Malle, B. F. (2018, February). *From binary deontics to deontic continua: The nature of human (and robot) norm systems.* Paper presented at the Robo-philosophy conference 2018, University of Vienna, Austria.

From Binary Deontics to Deontic Continua: The Nature of Human (and Robot) Norm Systems

Bertram MALLE[a]
[a] *Department of Cognitive, Linguistic, and Psychological Sciences, Brown University*

Abstract. To make artificial autonomous agents safe and beneficial contributors to society we may strive to equip them with norms. Previous models of what norms are and how they could be formalized have typically relied on binary deontic concepts (forbidden or not, obligatory or not, etc.). But human norms may not come as binaries; they may come as continua. In two studies, we show that people consistently and consensually distinguish between deontic phrases that span degrees of prohibition and degrees of prescription. In light of these results, formal systems for norms in robots must be expressive enough to handle such deontic continua.

Keywords. Robotics, moral norms, psychology, deontic logic

1. Introduction

As increasingly autonomous robots enter human communities at a rapid pace, we must ask how such entry can be safe and beneficial for human communities and society at large. One way to make artificial autonomous agents become safe and beneficial contributors to human communities is to equip them with norm competence, the ability to represent, learn, and follow the norms of its community [1]. Norms are not merely a subset of an agent's goals but rather constrain the agent's pursuit of goals. Were it not for the norms governing a particular context, an individual (human or artificial) would pursue a variety of personally beneficial actions that might not be beneficial for other community members. Norms therefore promise to be a natural mechanism to constrain robot behavior to make it fit in with other members of the community and be beneficial to them.

2. The Nature of Human Norms

But how would be go about building norm systems into robots? Our research group's approach has been to first understand in detail how humans cognitive represent and socially enact norm systems [1]–[3]. Naturally, robots might implement norms quite differently in their computational architectures from the way humans implement them in their neural architectures; but the fundamental principles of human norm systems must be maintained, or else people would not recognize robots as following social norms. Among these principles we have studied properties of norm representation (how norms are stored, activated, and organized), properties of norm learning (such as observation and instruction), and properties of norm enactment (how norm representations motivation action, how norm contradictions are reconciled).

But what are these "norms" that are represented, activated, learned, and reconciled? We have developed a working definition of norms that tries to capture the complexity of the norm concept in human social life, accommodates the fact that norms have to be somehow cognitively represented in human minds, and offers the prospect of formalizing norms (and nom systems) in artificial computational agents such as social robots. The definition builds on [1], inspired by [4] as well as [5]:

Definition: *A community's norm $\mathcal{N}$ is an instruction to (not) perform an action $\mathcal{A}$ in context $\mathcal{C}$, provided that a sufficient number of individuals in the community (1) indeed follow this instruction and (2) demand of each other to follow this instruction.*

That is, $\mathcal{N} := \mathcal{C} \rightarrow \mathbb{D}(\mathcal{A})$, whereby the deontic operator $\mathbb{D}$ is traditionally an instruction that $\mathcal{A}$ is "forbidden, "permissible," or "obligatory."

Numerous approaches to the formal representation of such instructions have been developed, refined, and debated within the field of deontic logic and in some cases applied to multiagent systems [6], [7], AI [8] or robots [9]. Alternative approaches have used variants of defeasible logic to aid in resolving norm conflicts [10], temporal logics to better represent contextual and temporal boundedness of norms [11], or prospective logic to represent more directly an agent's forward-looking action planning process [12]. All these and related approaches—indeed, virtually all formal representations of normative systems—have suffered from one major limitation: binary deontic concepts (for an exception, see [13]). In these systems, an action is either forbidden or not, permissible or not, obligatory or not. But human norms do not come as such clear-cut monoliths; they come in degrees. The second parameter in the above definition of norms likely refers to *degrees* of demand to (not) perform a certain action in a certain context. Indeed, the language of obligation includes degrees of demand expressed as actions that are "recommended," "encouraged," "required," or "mandatory"; the language of prohibitions includes actions that are "discouraged," "inappropriate," or "forbidden"; and the lists can easily be extended to permitted actions that are "tolerated," "allowed," or "optional."

One interpretation of this rich vocabulary of deontic phrases is that human language is vague and variegated and has many words for essentially the same underlying concepts or functions; such a muddled system should certainly not guide the design of moral robots. Another interpretation is that people use this rich vocabulary in systematic ways in influential norm communication: to teach norms of differential importance, affirm especially critical norms, enforce some norms more strictly than others, and negotiate norm conflicts, where a weaker norms is traded off against a stronger one. Under this interpretation, moral robots would certainly need to share this rich vocabulary and the graded deontic concepts underlying it—or else the robot would not count as norm competent.

