Silly rules enhance learning of compliance and enforcement behavior in artificial agents

Raphael Köster[a,1], **Dylan Hadfield-Menell**[b,c], **Richard Everett**[a], **Laura Weidinger**[a], **Gillian K. Hadfield**[c,d,e,f], **and Joel Z. Leibo**[a,1]

[a]DeepMind; [b]Department of Electrical Engineering and Computer Science, University of California Berkeley; [c]Center for Human-Compatible AI; [d]Schwartz Reisman Institute for Technology and Society, University of Toronto; [e]Vector Institute; [f]OpenAI

This manuscript was compiled on April 25, 2021

How do societies learn and maintain social norms? Here we use multi-agent reinforcement learning to investigate the learning dynamics of enforcement and compliance behaviors. Artificial agents populate a foraging environment and need to learn to avoid a poisonous berry. Agents learn to avoid eating poisonous berries better when doing so is taboo, meaning the behavior is punished by other agents. The taboo helps overcome a credit-assignment problem in discovering delayed health effects. By probing what individual agents have learned, we demonstrate that normative behavior relies on a sequence of learned skills. Learning rule compliance builds upon prior learning of rule enforcement by other agents. Critically, introducing an additional taboo, which results in punishment for eating a harmless berry, further improves overall returns. This "silly rule" counterintuitively has a positive effect because it gives agents more practice in learning rule enforcement. Our results highlight the benefit of employing a computational model focused on learning to implement complex actions.

Multi-agent reinforcement learning | Norms | Third-party punishment

One of the central attributes that differentiates human from other animal societies and accounts for the enormous gains of human ultra-sociality (1) is the presence of third-party enforced norms (2–4). Many of these norms generate direct benefits for individual and group well-being: norms that prescribe reciprocity, fair sharing of rewards, or non-interference with property properly claimed by another, for example, can coordinate behavior and sustain incentives for cooperation and investment. These are the norms that are the primary focus of most research into the properties and origins of human normativity (see (3) for a review.)

The normative landscape is also, however, populated by many norms that appear essentially arbitrary: norms about how and what we eat, how we greet each other, what clothes and body decorations we wear, and what rituals we observe (5, 6). People treat compliance with these norms as important and punish violations, but, except for effects generated by this socially-constructed salience, they have no direct or first-order impact on welfare. Fessler et al. (5) call the process by which patterns of behavior are imbued with moral sentiments that motivate sanctioning of violations of the pattern *normative moralization.* They use as an example the normative moralization of handedness. Most people are naturally right-handed but, particularly in societies with few specialized tools, whether someone is right- or left-handed generally has no material consequences for others. Nonetheless, many cultures treat using one's right hand as a morally approved category–denoting purity or politeness–and one's left hand as cause for opprobrium–revealing weakness or evil (7). Following Hadfield-Menell et al. (8) we call such social norms *silly rules.*

The ubiquity of silly rules provides a puzzle for functionalist accounts of norms (9); several explanations have been explored so far. One kind of explanation posits that silly rules may exist to serve as cheap signals of group membership and thus facilitate cooperation within the group (10). Another account holds that silly rules are stable because, in any society, the survival of each generation depends on the transmission from prior generations of a large amount of culture-specific and, importantly, *causally opaque* knowledge (11). This includes everything from which local plants produce edible versus poisonous fruit, to how best to organize to resolve disputes between family members. Most of the time individuals have no way of knowing which of the many rules they follow are critical for their well-being. Thus silly rules may remain stable by virtue of their incorporation into larger normative systems that also include important rules (1). Further support for this hypothesis is found in the tendency of human children to over-imitate adults, copying—and moralizing—even apparently irrelevant aspects of adult behavior (12). The sheer abundance of silly rules seems to require an account that grants the normative moralization of seemingly irrelevant actions a more significant role. It would seem that a society would do better to minimize costly efforts to punish and conform with norms that produce no material benefits, and so to economize on the number of silly rules used as markers or retained as a by-product of the cultural transmission of knowledge.

In this paper, we describe a new kind of functional explanation for silly rules based on the dynamics of learning in a society that lacks *a priori* knowledge of which of their rules are truly important (causal opacity). Our explanation relies on an essential asymmetry between the enforcement and compliance aspects of normative behavior. In short, the skills involved in third party norm enforcement readily transfer from norm to norm, while the skills involved in compliance are norm-specific. Thus adding a silly rule to a normative system that already contains some number of deeply important rules can be beneficial because the silly rule may provide greater opportunity to practice third party norm enforcement, a generic skill. Improved norm enforcement by the group then makes it easier for individuals to learn from experience the skills necessary for norm compliance, such as how to prospectively recognize and avoid specific taboos. Therefore, introducing a silly rule may positively impact the learnability of compliance behavior for all of a society's rules, including those that truly are important. The benefit of learning important rules faster

RK, RE, LW, JZL conducted the simulation and analysis. All authors contributed to the research design and writing of the paper.

The authors declare no conflicts of interest.

[1]To whom correspondence should be addressed. E-mail: rkoster@google.com, jzl@google.com

can easily outweigh the dead-weight loss created by the silly rule.

Silly rules can support the emergence and stability of a beneficial *normative social order* (13). In a normative social order, group behavior is patterned on a classification scheme (called a norm) that divides behaviors into approved and disapproved (taboo) categories. Here, we employ a computational approach to investigate the effects of silly rules on how well a normative social order is learned. Our model consists of a multi-agent reinforcement learning (RL) environment with eight artificial agents, all simultaneously learning and interacting with one another. Agents in our environment (Fig. 2) are faced with learning a foraging task: learning to find and consume food ("berries"). We assume that berries are relatively abundant so there is no competition between agents and no common pool resource problem. What makes the environment challenging is the presence of a poisonous berry which if eaten will then reduce the value of an agent's future consumption. But importantly, the deleterious effect only triggers after a significant delay. The delay introduces a credit-assignment problem, meaning it is difficult for our agents to learn which particular berry caused the negative effect and thus to learn to avoid it. In this setting, a taboo on the poisonous berry–however it may evolve–raises individual welfare. As Boyd et al. (11) emphasize, this is a critical pathway by which culture raises human well-being: through the transmission of cultural practices, such as the avoidance of harmful foods, even when agents lack direct causal awareness of why their practices are beneficial (11, 14). The mechanisms behind such social learning in humans may be multiple: a psychological propensity to conformity (15, 16), deliberate teaching practices (17), and third-party punishment of failures to follow norms (18). Our work focuses on the last of these. We show that agents are able to sustain the transmission of a valuable taboo in order to avoid a poisonous berry. For this, agents need to have learned to recognize when another agent has violated a taboo and to deliver a costly punishment to the violator.

