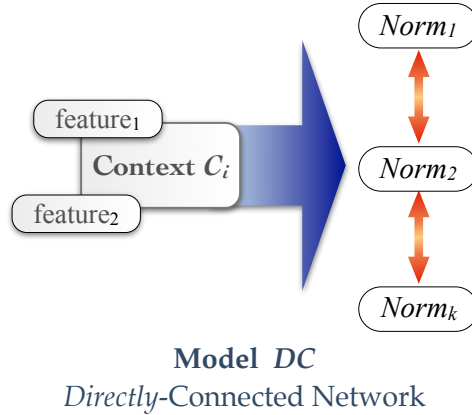
Let  $f_\Sigma$  be the set of features present in a given situation  $\Sigma$  and let  $f_C$  be the features that constitute the preconditions  $C$  for a given norm. We hypothesize that the degree of activation of the norm in a given situation  $\Sigma$  will be a function of the number of features shared between the situation and the preconditions of the norm (e.g.,  $|f_\Sigma \cap f_C|$ , where  $|\cdot|$  is the cardinality of a set, possibly weighted and scaled by factors depending on the contextual features and the norm). If this hypothesis is correct, then *all* norms that have any  $f \in \Sigma$  in their set of preconditions ( $f \in C$ ) will be activated to *some degree*. We will call these co-activated norms in a given situation  $\Sigma$  “norm bundles.”

Note that for two norms  $\mathcal{N}_1$  and  $\mathcal{N}_2$  in a norm bundle it could be the very same contextual property  $f_i \in \Sigma$  that is in both of their norm preconditions ( $f_i \in C_1$  and  $f_i \in C_2$ ). Alternatively, different features in the situation ( $f_1, f_2 \in \Sigma$ ) could activate different norms ( $f_1 \in C_1$  and  $f_2 \in C_2$ , but  $f_2 \notin C_1$  and  $f_1 \notin C_2$ ). Hence, norm bundles do not necessarily have to share any particular situational features, even when their constituent norms are co-activated, as long as there are reliable co-variations of situational features. From a computational perspective the question then arises exactly how these co-variations are represented; that is, whether norms in norm bundles are represented in the human cognitive architecture as connected *directly* with one another or connected only *indirectly*, via the shared preconditions between the norms and the actual context features that trigger them. Hence, there are at least two different models of how such covariation among norms in a norm bundle can come about—models that specify in what way norms are part of a “bundle” (see Figure 1.2).

### 1.4.2 Two Models of Norm Covariation

In a *directly-connected* network (Model *DC* in Figure 1.2), norms and their co-activation are represented as nodes and edges in a mathematical network, where each edge is given a weight indicating the strength of association between the nodes (Harvey and Enzle, 1981), possibly built up through learning and repeated co-activation. A given context (constituted by a fuzzy set of features) activates a particular norm network in part because the context activates some norms and these norms activate other, connected norms.

Alternatively, in an *indirectly-connected* network (Model *IC* in Figure 1.2), specific features (e.g., objects in a scene) independently activate specific norms, and sets of norms covary as bundles solely because the real-world features that activate



**Fig. 1.2** Two models of how contexts can activate “bundles” of norms. Under the first, but not the second, model,  $Norm_k$  would be activated

them typically co-vary within contexts, not because of direct connections among the norms themselves. In this more minimalist network, no additional concept of a “context” (above and beyond an extensional class of features) needs to be postulated. The “affordances” of objects and properties in situations suffice to activate the right kinds of norms.

For example, holding the fork a certain way and holding the knife a certain way while eating at the table may be a connected pair of norms that is activated as a bundle by the sight of a set table; alternatively, the fork may activate *its* norm of use and the knife may activate *its* norm of use, and the two norms are co-activated merely because, in the real world, knives and forks are typically co-present.



### 1.4.3 Different Empirical Predictions

Although both models account for “norm bundling,” the two models make different predictions about norm activation patterns in unusual situations. Consider a situation  $\Sigma$  (e.g., eating at a fine-dining restaurant) that is normally constituted by a sufficient subset from the set of features  $f_1$  to  $f_6$  and, if recognized as a particular context  $C$ , reliably activates the bundle of norms  $\mathcal{N}_1$  to  $\mathcal{N}_4$ , which all have  $C$  as their precondition. Now suppose that the perceptual input is impoverished (e.g., bad lighting or intense noise), making only features  $f_1$  to  $f_3$  available in this particular case. According to the *directly-connected* model, such an impoverished situation would still be likely to activate the whole bundle of norms, because even a few directly activated norms would themselves activate other norms with which they are directly connected. By contrast, according to the *indirectly-connected* model, norms are activated only by specific features (e.g., objects) in a scene, and therefore the impoverished situation would elicit “incomplete” norm bundles—only those that are individually activated by features  $f_1$  to  $f_3$ .