Deciding on which of these interpretations is correct will be aided by empirical data on one fundamental question: Do people make consistent and consensual distinctions among verbal phrases that denote degrees of prescription, prohibition and permission? If there is such consensus and consistency, we need to understand how people represent these continua and must develop formal systems that are expressive enough to handle them. This, perhaps, would be the end of deontic logic as we know it.

3. Empirical Evidence for Continuous Norm Concepts

We report on two studies that answer the questions of consistency and consensus of deontic continua in (at least English) language users. In Study 1 we presented 57 participants with 22 phrases of prohibition (including some of permission) and asked them to rank-order them from “this means not prohibited” to “this means completely prohibited.” In addition, we presented a second group of 57 participants with 22 phrases of prescription (including some of permission) and asked them to rank-order them from “this means not prescribed” to “this means very strongly prescribed.” Figure 1 displays the results of prohibitions, suggesting that some families of phrases are near-synonyms but that the families themselves are clearly distinct and form a scale of weak to strong prohibitions.

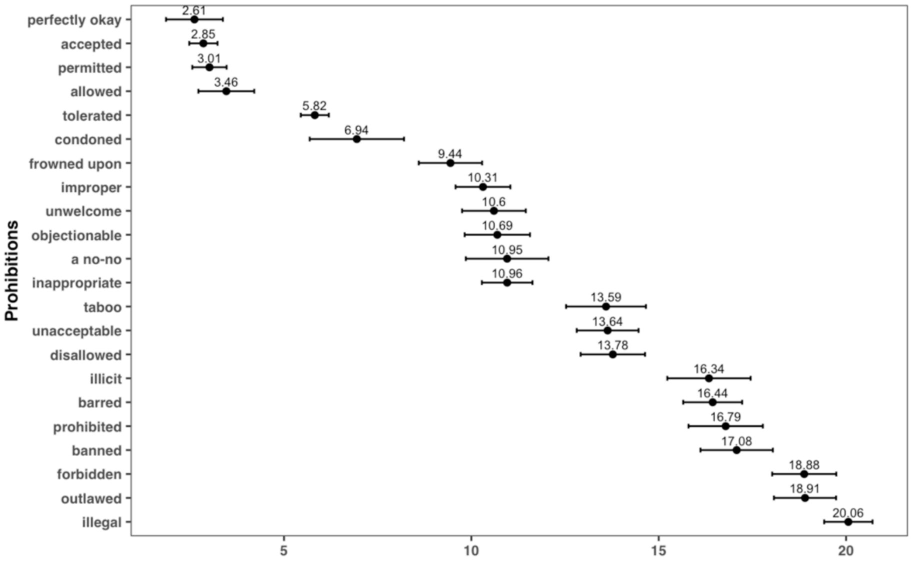

Figure 1. Twenty-two phrases of prohibition (including some of permission) are shown with their mean ranks (across 57 participants) and 95% confidence intervals around those means. At least six levels can be clearly distinguished, with the first two corresponding to permissions.

In Study 2 we selected many of the same phrases and added a few new ones to create two sets of prohibition (one with 12 phrases, the other with 13) and two sets of prescription (on with 10 phrases, the other with 11). Each set was presented to a group of 99 to 101 participants who were asked to assign each phrase a scale value from 1 to 10, corresponding to the range from “this means not at all prohibited” to “this means completely prohibited” or, for prescriptions, from “this means not at all prescribed” to “this means very strongly prescribed.” The results were largely consistent with the first study. Indeed, for those phrases that appeared in both the rank methodology of Study 1 and the rating methodology of Study 2, the correlation of average rank and average rating was r = .96 for prescriptions and r = .97 for prohibitions. At the top end of the rating scale, people found it more difficult to differentiate the most extreme deontic phrases from one another, likely because of a measurement ceiling effect. We therefore recently initiated a third study that modified the rating scale to create more graphical and psychological room at the top and instructed participants to more evenly distribute their ratings. We expect a full replication of our findings.