Because our model allows us to separate the learning of enforcement and compliance behaviors from the learning of norm content itself, we designed an experiment in which norm content was fixed in advance by the experimenter (which color berries were taboo). By varying the content of the norms, we can study the downstream effects on how the normative social order (enforcement and compliance behavior) is learned. If a player breaks a taboo they change color and become 'marked'. We assumed all agents have a form of mutual knowledge of the rule in the sense that they may perceive violations of other players via the marking. Note that players cannot directly perceive their own marking, otherwise the self-marking would trivially solve the credit assignment problem. If a player is marked, other players can collect a reward for punishing them. This creates an incentive for players to learn to punish rule violations, and for players to not violate any rules. This reflects situations in which there is a centralized scheme that labels transgressive behavior, but the enforcement is decentralized. For example, in medieval iceland the 'law speaker' would label acts as unlawful. Individuals would be declared 'outlaws', and others could take their property without repercussions (19).

Our environment is intended to capture the evolutionary challenge humans faced in developing the behavioral and cognitive repertoires of normativity and third-party punishment, attributes that distinguish humans from other primates and account for the gains humans enjoy from ultrasociality and extraordinary gains from cooperation (2, 3, 20, 21). It is the challenge of discovering and learning these behaviors–to punish violators and to avoid punishment through behavior modification—that animates our study. We demonstrate that simple RL agents that lack any "mental" models of rules, violations, or punishment, will nonetheless learn to enforce the rules. The enforcement subsequently enables agents to learn to comply with a taboo against eating the poisonous berry.

As a precondition to our main analysis, we show that individuals achieve higher overall welfare in a world where eating the poisonous berry is taboo, relative to a world in which there are no taboos so avoidance of the poisonous berry must be learned only through individual experience. We show that even with the cost of enforcement, overall group welfare is higher with a norm than without. Thus, the normative social order is valuable. We then show our main result: that the value of a normative order is *higher* if the norms in this regime include not only important rules—such as the rule against eating poisonous berries—but also silly rules which make the eating of a harmless berry taboo and bring about the same third-party punishment. In our environment, agents learn to enforce, and comply with, norms more quickly if the rule system includes two taboos—one against eating the poisonous berry, and one against eating a harmless berry.

Our results demonstrate a new account of the ubiquity of rules that have no first-order impact on well-being. They also provide a formalization of normativity in a computational setting that we think will expand the tools available both for understanding how human normativity operates, and how artificial agents that are capable of participating in human normative social orders might be built. In this sense, this work is part of a research program that ultimately aims to develop models capable of capturing distinctive features of human intelligence such as the origin of institutions (22).

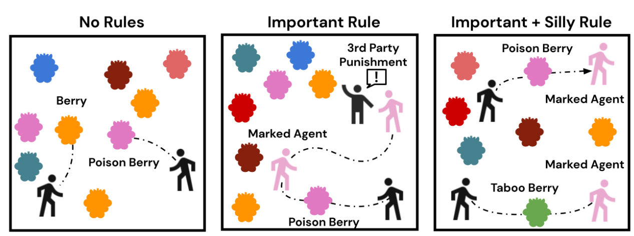

Fig. 1. Schematic overview of the experimental conditions. In the *no rules* condition, agents collect berries for reward. One berry-color is poisonous and after a time delay reduces reward obtained from consumed berries. Being poisoned is invisible to all players and a hard credit-assignment problem. In the *important rule* condition, eating the poisonous berry is a social taboo. When eaten, the player who ate the berry immediately gets marked, which is only visible to other agents. Other agents can collect a reward by punishing a marked agent. In the *important+silly rule* condition, the same taboo against the poisonous berry is in place, but additionally there an identical taboo on a berry that is not poisonous. Therefore, there are two taboos for which agents can experience punishment by others. Our experiment sets out to study the effects of these different normative schemes.

Studying social norms with deep reinforcement learning. Computational simulations of populations and cultural development typically use an abstracted or idealized space to encode normative structure (23–26). Agents are usually mod-

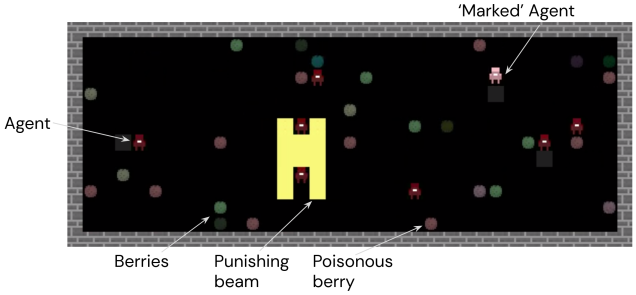

Fig. 2. Depiction of the environment. The agents inhabit a grid world. Agents earn reward for eating berries, which regrow probabilistic after being harvested. One type of berry is poisonous and if collected by an agent, it diminishes the agent's ability to gather rewards from other berries, after a delay period. If an agent eats one of the poisonous berries in the *important rule* condition, the agent immediately gets "marked" and appears in a different color to the other agents. In the *important+silly rule* condition, one additional, non-poisonous, berry also triggers an agents' marking. Agents are able to punish each other using a "punishing beam", causing a loss to themselves and a large loss to the punished agent. If a "marked" agent is punished, the punishing agent receives a large reward.

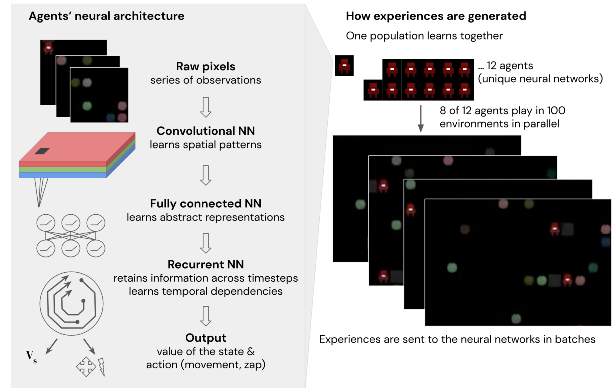

Fig. 3. Agent architecture and training procedure. Agents learn together in one population of 12 agents, 8 of which are selected to play in one episode in order to generate experiences (in multiple parallel environments). Each agent contains an independent neural network that receives a batch of its own experiences from these environments to update its neuronal weights. The inputs the agents receive are the raw pixels from their field of view. The network architecture of each agent consists of a convolutional neural network that learns to decompose the input into spatial patterns. This projects to fully connected layers that learn more abstract representations of game states, followed by a recurrent network that is able to retain and transform information over multiple timesteps. The output of neural network on each timestep is a prediction of the value of the current state and an action (movement or zap). Network weights are gradually adjusted to maximize long term cumulative reward.

eled either as choosing strategies within a game theoretic framework in which they are supplied with a set of available actions and associated payoffs, or as implementing behavioral rules in competition with other similarly-constructed agents. In these approaches, agents choose what to do (e.g. cooperate or defect), but the models cannot capture phenomena related to how they learn to implement their choice.