Likewise, the models make different predictions for cases in which a foreign object is present in an otherwise familiar situation (e.g., a baseball in a fine-dining restaurant). According to the *DC* model, a foreign object would have little effect on the activated norms, because once an overall context triggers its bundle of norms, any effect of specific (additional) features would be drowned out (or at least mitigated) by all the interconnected norms. Not so for the *IC* model, according to which norms are activated individually by specific features (e.g., objects) in the scene. The baseball in the restaurant would have a marked effect on the set of activated norms, because people cannot help but bring to mind whatever one may (or may not) do with a baseball, even in a fine-dining restaurant.

### 1.4.4 Implications for Cognitive Robotic Architectures

Implementations of the *DC* model in cognitive robotic architectures could be analogous to networks of spreading activation (e.g., in the spirit of the declarative memory in ACT-R) where a given context (constituted by a sufficient subset of features) will spread activation to the norms that have this context as a precondition. As mentioned, the norms in a given bundle need not have a single precondition that is shared among all of them—as long as some of the norms share some preconditions with other norms and some subset of these partially shared preconditions are present, the bundle will be activated through spreading activation. The main advantage of directly-connected norm bundles is that partial matches or inaccurate perceptions may still be sufficient for activating all norms in a bundle. This is because the direct connections among norms within a bundle will spread activation to each other, so as long as some of the norms are immediately activated (e.g., through perceptions, inferences, etc.), the other ones will eventually become activated as well. The main disadvantage of directly-connected norm bundles is that some norms might become

inappropriately activated (i.e., without there being a contextual feature to which the norm applies), simply because direct linkages can drag one norm along with another.

Implementations of the *IC* model, on the other hand, do not require representational mechanisms such as spreading activation, as all norms in a bundle will be solely activated by the features in the situation (e.g., objects) that match their context preconditions. Hence, the main advantage of indirectly connected norm bundles is that the norms are activated in close correspondence to situations and their recognizable or inferable features. Such a network need not engage in inferences about “contexts” as separate constructs, because contexts are merely extensional classes of features. Of course, if features are highly correlated, such extensional classes could be learned as higher-level categories, but they do not have to be separately represented each time a norm is activated. The main disadvantage of indirectly-connected norm bundles is that acute and fast perceptual processes are required that recognize all relevant objects and properties in the environment so as to activate their corresponding norms (e.g., permissible ways of handling a fork, a knife, a spoon, a napkin,...).

Critically, however, both models require ways to arbitrate among activated norms that have mutually contradictory implications. For example, norm  $\mathcal{N}_1$  might impose an obligation to do *A* while  $\mathcal{N}_2$  might impose an obligation to do *B*, yet either doing *A* and *B* is not possible at the same time, or doing one of them will undo prerequisites of the other in a way that the other action can no longer be performed.

Deciding between the two models will also influence the general logical form of norms. If there are direct connections between, say  $\mathcal{N}_1$  and  $\mathcal{N}_2$  (above and beyond shared preconditions, i.e., context features), how are these connections represented? Are they continuous and/or probabilistic? And what implications does such a representation have for logical reasoning on deontic operators? If, on the other hand, there are no connections among norms themselves, can we completely characterize norm networks as arrays of context features that do or do not activate specific norms? We now explore these possibilities in more detail.

### 1.4.5 What Would Constitute Norm Connections?

Figures 1.3 and 1.4 illustrate how quantitative predictions for norm co-activation strength can be derived from each model. Figure 1.3 shows a hypothetical norm system represented in a table where rows index features  $f_1$ – $f_6$  that constitute contexts  $C_1$  to  $C_4$  and columns index norms that can be activated by these features. A cell is 1 if the corresponding norm is activated in the presence of the feature (and 0 otherwise, but left empty in the table for better readability).