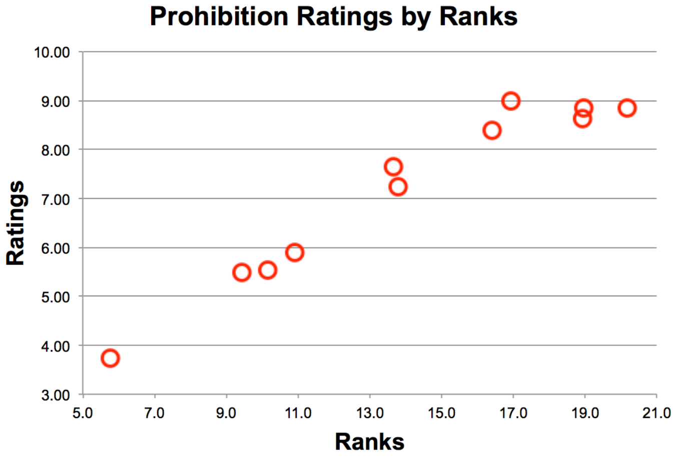

Figure 2. Eleven phrases of prohibition that were used both in Study 1 (using a rank-order methodology) and in Study 2 (using a rating methodology), plotted with their mean ranks (x axis) and their mean rating (y axis). The correlation between the ranks and ratings was r = .97

In light of these results we will be able to construct a measurement scale that uses has distinct verbal phrases to capture prohibitions of decreasing degrees at one end, permissions in the middle, and prescriptions in increasing degrees at the other end. For any given action in a given context we will then be able to assess where on that scale the deontic force for this action falls. Moreover, we will attempt to project this full scale of deontic force onto a single measurement interval from 0 (maximally prohibited) to 0.5 (optional) to 1.0 (maximally prescribed), making it amenable to a decision-theoretic framework we recently refined [14], [15], allowing us to formally represent a variable number of norms with variable degrees of normative force.

References

[1] B. F. Malle, M. Scheutz, and J. L. Austerweil, “Networks of social and moral norms in human and robot agents,” in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, Eds. Cham, Switzerland: Springer International Publishing, 2017, pp. 3–17.

[2] B. F. Malle and M. Scheutz, “Moral competence in social robots,” in *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics’2014*, Chicago, IL: IEEE, 2014, pp. 30–35.

[3] B. F. Malle, “Integrating robot ethics and machine morality: the study and design of moral competence in robots,” *Ethics and Information Technology*, vol. 18, no. 4, pp. 243–256, Nov. 2016.

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

[5] G. Brennan, L. Eriksson, R. E. Goodin, and N. Southwood, *Explaining norms*. New York, NY: Oxford University Press, 2013.

[6] R. Conte, R. Falcone, and G. Sartor, “Introduction: Agents and Norms: How to fill the gap?,” *Artificial Intelligence and Law*, vol. 7, no. 1, pp. 1–15, Mar. 1999.

[7] F. Dignum, “Autonomous agents with norms,” *Artificial Intelligence and Law*, vol. 7, no. 1, pp. 69–79, Mar. 1999.

[8] M. Anderson and S. L. Anderson, Eds., *Machine ethics*. New York, NY: Cambridge University Press, 2011.

[9] K. Arkoudas, S. Bringsjord, and P. Bello, “Toward ethical robots via mechanized deontic logic,” in *Machine Ethics: Papers from the AAAI Fall Symposium 2005*, vol. FS-05-06, 2005, pp. 17–23.

[10] G. Governatori and A. Rotolo, “BIO logical agents: Norms, beliefs, intentions in defeasible logic,” in *Normative Multi-agent Systems*, G. Boella, L. van der Torre, and H. Verhagen, Eds. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

[11] T. Ågotnes, W. V. D. Hoek, J. A. Rodriguez-Aguilar, C. Sierra, and M. Wooldridge, “On the logic of normative systems,” in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI ’07)*, M. Veloso, Ed. Palo Alto, CA: AAAI Press, 2007, pp. 1181–1186.

[12] L. M. Pereira and A. Saptawijaya, “Modelling morality with prospective logic,” in *Progress in Artificial Intelligence*, J. Neves, M. F. Santos, and J. M. Machado, Eds. Springer Berlin Heidelberg, 2007, pp. 99–111.

[13] M. Nickles, “Towards a logic of graded normativity and norm adherence,” in *Normative Multi-agent Systems*, G. Boella, L. van der Torre, and H. Verhagen, Eds. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2007.

[14] V. Sarathy, M. Scheutz, and B. F. Malle, “Learning behavioral norms in uncertain and changing contexts,” in *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen, Hungary, 2017.

[15] V. Sarathy, M. Scheutz, Y. N. Kenett, M. M. Allaham, J. L. Austerweil, and B. F. Malle, “Mental representations and computational modeling of context-specific human norm systems.,” in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, London*, 2017.