Here we apply a more generalized framework, a multi-agent RL approach that has been successfully used to study intertemporal (sequential) social dilemmas (27–37). Agents in this framework are artificial neural networks, which learn behavioral policies (associating actions to states) and obtain rewards from an environment. State transitions in the environment are generated by the actions of all agents combined. Agents inhabit a 2-D world in which they and other objects are located at coordinates in space. An agent's action-space consists of moving up, down, left, right, rotating left and right and using a "punishing beam" directed at an adjacent agent (Fig. 3). Use of the punishing beam costs the punisher 20 points and inflicts a cost of 35 on the punished agent. A variety of 'berries' of different colours are distributed randomly throughout the world. An agent receives a reward of 4 if it navigates to a square with a berry, interpreted as "eating it". Berries grow in sufficient abundance that there is no competition between agents. Pink berries are poisonous: 100 timesteps after consumption they reduce reward gained by future berries to 1 point.

The environment can include a latent classification scheme (13) that designates some berries (colors) as "taboo". This normative classification is implemented by inducing a change in the color of an agent (visible only to other agents) who consumes a taboo berry and changing the payoff associated with use of the punishing beam against such an agent. Punishing a marked agent generates a reward for the punisher of 15 instead of a loss of 20 (note that punishment is always net negative for the collective reward of the group). We consider the environment in three conditions corresponding to three different classification schemes. In the *no rules* condition, no berries are designated as taboo. Agents never become marked and punishing is never profitable. In the *important rule* condition the poisonous berry is taboo. In the *important+silly rule* condition, both the poisonous berry and another, harmless, berry are taboo. We manipulate the classification scheme to assess its causal effect on learning dynamics. We hypothesize that overall returns are improved by adding the important rule, and are further improved by adding the silly rule.

An agent in our environment has no prior knowledge of game rules or states. It has no model for the classification scheme or the potential rewards for appropriately-directed use of the punishing beam. The agent has to learn how its actions, its observations (raw pixels), and the rewards it receives relate to each other entirely from scratch. It can do this by learning representations that allow it to generalize between similar situations. Given the enormous game-space and the fact that all agents have incomplete information about the workings of the environment, classical game-theoretical analysis is not tractable. Further, this approach allows us to analyze learning dynamics, not just equilibria (31).

In this framework, behavior is driven by individuals learning to maximize the expected value of all future rewards they will obtain from their environment (e.g. by collecting berries, avoiding and delivering punishment). This learning over time is accomplished by incremental adjustment of neural network weights (38). It generates 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 are themselves learning simultaneously. 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 (determining which actions contribute to rewards over time), as well as figuring out how to perform complex action sequences are difficult. The dynamics of how norms are learned and implemented are endogenous to the multi-agent learning model. This leads to a number of important differences from more abstracted simulations like matrix games. We argue that, by focusing on learning, this computational model may be particularly appropriate to model anthropological phenomena like the the emergence and importance of social norms. In particular the model creates rich learning dynamics for individual agents as well as groups that could not otherwise be approached:

1. **Complex action sequences** Punishing other agents' behavior, observing a rule violation or complying with a rule are complex sequences of atomic actions that can look different each time they are performed or observed.
2. **Skills build on each other** As agents have to learn to implement complex behaviors, we can expect a temporal dependency and sequentiality among these behaviors. For example, for agents to learn to avoid a taboo, agents will first need to learn how to effectively apply punishing, in order to motivate rule compliance.
3. **Opportunity cost** As agents are driven by maximizing total reward, whether or not an agent engages in social punishing depends on the opportunity cost of the action sequence, the agent's skill in implementing it, and the reward gained by punishing the other agent's transgression. This means there is an intrinsic economy to behavior that is bounded by what agents have learned.
4. **Generalization** Since the social dynamics are learned in neural networks from scratch they afford the opportunity for, or even necessitate, a degree of generalization. In particular, as punishment is identical for the consequences of transgressing against an important or silly rule, there is an opportunity for generalization of enforcement behavior learned from both rules.
5. **Endogenous errors** As social punishing of silly or important rules is implemented in the same way, a confusion between the two can arise. Similarly, punishing might be misdirected at agents that did not break a social taboo. These costly false-positive incidents provide an intrinsic counterweight to the development of an indiscriminate social punishing dynamic. Importantly, unlike other frameworks, multi-agent RL does not require us to model mistakes in behavior as random noise (37, 39). Instead, mistakes in multi-agent RL are emergent from the learning dynamics and the inherent difficulty of implementing an effective behavior policy.

Results

As displayed in Fig. 4, we examine group-level metrics about agent-populations over the trajectory of learning. We plot the average trajectory per condition. As visible in Fig. 4A, the first thing agent populations learn is to reduce the frequency with which 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 actions become less random. As can be seen in Fig. 4F, this rapid initial learning increases the collective return (the sum of rewards gained by all agents). Note that the suppression of misdirected punishing happens fastest in the *no rules* condition. This is unsurprising, as in this condition there is no direct incentive to punish any other players at all, because there are no taboos that lead to marked players.

The second important learning dynamic is that the number of times marked players get successfully punished initially strongly increases before it decreases (Fig. 4B). 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 displayed in Fig. 4C, the amount of time agents spend marked is steadily declining. However, taken by itself, this metric does not differentiate between whether this decline is driven by agents becoming better at avoiding rule violation, or whether agents get better at punishing rule breakers and thereby removing their mark. As can be seen in Fig. 4E, the decline of successful punishments coincides with a decline in the number of taboo berries eaten. This shows that there is a sequence 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 silly rule (one additional taboo berry) most clearly. Early in learning, it is unsurprising that doubling the number of taboo berries leads to a higher number of taboo berries eaten and subsequent punishing. But 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. This is evident in later stages of learning where agents eat fewer taboo berries in the condition in which there are twice as many.