According to the *indirectly*-connected model, the strength of co-activation of norm  $\mathcal{N}_i$  with  $\mathcal{N}_j$ , the formula  $r_f(\mathcal{N}_i, \mathcal{N}_j)$ , can be written as:

$$r_f(\mathcal{N}_i, \mathcal{N}_j) = \frac{\sum_C \sum_f I(f \in \mathcal{N}_i \wedge \mathcal{N}_j)}{\sum_C \sum_f I(f \in \mathcal{N}_i)}, \quad (1.2)$$

Context	feature	$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
<b>C<sub>1</sub></b>	$f_1$		1	1	
	$f_2$		1		
<b>C<sub>2</sub></b>	$f_3$	1		1	1
	$f_4$			1	1
	$f_5$	1			1
<b>C<sub>3</sub></b>	$f_6$	1	1		
<b>C<sub>4</sub></b>	$f_3$	1		1	1
	$f_2$		1		
	$f_6$	1	1		
	$f_5$	1			1

**Fig. 1.3** Contexts ( $C_1$  to  $C_4$ ) and their features ( $f_1$ – $f_6$ ) that activate specific norms  $\mathcal{N}_1$  to  $\mathcal{N}_4$ . Cells with unique colors indicate co-activation of two or more norms by a particular feature.

where  $I(\cdot)$  is the identity function that returns 1 when its argument is true and 0 otherwise. According to equation 1.2, the strength of co-activation of  $\mathcal{N}_i$  with  $\mathcal{N}_j$  is the number of features (repeating features over contexts) they have in common, normalized by the number of features that are preconditions for  $\mathcal{N}_i$  (again repeating features over contexts). For example, focusing on context  $C_2$ , feature  $f_3$  co-activates norms  $\mathcal{N}_1$ ,  $\mathcal{N}_3$ , and  $\mathcal{N}_4$ ; feature  $f_4$  co-activates  $\mathcal{N}_3$  and  $\mathcal{N}_4$ ; and feature  $f_5$  co-activates  $\mathcal{N}_1$  and  $\mathcal{N}_4$ . Features can reappear across contexts, and this is illustrated above by the fact that  $f_3$  and  $f_5$  also help constitute context  $C_4$ . All these co-activation patterns of norms, triggered by features, lead to the feature-level co-activation matrix on the top table of Figure 1.4.

According to the *directly*-connected model, what counts are not feature-level co-activations but context-level co-activations. Contexts, latent factors inferred from slightly varying sets of features, activate their norms *as a set*, with some norms activated by already activated other norms, not by features. Thus, according to the *directly*-connected model, the strength of co-activation between norm  $\mathcal{N}_i$  and  $\mathcal{N}_j$ , the formula  $r_c(\mathcal{N}_i, \mathcal{N}_j)$ , is:

$$r_c(\mathcal{N}_i, \mathcal{N}_j) = \frac{\sum_{\mathcal{C}} I(\exists f \in \mathcal{C} : f \in \mathcal{N}_i \wedge f \in \mathcal{N}_j)}{\sum_{\mathcal{C}} I(\exists f \in \mathcal{C} : f \in \mathcal{N}_i)}. \quad (1.3)$$

According to equation 1.3, strength of co-activation is the ratio of the number of contexts where both  $\mathcal{N}_i$  and  $\mathcal{N}_j$  are applicable to the number of contexts where  $\mathcal{N}_i$  is applicable. For example,  $f_3$  and  $f_4$  would be taken as sufficient evidence for the presence of  $C_2$ , and  $C_2$  would activate, as a set,  $\mathcal{N}_1$ ,  $\mathcal{N}_3$ , and  $\mathcal{N}_4$ . No matter which features in a scene allow a given context to be inferred, all of its norms (the norms

		$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
<b>Feature-level co-activation</b>	$\mathcal{N}_1$ with		2/6	2/6	4/6
	$\mathcal{N}_2$ with	2/5		1/5	0/5
	$\mathcal{N}_3$ with	2/4	1/4		3/4
	$\mathcal{N}_4$ with	4/5	0/5	3/5	

		$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$	$\mathcal{N}_4$
<b>Context-level co-activation</b>	$\mathcal{N}_1$ with		2/3	2/3	2/3
	$\mathcal{N}_2$ with	2/3		2/3	1/3
	$\mathcal{N}_3$ with	2/3	2/3		2/3
	$\mathcal{N}_4$ with	2/2	1/2	2/2	

**Fig. 1.4** Computation of norm co-activation at the level of features (top table) and at the level of contexts (bottom table).

that have that context as a precondition) are activated, and co-activation among these norms leads, over time, to norm interconnections. Those are represented as context-level connection strengths (again normed against number of norm occurrences) in the bottom table of Figure 1.4.

We see that the two matrices are quite different, so they should in principle be empirically distinguishable. Mere feature-caused co-activation predicts far smaller co-occurrence frequencies than context-caused co-activation with subsequent connection formation. If we can measure such norm co-activations (and we are currently developing a paradigm to do so), we have yet another way of arbitrating between the two models, which would teach us about the underlying principles of human norm networks and provide benchmarks for corresponding norm networks in robotic architectures.

