Silly Rules Improve the Capacity of Agents to Learn Stable Enforcement and Compliance Behaviors

Extended Abstract

Raphael Köster
DeepMind
rkoster@google.com

Dylan Hadfield-Menell
Department of Electrical Engineering and Computer Science,
University of California Berkeley
Center for Human-Compatible AI
dhm@eecs.berkeley.edu

Gillian K. Hadfield
Schwartz Reisman Institute for Technology and Society,
University of Toronto
Vector Institute
Center for Human-Compatible AI
OpenAI
g.hadfield@utoronto.ca

Joel Z. Leibo
DeepMind
jzl@google.com

KEYWORDS

multi-agent, deep reinforcement-learning, norms

ACM Reference Format:
Raphael Köster, Dylan Hadfield-Menell, Gillian K. Hadfield, and Joel Z. Leibo . 2020. Silly Rules Improve the Capacity of Agents to Learn Stable Enforcement and Compliance Behaviors. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 2 pages.

How can societies learn to enforce and comply with social norms? Many if not most human norms are functional. Rules that punish non-cooperative behavior, for example, support cooperation. An intriguing feature of human normativity is that many social norms concern behaviors that have no direct impact on material well-being. Examples include rules about what color clothing one wears to a funeral [7] or whether one uses one's left or right hand in particular tasks [2]. Such apparently pointless rules are ubiquitous, often acquiring great social meaning despite the absence of functionality. Hadfield-Menell et al. (2019) call these norms "silly rules" and distinguish them from "important rules," such as rules that govern resource sharing or prohibit harmful conduct, that directly impact welfare [3].

Here we investigate the learning dynamics and emergence of compliance and enforcement of social norms in a foraging game, implemented in a multi-agent reinforcement learning setting. In this spatiotemporally extended game, individuals are incentivized to implement complex berry-foraging policies and punish transgressions against social taboos covering specific berry types. Agents inhabit a 2-D grid-world in which they and other objects are located at coordinates in space. The atomic actions in an agent's action-space are moving up, down, left, right, rotating left and right and using a "punishing beam" (which allows players to remove rewards from other players, at a smaller cost to themselves, akin to third-party punishment). An agent perceives raw pixels. How these pixels relate to other agents or their actions must be learned. The behavior of the agent is driven by its learning to maximize the expected value of all future rewards it will obtain from its environment (e.g. by collecting berries). This learning over time is accomplished by incremental adjustment of neural network weights. This forms distributed neural representations that produce reward-maximizing behavior in response to visual input of the current situation. Agents learn continuously while being exposed to episode after episode, inhabiting the same environment with a population of other agents who learn simultaneously with them. In order to do this effectively, agents need to correctly assign credit to current stimuli and actions based on subsequent rewards they receive. This creates a rich dynamic in which every part of a behavior has to be learned, and strategic decisions have to be *implemented* via a behavioral policy . Both the cognitive challenge of correct credit assignment as well as performing complex action sequences are difficult and the dynamics of how norms are learned and implemented are endogenous to the multi-agent learning model.

The populations of agents are initialized under different conditions: 1. The "norm-free" condition has a poisonous berry but no social taboos. 2. In the "important rule" condition, consuming the poisonous berry is a social taboo. Agents who eat the berry are marked and can get punished by other agents for a reward. 3. The "silly rule" condition, which additionally to the "important rule", the poisonous berry being taboo, has a non-poisonous berry that also triggers the taboo.

Preprint is available at https://arxiv.org/abs/2001.09318

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

Video of example episode: https://youtu.be/Xn2eTSX-4GU. Consumption of taboo berry and subsequent punishment at 23-25 seconds. Note that agents see a lower resolution version of the environment in which each entity is represented by a single pixel.

The first thing agent populations learn is to reduce the amount of times that unmarked players are punished. Punishing unmarked players is costly to both the punished and the punishing agent, so it is unsurprising that this behavior does not persist long once agents learn less random policies.