As can be seen in Fig. 4D, in terms of avoiding getting poisoned, having two taboos instead of one consistently leads to better results. Additionally, we can see that the credit assignment problem of avoiding the poisonous berry without the help of a social punishing mechanism is prohibitively hard (the *no rule* condition shows no decrease). However, avoiding poisonous berries is not in itself enough to increase overall returns (Fig. 4F). In this environment, overall returns are positively affected by gathering berries and negatively affected by eating poisonous berries and agents punishing each other (every punishment event creates a net loss for the group). Therefore, in order for the additional silly rule to yield an overall benefit in collective return, it needs to not only help to avoid poisonous berries, but it also needs to not add too much dead-weight loss by either hampering consumption of healthy berries or increasing punishment events. As shown in Fig. 4F, there is an overall benefit on collective return in the intermediate learning stages. We test the difference in collective returns between the *important rule* and *important+silly rule* condition in 10 separate timebins. There is a significant benefit of the arbitrary rule condition in the 3rd, 4th and 5th timebin (3rd: $t(28)=3.94$, $p=0.0005$, 4th: $t(28)=3.26$, $p=0.003$, 6th: $t(28)=2.43$, $p=0.022$. The 3rd and 4th timebin remain significant after Bonferonni-correction for 10 multiple

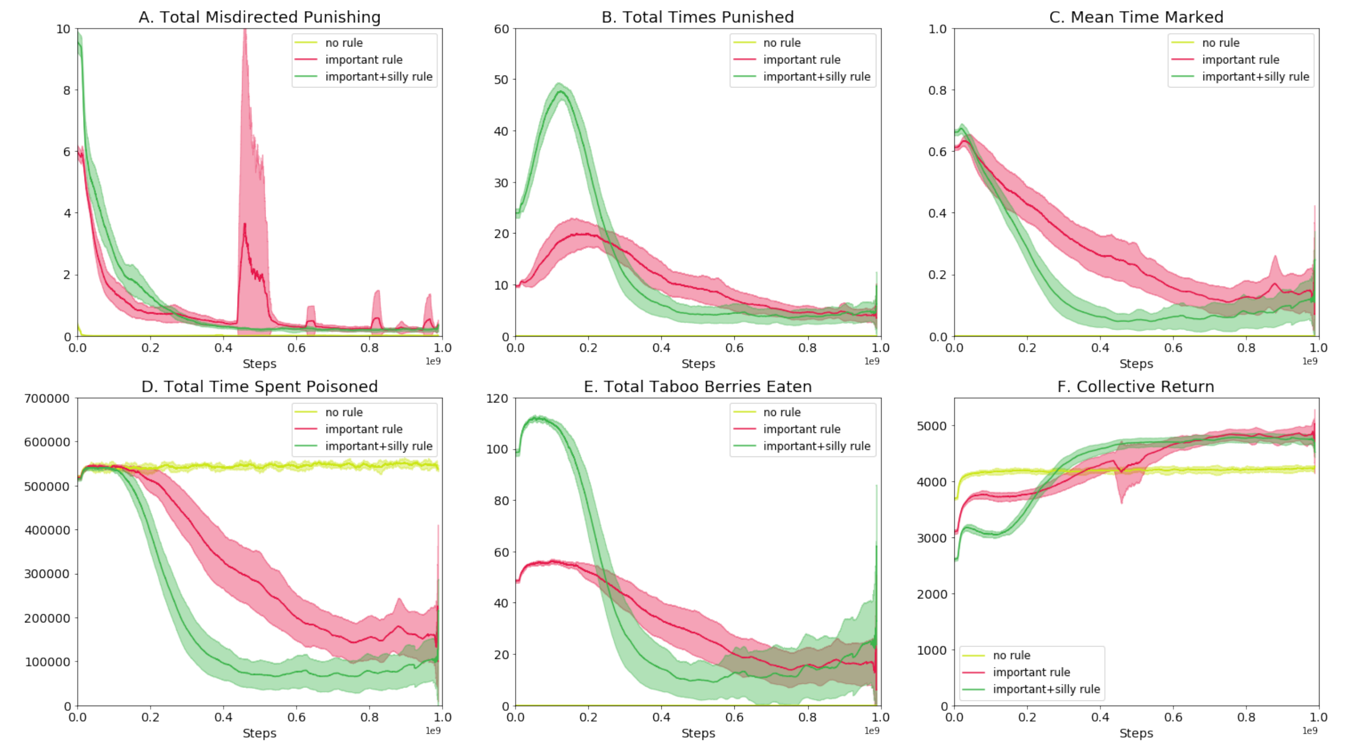

Fig. 4. Learning dynamics: We are examining group-level metrics about agent-populations (y-axis) over the trajectory of learning (x-axis in timesteps). We plot the average trajectory per condition (with 99% confidence interval). A. Number of times unmarked agents are punished (agents that have not broken a taboo). B. Number of times marked agents are punished (agents that have broken a taboo). C. Time spent marked after breaking a taboo. D. Time agents spent poisoned (timesteps after eating the first poisoned berry). E. The number of "taboo" berries eaten (poisonous and non-poisonous combined, if available in the condition). F. Total sum of reward gained by group (including costs of punishing). In total, we observe a benefit of the *important+silly rule* condition in the intermediate stages of learning, driven by an increased ability to avoid poisonous berries. We also see a temporal order to learned behaviors, e.g. an increase in social punishment that then declines together with a decrease in number of taboo berries eaten.

comparisons).

The results in Fig. 4 suggest that more frequent punishment early in learning is associated with less time spent poisoned in the middle stages of learning. We directly test this hypothesis by exploiting the variance across different training runs (i.e. separate populations, Fig. 5). In both conditions we find that high rates of punishment in the early stages of learning (mean over time, timesteps 0 to 2e8) are related to low amounts of time spent poisoned in subsequent stages (timesteps 2e8 to 4e8) (*important rule*: $r = -0.79$, $p = 0.0004$; *important+silly rule*: $r = -0.5$, $p = 0.057$, $n = 15$). Note that the correlation is lower in the *important+silly rule*, but that the magnitudes of both measures differ strongly. It is possible the correlation is less pronounced because adding the silly rule increases the magnitude and restricts range of the rate of punishment (on the y-axis).

Probing what agents have learned. Large-scale observational longitudinal studies with multiple actors face the problem that all actions taken are entangled and interdependent because agents react to other agents. Studying the effects of multiple agents' interactions over time allows us to investigate the effects of social norm enforcement on the population at large, but does not enable conclusions about what specific mechanisms cause an individual agent's behaviour. For humans, psychology experiments address this issue by isolating specific mechanisms and testing these in controlled conditions, such as testing reactions to particular stimuli in laboratory experiments. Our simulation allows us to follow this logic and confront our artificial agents with tightly controlled experimental environments inspired by lab-testing to directly probe what the agents have learned. As shown in Fig. 6A, we implement these quasi-lab experiments by extracting agents at different points in training and recording their actions when placed in a simple empty environment with no other agents, and only one stimulus to interact with. Critically, the agent is not learning in this environment. Running this experiment with multiple copies of the same agent allows us to run multiple trials to probe an agents' response to a particular game object in isolation. This tests what the agent has learned at different stages of training. Even though these tests constitute environments that the agent has not seen during training, the behavioral results align with what the agent is expected to learn in its training environment. This is particularly interesting because it requires a degree of generalization from the agents ('zero shot', as the agents do not learn during the probe). Their successful transfer of behaviors learned in a large complex environments to an empty testing environment indicates that they learned robust behavioral responses to game objects.