We should add that the two models *DC* and *IC* also make different predictions about the process of norm updating (Property 4 mentioned earlier). When a context is added as an additional precondition for a given norm, the *DC* would predict that this norm soon picks up new connections with other norms, because the co-activation (bundling) likelihoods among norms are a direct function of the number of shared preconditions between norms. According to the *IC* model, by contrast, the norm co-activation pattern changes more slowly, and only to the extent that the pattern of feature co-occurrences changes.

Clearly, a number of hybrid models could be constructed as well. For example, one model could allow norm-to-norm interconnections without postulating contexts as latent factors inferred from features. In this case, features directly cause norm

co-activation *and thereby* cause formation of real norm interconnections, so norms could also be activating each other (e.g.,  $f_4 \rightarrow \mathcal{N}_4 \rightarrow \mathcal{N}_3$ ). The problem that arises for a network with these characteristics is that norms could activate other norms that are *not* appropriate for a given context. Consider  $C_1$  in the example norm system of Figure 1.3. If  $f_1 \rightarrow \mathcal{N}_3$  and, because of the strong interconnection  $r_f(\mathcal{N}_3, \mathcal{N}_4)$ , also  $f_1 \rightarrow \mathcal{N}_4$ , then the norm  $\mathcal{N}_4$  is activated in  $C_1$  even though, by assumption for this example network, it shouldn't be active in this context. Thus, the model may have to incorporate inhibitory connections in addition to excitatory connections—which would then lead to interesting new predictions.

A final question is how the human norm network cognitively instantiates an intuitive requirement: that contexts reliably activate the “right” bundle of norms, not just some bundle of previously co-occurring norms. Achieving this reliability is made difficult by the fact that contexts are likely to show fluctuation in the specific set of features that instantiate a context in any particular case. A *DC* network relies on inferred context categories built right into the cognitive system, which creates robust invariance across feature fluctuations (because the learned norm-to-norm interconnections maintain the identity of context categories). An *IC* network would be far more sensitive to feature fluctuations. Every time a new feature combination emerges, it triggers a slightly different set of norms. So equivalence classes for what is the “same context” would be difficult to form. But because the *IC* model does not rely on abstract context representations and instead responds to natural, complex feature intercorrelations (that may, in reality, constitute true contexts), the reliability and invariance of the norm network is a direct function of the reliability and invariance of the world itself—the more the world fluctuates, the more an *IC* network offers finely adjusted sets of activated norms.

#### 1.4.6 In Dictu Norm Activation

So far we have analyzed norm activation *in situ*—that is, in real-world situations that offer a rich array of features, which can constitute contexts. But norm activation (and indeed, norm learning) often occurs *in dictu*, when one person tells another person to (not) act in a certain way “in church” or “when adults are around” or “when somebody just experienced a loss.” What would the *indirectly-connected model* say about such situations? Where are the specific features that would trigger the specific norms? Is this not a case in which contexts are like latent factors that directly trigger a bundle of norms that have become interconnected?

This situation does not actually cause a problem for the *IC model*. A minimalist model about norm interconnections does not have to be minimalist about concept-feature and feature-feature interconnections. It would be strange to deny, in light of the semantic network and category literature, that concepts such as “in church” could not activate a large number of features that then directly activate norms. The idea that context categories directly activate norms is in fact less plausible because the fuzziness of categories such as “in church” (in the physical building? in a cathe-

dral? during mass?) doesn't easily select for specific bundles of norms. The addressee would have to disambiguate the vague category (either in their own mind or by asking questions) and thereby "fix" the relevant features, which in turn would activate relevant norms.

### ***1.4.7 Context and Structured Organization***

We have illustrated how context interacts with the horizontal structural organization of norms—their direct connections or indirect co-activation patterns. Context can also exert a powerful influence on norm activation by means of vertical (hierarchical) structures in the norm system. When planning to go to a business meeting, for example, abstract norms such as "be respectful" might be activated merely by thinking about the meeting in advance. "Be respectful" by itself does not have specific action instructions, but when a moment arises in which a business partner says something obviously incorrect, the norm may translate down the abstraction hierarchy into a concrete instruction to remain quiet or to be expressly hesitant in one's corrective remark.

Context categories can also exert a powerful influence on norm activation by means of temporal structures. Driving up to the restaurant and reading the "Valet Parking" sign triggers a normative sequence of actions (parking the car at the sign, greeting the valet, passing the key, accepting a number tag, etc.). The activated norm may "reel off" a series of sequential instructions that are associated with one another, not necessarily as *norm* interconnections but as well-practiced *action* interconnections.