The second important learning dynamic is that the number of times 'marked players' get successfully punished initially strongly increases before it decreases. We interpret the increase as an improvement in the agents' skill at enforcing the social norm, i.e. being increasingly skilled at effectively punishing marked agents. As a result, the amount of time agents spend marked is steadily declining.

This shows that there is a hierarchy in the learned behaviors, as first the social punishing system needs to be successfully implemented before it is possible for agents to learn that they should avoid breaking the social norm. In these two measures (successful punishments and taboo berries eaten) we see the role of the arbitrary taboo (one additional taboo berry) most clearly. Early in learning, it is unsurprising that double the amount of taboo berries leads to a higher amount of taboo berries eaten and subsequent punishing. Interestingly, once these quantities start to decline, they decline more rapidly in the condition with two taboos instead of one and in fact reach a lower level. So, it appears that increased exposure to taboo berries and punishing early leads to more robust learning. In terms of avoiding getting poisoned, having two taboos instead of one consistently leads to better results. However, the consistent benefit of the additional arbitrary taboo in terms of avoiding the poisonous berries does not in itself translate into a benefit in collective return. Collective return sums all rewards gained by all agents. If poisonous berries were avoided by agents just standing still or moving more slowly, the collective return would reveal that agents have not learned to forage successfully. Similarly, collective return factors in the cost of social punishing. This means that in order to achieve a benefit in collective return, the avoidance of poisonous berries has to be so substantial that it surpasses the costs associated with the social punishment scheme. This is actually the case in the intermediate learning stages. A series of experiments show that this benefit is larger when group sizes are large, and the credit assignment problem is harder (more berry types and longer delay after poison takes effect).

We demonstrate that a more complex rule set containing an arbitrary taboo, or silly rule, can lead to faster and more stable learning for reinforcement learning agents, supporting the initial finding in [3]. While the arbitrary taboo provided a consistent benefit in avoiding poisonous berries, it is worth noting that the benefit of the arbitrary rule on the overall prosperity of the group was only present in the intermediate stages of learning. This could be associated with the dead-weight cost of maintaining a social norm that does not serve a direct material function, or imprecise strategies to avoid poison (i.e. moving more slowly in general).

This line of research connects to the study of human cultural evolution and social norms that support complex group behaviors such as cooperation. We offer an explanation why arbitrary taboos may appear and are maintained, grounded in the mechanics of learning within a single group. This account is independent of, but not necessarily inconsistent with, existing explanations centered around in-group/out-group classification and group cohesion [5]. Our findings also echo results from the literature on cultural evolution that suggest larger group sizes can benefit learning and accumulation of culture [1, 4, 6]. In our account, this is due to the fact that larger groups, with higher population density, assist agents in learning to participate in the fundamental enforcement scheme. A higher density of agents could benefit learning by providing shorter paths to marked agents and increased number of observations of rule violations and subsequent punishing.

REFERENCES

[1] Maxime Derex, Marie-Pauline Beugin, Bernard Godelle, and Michel Raymond. 2013. Experimental evidence for the influence of group size on cultural complexity. *Nature* 503, 7476 (2013), 389.

[2] Daniel MT Fessler and Carlos David Navarrete. 2003. Meat is good to taboo. *Journal of Cognition and Culture* 3, 1 (2003), 1–40.

[3] Dylan Hadfield-Menell, McKane Andrus, and Gillian Hadfield. 2019. Legible Normativity for AI Alignment: The Value of Silly Rules. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 115–121.

[4] Joseph Henrich. 2004. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses - the Tasmanian case. *American Antiquity* 69, 2 (2004), 197–214.

[5] Victor Benno Meyer-Rochow. 2009. Food taboos: their origins and purposes. *Journal of Ethnobiology and Ethnomedicine* 5, 18 (2009).

[6] Adam Powell, Stephen Shennan, and Mark G Thomas. 2009. Late Pleistocene demography and the appearance of modern human behavior. *Science* 324, 5932 (2009), 1298–1301.

[7] Michael Tomasello and Amrisha Vaish. 2013. Origins of human cooperation and morality. *Annual Review of Psychology* 64, 1 (2013), 231–255.