Fig. 6 B, C & D displays how many timesteps it takes agents to approach different berries when confronted with the berries in isolation. These approach-behaviors vary over the course of training and by the conditions the agents have learned in. As expected (cf. Fig. 4D), agents learn to avoid the poisonous berry (pink) in *important rule* and *important+silly rule* but not in *no rule*. Additionally, agents learn to avoid

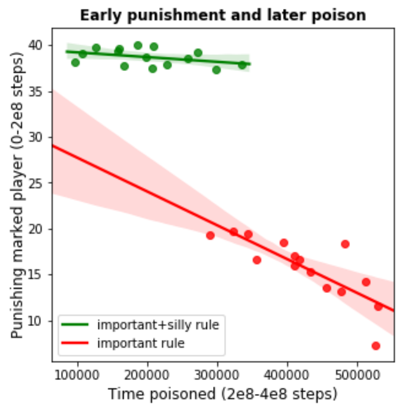

Fig. 5. Higher rates of early punishment is related to less time spent poisoned later in training. Each marker is an independent population of agents. On the y-axis we plot how often players are punished early in training (between timesteps 0 and 2e8). On the x-axis we plot the amount of time players spend poisoned subsequently (mean of the timesteps 2e8 to 4e8). The results are consistent with the interpretation that a high peak in punishment early in training is followed by more avoidance of the poisoned berry later.

the non-poisonous taboo berry (green) in *important+silly rule*. Fig. 6E overlays the poisonous berry lines from panel B and D and illustrates that agents learn to avoid the poisonous berry more in *important+silly rule* than *important rule*. Similarly, Fig. 6F illustrates that agents punish marked players more during learning in *important+silly rule* than *important rule*.

Again, we set out to test the hypothesis that learning about punishment early in training is associated with subsequent avoidance of the poisoned berry. Fig. 6G mirrors the results Fig. 5, demonstrating that the single-player probes are consistent with the behavior observed in the multi-agent setting. We correlate the degree to which a probed agent punishes the marked player during the early stages of training (mean of timewindow 0 to 2e8 steps, marked in panel F) with that agent's subsequent tendency to consume the poisonous berry (mean of timewindow 2e8 to 4e8 steps, marked in panel E). In both conditions we find a negative relationship (*important rule*: $r = -0.86$, $p < 0.0001$; *important+silly rule*: $r = -0.46$, $p = 0.085$, $n = 15$). Again, note that the absolute magnitude of the values differs across conditions; all datapoints in *important+silly rule* are restricted to relatively high punishment and low rates of approaching the poisonous berry. In sum, these results support the conclusions drawn from the full multi-agent simulation: the additional taboo leads to more frequent punishing events earlier during training, which in turn supports agents' learning to avoid the poisonous berry. Crucially, these results were obtained in a controlled experimental setting that directly probed what agents have learned by observing their reactions to single objects. These results suggest a sequential, social acquisition of skills, explaining why silly rules help agents learn and behave according to meaningful rules.

Discussion

This work contributes a functional account of why human normative systems contain so many silly and arbitrary rules that is grounded in the mechanics of learning within a single group. The presence of silly rules creates the potential for a larger number of norm violations. From the perspective of an agent learning the skills necessary to effectively enforce their society's norms the additional violations constitute additional opportunity for practice, and thus promote a faster rate of improvement in their command of the mechanics of third-party punishment. On the compliance side, the rate at which individuals may learn by trial-and-error to avoid violating taboos depends on the enforcement environment they inhabit. When their groupmates implement highly effective third-party enforcement strategies then exploratory taboo violations are punished both swiftly and surely. Since both speed and certainty of reward (or punishment) are factors known to improve trial and error learning (40, 41), highly "effective" compliance policies (i.e. policies that avoid violating taboos) can be learned rapidly under these conditions. On the other hand, when third-party enforcement is ineffective, then exploratory taboo violations frequently go unpunished or their punishment comes only after a substantial delay. Such conditions are known to make trial and error learning very difficult and slow. Enforcement and compliance are asymmetric in the sense that the former is a skill that may be applied without modification to any norm since many of the sub-behaviors involved in third-party punishment are directed toward the violator (e.g. chasing them), not toward the event of the violation itself. Thus they are "transferable skills", generically applicable to any norm. Compliance, on the other hand, requires learning to recognize for oneself what would constitute a violation. Now consider also that every society contains a certain number of deeply important rules for which ensuring compliance is of paramount importance. The interpretation of our key result is that the functional role of silly rules in human normative systems may (in part) be to help train a society's ability to comply with important rules. Adding silly rules into a normative system that already contains deeply important rules can be expected to improve the learning of enforcement for all rules, thereby improving the learning of compliance for all rules, including the rules that truly matter.

While this account is consistent with previous findings on the potential benefits of silly rules (8), the present study demonstrates a novel, mechanistic benefit of silly rules where silly rules improve the scale of enforcement practice, causing a concomitant improvement in the learnability of compliance. We interpret this result as indicating that silly rules enrich the information environment for agents that face a learning challenge. In Hadfield-Menell et al. (8) agents faced a challenge of estimating the likelihood that there are enough agents willing to punish rule violations in a group. In our paper, agents face the more fundamental challenge of learning the relationship between the visual information and actions available to them and the consequences of the two in terms of reward. The silly rule enriches this learning environment with more opportunities to learn about the relationship between punishment behavior and the associated reward for the punisher, as well as the negative consequences for the punished. This account is also independent of, but not necessarily inconsistent with, existing explanations centered around in-group/out-group classification and group cohesion (42). In the real world, adding important rules may be difficult, as they require causal insights into how to avoid undesirable outcomes. Silly rules can be created as needed and, if they are not too costly, the normative order may benefit from the practice that violations against silly rules provide.