## **1.5 SUMMARY AND CONCLUSION**

For a robot to become ethical it will need to have a *norm capacity*—a capacity to learn, represent, activate, and apply a large number of norms that people expect one another to obey and, in all likelihood, will expect robots to obey. To build such a norm capacity we will need to make critical decisions about how such a norm system is organized and implemented in the robot's cognitive architecture. We have focused on the contrast between two models of how such a norm system might be organized—as *directly* or *indirectly* connected networks—and illustrated some of the questions that this contrast raises. At the same time, we have set aside countless other questions. For example, in designing a robot's norm network, how would the specific norms that apply within a community, as well as their triggering contexts, be identified? How would a computational norm network handle norm conflict—that is, cases in which features in a given situation activate norms with contradictory action instructions or incompatible state goals. And exactly how can a system expand and refine its norm network without suffering from serious interference among

its norms? Despite the many unanswered questions, we hope that delineating key properties of the human norm system and beginning to analyze logical and computational characteristics of this system will prove fruitful in the endeavor to make robots socially and morally acceptable participants in society.

## ACKNOWLEDGEMENTS

This work was in part funded by ONR MURI grant #N00014-14-1-0144.

## References

- Aarts H, Dijksterhuis A (2003) The silence of the library: Environment, situational norm, and social behavior. *Journal of Personality and Social Psychology* 84(1):18–28
- Andrighetto G, Villatoro D, Conte R (2010) Norm internalization in artificial societies. *AI Communications* 23(4):325–339
- Bicchieri C (2006) *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, New York, NY
- Bower GH (1970) Organizational factors in memory. *Cognitive Psychology* 1(1):18–46, DOI 10.1016/0010-0285(70)90003-4
- Brennan G, Eriksson L, Goodin RE, Southwood N (2013) *Explaining norms*. Oxford University Press, New York, NY
- Cialdini RB, Kallgren CA, Reno RR (1991a) A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In: Zanna MP (ed) *Advances in Experimental Social Psychology*, vol 24, Academic Press, San Diego, CA, pp 201–234
- Cialdini RB, Kallgren CA, Reno RR (1991b) A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In: Mark P Zanna (ed) *Advances in Experimental Social Psychology*, vol 24, Academic Press, San Diego, CA, pp 201–234
- Fehr E, Fischbacher U (2004) Social norms and human cooperation. *Trends in Cognitive Sciences* 8(4):185–190, DOI 10.1016/j.tics.2004.02.007
- Goldstein NJ, Cialdini RB, Griskevicius V (2008) A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research* 35(3):472–482, DOI 10.1086/586910
- Harvey MD, Enzle ME (1981) A cognitive model of social norms for understanding the transgressionhelping effect. *Journal of Personality and Social Psychology* 41(5):866–875, DOI 10.1037/0022-3514.41.5.866
- Hechter M, Opp KD (eds) (2001) *Social norms*. Russell Sage Foundation, New York, NY

- Malle BF (2015) Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology* [online first], DOI 10.1007/s10676-015-9367-8
- Malle BF, Scheutz M (2014) Moral competence in social robots. In: *IEEE International Symposium on Ethics in Engineering, Science, and Technology*, IEEE, Chicago, IL, pp 30–35
- Moor JH (2006) The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21, DOI 10.1109/MIS.2006.80
- Nourbakhsh IR (2013) *Robot futures*. MIT Press, Cambridge, MA
- Schank RC, Abelson RP (1977) *Scripts, plans, goals, and understanding*. Erlbaum, Hillsdale, NJ
- Sullins JP (2011) Introduction: Open questions in roboethics. *Philosophy & Technology* 24(3):233, DOI 10.1007/s13347-011-0043-6
- Ullmann-Margalit E (1977) *The emergence of norms*. Clarendon library of logic and philosophy, Clarendon Press, Oxford
- Šabanović S (2010) Robots in society, society in robots. *International Journal of Social Robotics* 2(4):439–450, DOI 10.1007/s12369-010-0066-7
- Van Berkum JJA, Holleman B, Nieuwland M, Otten M, Murre J (2009) Right or wrong? The brains fast response to morally objectionable statements. *Psychological Science* 20(9):1092–1099, DOI 10.1111/j.1467-9280.2009.02411.x
- Veruggio G, Solis J, Van der Loos M (2011) Roboethics: Ethics applied to robotics. *IEEE Robotics Automation Magazine* 18(1):21–22, DOI 10.1109/MRA.2010.940149
- Wallach W, Allen C (2008) *Moral machines: Teaching robots right from wrong*. Oxford University Press, New York, NY