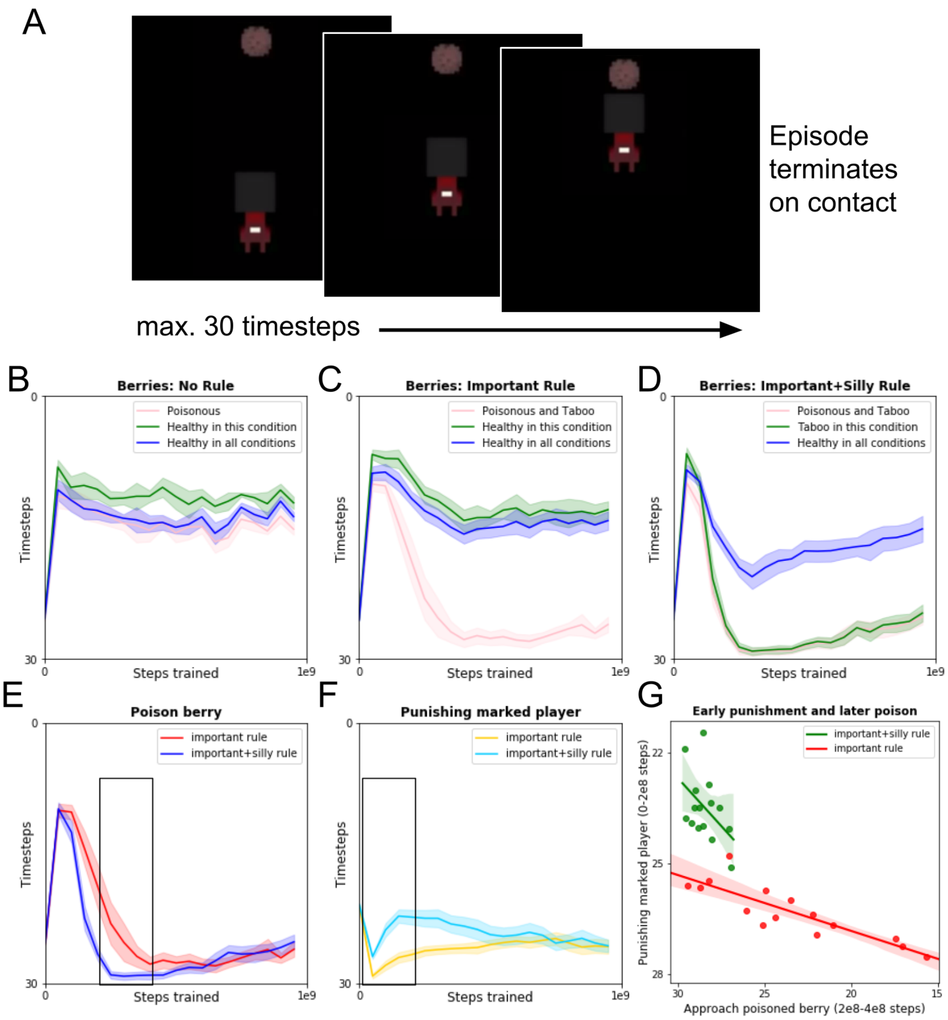

Fig. 6. Single target probes or 'zero shot generalization'. A. Depiction of probe. An agent is placed in an empty room with just one other object (berry or agent) and we measure how many timesteps it takes to eat the berry or zap the player. B, C & D. Berry types across the 3 different conditions. Agents learn to avoid berries that are taboo. Lines depict the mean across populations of how quickly the agent interacts with the object (y-axis) over learning (x-axis). Error bars represent SEM over different independent populations. E, F. Difference between 'important rule' and 'silly rule' for approaching the poisoned berry (same as in C & D) and punishing the marked player. Agents are faster to learn to avoid the poisoned berry and punish taboos in the *important+silly rule* condition. G. Early punishing (mean 0 to 2e8 steps) of the marked player is associated with reduced consumption of the poisoned berry (mean 2e8 to 4e8 steps) later in training.

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) (43). These costs suggests a strong counterweight to the usefulness of silly rules in the real world.

A clear limitation of this work is that we have not shown the emergence of the social norms themselves. We supplied in the environment the causal relationship between an action—eating a particular berry—and the trigger for social punishing: becoming marked in the view of other agents and generating a reward for an agent who successfully aimed the punishing beam at the transgressor. The next steps in this line of work are therefore to study the emergence of particular patterns of marking—norms—and the capacity for norms to change in response to changes in the environment or other sources or variation including natural drift. We hypothesize that learning how to follow and maintain social norms can assist agents in adapting to variation in the environment. This social technology of benefiting from norms is closely related to the cultural niche (11) inhabited by humans, and to humanity's intelligence and success. Understanding how this technology emerges in multi-agent settings may play a critical role in understanding the emergence of human-level intelligence.

Materials and Methods

Multi-Agent Reinforcement Learning. We consider multi-agent reinforcement learning in partially-observable general-sum Markov games (44, 45). In each game state, agents take actions based on a partial observation of the state space and receive an individual reward. Agents must learn through experience an appropriate behavior policy while interacting with one another. We formalize this as follows: an N-player partially observable Markov game $\mathcal{M}$ defined on a finite set of states $\mathcal{S}$. The observation function $\mathcal{O} : \mathcal{S} \times \{1, ..., N\} \to \mathbb{R}^d$, specifies each player's d-dimensional view on the state space.

In each state, each player i is allowed to take an action from its own set $\mathcal{A}^i$.

Following their joint action $(a^1, ..., a^N) \in \mathcal{A}^1 \times ... \times \mathcal{A}^N$, the state changes obeys the stochastic transition function

$\mathcal{T} : \mathcal{S} \times \mathcal{A}^1 \times ... \times \mathcal{A}^N \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of discrete probability distributions over $\mathcal{S}$, and every player receives an individual reward defined as

$r^i : \mathcal{S} \times \mathcal{A}^1 \times ... \times \mathcal{A}^N \to \mathbb{R}$ for player i. Finally, let

$o^i = \{\mathcal{O}(s, i)\}_{s \in \mathcal{S}}$ be the observation space of player i.

Each agent learns, independently through its own experience of the environment, a behavior policy $\pi^i : \mathcal{O}^i \to \Delta(\mathcal{A}^i)$ (written $\pi(a^i|o^i)$) based on its own observation $o^i = \mathcal{O}(s, i)$ and extrinsic reward $r^i(s, a^1, \ldots, a^N)$. Each agent's goal is to maximize a long term γ-discounted payoff defined as follows:

$$V^i_{\vec{\pi}}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, \vec{a}_t) | \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right]. \quad [1]$$

Experiment and conditions. We consider a foraging task implemented as a partially observable Markov game on a 2D grid (see Fig. 2). Agents gain reward by collecting berries that stochastically respawn. The respawn probabilities are high, so there is little competition for resources. Moving onto the coordinates of a berry, agents earn a reward of 4 points. Each berry type is consistently mapped to a color (24 different types). One berry type is "poisonous". There is no other signal of which berry type is poisonous that is observable to an agent at the time of consumption, except the color that remains consistent for all episodes. If collected by a player, this player is "poisoned" after a delay of a fixed number of timesteps (100 timesteps). Poisoning reduces a player's ability to absorb nutrition: after poisoning sets in each subsequent berry the player collects yields a reward of 1 instead of 4. Besides moving, agents have in their behavioral repertoire the ability to apply a "punishing beam". If successfully targeted at another player, the user of the beam loses a reward of 20 (the cost of punishing, in addition to the opportunity cost of time spent aiming and firing the beam instead of collecting berries) and the punished player loses a reward of 35*.

Each instance of the training regime is initialized in one of 3 different conditions. This is a between-subjects design: each agent population only experiences one of these 3 conditions. The conditions differ in the content of the classification scheme that marks agents if they have broken a taboo. We consider three conditions: no berry is taboo (*no rule*), the poisonous berry is taboo (*important rule*), the poisonous berry and one harmless berry are taboo (*important+silly rule*).

In *no rule* there are no additional mechanics to the game beyond what is described above. Agents have to learn which berry is poisonous without any additional information.

In *important rule* we introduce a group rule against eating the poisonous berry type. In this condition, a player that eats a poison berry is "marked": from the perspective of other agents in the environment, the marked player changes color. This color change is not visible to the marked player. This color change implements the idea that other agents evaluate the consumption behavior of the agent that has chosen to eat a "taboo" food. This marking then interacts with the punishing capacity of other agents. If a marked player is successfully targeted by another player with a punishing beam, the punishing player gets a reward of 35—effectively transferring reward from the marked player to the punishing player, for a net payoff to the punishing player of 15 points (note that when considering the sum of rewards of the whole group, a successful punishment results net-loss for the group of 20 points because of the cost of using the punishment beam). Aiming punishment at a non-marked player is costly to both as in the *no rule* condition. Once punished, the marking disappears.

In *important+silly rule*, we augment the important rule with an additional silly rule, or arbitrary taboo. Players become marked not only if they consume the poisonous berry but also if they consume another designated, but harmless, berry. As in the *important rule* condition, successful punishing of an agent that has violated the silly rule by consuming the designated harmless berry earns the punishing agent a net of 15 points and costs the transgressing agent 35 points. Thus, from the perspective of the agents, the "important" and "silly" rules are isomorphic if they have not integrated knowledge of the actual poisoning dynamic.

Note that in these settings classification scheme is implemented by the environment. We have not modeled the emergence of the rules in themselves. Agents are incentivised to learn policies that implement the behaviors of collecting berries, delivering third-party punishment, and avoiding taboo berries that create a risk of punishment.

Agent architecture and training method. Each instance of the training regime contained a population of 12 learners. The environment is a gridworld of size 33×12 pixels and agents observe a 15×15 pixels RGB window, centered on their current location (note that the depictions in this paper are higher resolution for display purposes). On each episode, a subset of learners was drawn without replacement to play in the current episode (8 players in each episode). Each episode lasted for 1000 steps. For each timestep s, each learner i in the population produced a policy π^i and an estimate of the value $V^i_{\vec{\pi}}(s)$ with a neural network, implemented on a GPU. This neural network was trained by receiving importance-weighted policy updates (46) sampled from a queue of trajectories. These trajectories were created by 64 simultaneous environments on CPUs that play the game (with 8 players, which used policies sampled uniformly from the population of learners without replacement). The learners received truncated sequences of 100 steps of trajectories in batches of 16.

The neural network's architecture consisted of a visual encoder (2D-convolutional neural net with 6 channels, with kernel size and stride size 1) followed by a 2-layer fully connected MLP with 64 RELU-neurons in each layer, an LSTM (128 units) and finally linear policy and value heads, outputting the value of the current state and a probability over actions to be chosen. We used a discount-factor of 0.99, the learning rate was 0.0004, and the weight of entropy regularisation of the policy logits was 0.003. We used the RMS-prop optimiser (learning rate=0.0004, epsilon=1e-5, momentum=0.0, decay=0.99). The agent also minimized a CPC loss (47) in the manner of an auxiliary objective (48).

Statistical analysis of observational data. In order to assess the difference between conditions, we divide the learning timecourse into 10 bins and average the collective returns for each instance of agent populations in each bin. We use a t-test to compare the *important rule* and *important+silly rule* conditions in each bin. We correct the results with a Bonferonni-correction for 10 multiple comparisons (10 timebins).

For the *important rule* and *important+silly rule* conditions we extracted the mean values for each population of early punishment (mean of the timesteps 0 to 2e8) and subsequent (mean of the timesteps 2e8 to 4e8) time spent poisoned. These two measures where then correlated within each condition.

Note that all statistics are done with the datapoints corresponding to entire populations that each contain 12 agents. This is done because only the data of the entire populations is independent of each other (the agents within one population affect each other, therefore do not produce independent data).

Probe methods. For each agent in each population, the agent's unique neural networks were loaded from 20 evenly spaced timepoints spanning the training run. The agent was then placed in a small empty black environment that contained only one sprite placed in front of the agent (the sprite of a berry or agent). Each episodes episode terminates when the agent interacts with the sprite, or after 30 timesteps (timeout). Valid interactions with sprites are "eating" (upon contact) when the sprite is a berry, and "zapping" with the

*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.

punishment beam when the sprite is an agent. The duration of an episode is our metric for measuring the agent's tendency to interact with the sprite, akin to a "revealed preference" for interacting with a game object. Shorter episode duration indicates a higher preference of the agent to interact with the sprite. Note that the agents do not learn in these episodes. In these probe-episodes, agents are exposed to the sprite of the pink poisonous berry, a green berry that is taboo in the *important+silly rule* condition, four berries that are neither poisonous nor taboo, and the sprite of the "marked" player. The 20 samples per agent from different timepoints during training are probed individually with each sprite. Each probe is repeated 20 times and the duration of all episodes is averaged The results for each timepoint are then averaged across all 12 agents in the population, resulting in 20 datapoints of each population's probe performance over the course of training (15 each for *important rule* and *important+silly rule* and 5 for *no rule*).

Statistical analysis. Mirroring the observational data, we extracted the mean values for each population of early punishment (mean of the timesteps 0 to 2e8) and subsequent (mean of the timesteps 2e8 to 4e8) approach of the poisoned berry for the *important rule* and *important+silly rule* conditions. These two measures where then correlated within each condition.

ACKNOWLEDGMENTS. The authors would like to thank Jane X. Wang for helpful discussion.

1. PJ Richerson, R Boyd, *Not by genes alone: How culture transformed human evolution.* (University of Chicago press), (2008).
2. K Riedl, K Jensen, J Call, M Tomasello, No third-party punishment in chimpanzees. *Proc. Natl. Acad. Sci. United States Am.* **109**, 14824–14829 (2012).
3. M Tomasello, A Vaish, Origins of human cooperation and morality. *Annu. Rev. Psychol.* **64**, 231–255 (2013).
4. JW Buckholtz, R Marois, The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. neuroscience* **15**, 655–661 (2012).
5. DM Fessler, CD Navarrete, Meat is good to taboo. *J. Cogn. Cult.* **3**, 1–40 (2003).
6. J Henrich, N Henrich, The evolution of cultural adaptations: Fijian food taboos protect against dangerous marine toxins. *Proc. Royal Soc. B: Biol. Sci.* **277**, 3715–3724 (2010).
7. MC Corballis, Laterality and myth. *Am. Psychol.* **35**, 284 (1980).
8. D Hadfield-Menell, M Andrus, G Hadfield, Legible normativity for ai alignment: The value of silly rules in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* (ACM), pp. 115–121 (2019).
9. G Brennan, L Eriksson, RE Goodin, N Southwood, *Explaining norms.* (Oxford University Press), (2013).
10. R McElreath, R Boyd, P Richerson, Shared norms and the evolution of ethnic markers. *Curr. anthropology* **44**, 122–130 (2003).
11. R Boyd, PJ Richerson, J Henrich, The cultural niche: Why social learning is essential for human adaptation. *PNAS* **108**, 10918–10925 (2011).
12. B Kenward, Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. *J. experimental child psychology* **112**, 195–207 (2012).
13. GK Hadfield, BR Weingast, Microfoundations of the rule of law. *Annu. Rev. Polit. Sci.* **17**, 21–42 (2014).
14. M Derex, JF Bonnefon, R Boyd, A Mesoudi, Causal understanding is not necessary for the improvement of culturally evolving technology. *Nat. human behaviour* **3**, 446 (2019).
15. R Bond, PB Smith, Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. *Psychol. bulletin* **119**, 111 (1996).
16. J Henrich, R Boyd, The evolution of conformist transmission and the emergence of between-group differences. *Evol. human behavior* **19**, 215–241 (1998).
17. W Hoppitt, KN Laland, *Social learning: an introduction to mechanisms, methods, and models.* (Princeton University Press), (2013).
18. C Tennie, J Call, M Tomasello, Ratcheting up the ratchet: on the evolution of cumulative culture. *Philos. Transactions The Royal Soc. B* **364**, 2405–2415 (2009).
19. GK Hadfield, BR Weingast, Law without the state: legal attributes and the coordination of decentralized collective punishment. *J. Law Court.* **1**, 3–34 (2013).
20. P Richerson, R Boyd, The evolution of human ultra-sociality in *Ideology, Warfare, and Indoctrinability*, eds. I Eibl-Eibisfeldt, F Salter. (Berghan Books, Oxford), pp. 71–95 (1998).
21. J Henrich, M Muthukrishna, The origins and psychology of human cooperation. *Annu. Rev. Psychol.* **72**, 207–240 (2021).
22. JZ Leibo, E Hughes, M Lanctot, T Graepel, Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742* (2019).
23. R Boyd, H Gintis, S Bowles, PJ Richerson, The evolution of altruistic punishment. *Proc. Natl. Acad. Sci.* **100**, 3531–3535 (2003).
24. LC Brooks, W Iba, S Sen, Modeling the emergence and convergence of norms in *Twenty-Second International Joint Conference on Artificial Intelligence.* (2011).
25. S Mahmoud, S Miles, M Luck, Cooperation emergence under resource-constrained peer punishment in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems.* (International Foundation for Autonomous Agents and Multiagent Systems), pp. 900–908 (2016).
26. N Ajmeri, H Guo, PK Murukannaiah, MP Singh, Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. in *IJCAI.* pp. 28–34 (2018).
27. M Kleiman-Weiner, MK Ho, JL Austerweil, ML Littman, JB Tenenbaum, Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction in *CogSci.* (2016).
28. JZ Leibo, V Zambaldi, M Lanctot, J Marecki, T Graepel, Multi-agent reinforcement learning in sequential social dilemmas in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems.* (International Foundation for Autonomous Agents and Multiagent Systems), pp. 464–473 (2017).
29. A Lerer, A Peysakhovich, Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017).
30. A Peysakhovich, A Lerer, Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975* (2017).
31. J Perolat, et al., A multi-agent reinforcement learning model of common-pool resource appropriation in *Advances in Neural Information Processing Systems.* pp. 3643–3652 (2017).
32. E Hughes, et al., Inequity aversion improves cooperation in intertemporal social dilemmas in *Advances in neural information processing systems.* pp. 3326–3336 (2018).
33. J Foerster, et al., Learning with opponent-learning awareness in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* (International Foundation for Autonomous Agents and Multiagent Systems), pp. 122–130 (2018).
34. A Peysakhovich, A Lerer, Prosocial learning agents solve generalized stag hunts better than selfish ones in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems.* (International Foundation for Autonomous Agents and Multiagent Systems), pp. 2043–2044 (2018).
35. JX Wang, et al., Evolving intrinsic motivations for altruistic behavior in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems.* (International Foundation for Autonomous Agents and Multiagent Systems), pp. 683–692 (2019).
36. B Baker, Emergent reciprocity and team formation from randomized uncertain social preferences. *Adv. neural information processing systems (NeurIPS)* (2020).
37. R Boyd, S Mathew, Arbitration supports reciprocity when there are frequent perception errors. *Nat. Hum. Behav.*, 1–8 (2021).
38. Y LeCun, Y Bengio, G Hinton, Deep learning. *nature* **521**, 436–444 (2015).
39. RB Myerson, Refinements of the nash equilibrium concept. *Int. journal game theory* **7**, 73–80 (1978).
40. RS Sutton, AG Barto, A temporal-difference model of classical conditioning in *Proceedings of the ninth annual conference of the cognitive science society.* (Seattle, WA), pp. 355–378 (1987).
41. W Dabney, et al., A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
42. VB Meyer-Rochow, Food taboos: their origins and purposes. *J. Ethnobiol. Ethnomedicine* **5** (2009).
43. E Nisioti, C Moulin-Frier, Grounding artificial intelligence in the origins of human behavior. *arXiv preprint arXiv:2012.08564* (2020).
44. LS Shapley, Stochastic Games. *In Proc. Natl. Acad. Sci. United States Am.* (1953).
45. ML Littman, Markov games as a framework for multi-agent reinforcement learning in *Proceedings of the 11th International Conference on Machine Learning (ICML).* pp. 157–163 (1994).
46. L Espeholt, et al., Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures (2018).
47. Avd Oord, Y Li, O Vinyals, Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
48. M Jaderberg, et al., Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397* (2016).