Spurious normativity enhances 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,g,h], and Joel Z. Leibo[a,1]

[a]DeepMind, London EC4A 3TW, United Kingdom; [b]Computer Science and Artificial Intelligence Laboratory, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Center for Human-Compatible AI, University of California, Berkeley, CA 94720; [d]Faculty of Law, University of Toronto, Toronto, ON M5S 3E6, Canada; [e]Rotman School of Management, University of Toronto, Toronto, ON M5S 3E6, Canada; [f]Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, ON M5G 1L7, Canada; [g]Vector Institute for Artificial Intelligence, Toronto, ON M5G 1M1, Canada; and [h]OpenAI, San Francisco, CA 94110

Edited by Brian Skyrms, University of California, Irvine, CA; received April 13, 2021; accepted October 13, 2021

How do societies learn and maintain social norms? Here we use multiagent 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. 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. 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. Our results highlight the benefit of employing a multiagent reinforcement learning computational model focused on learning to implement complex actions.

multiagent reinforcement learning | norms | third-party punishment | cultural evolution | social norms

One of the central attributes that differentiates human from other animal societies and accounts for the enormous gains of human ultrasociality (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 noninterference 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 ref. 3 for a review).

The normative landscape is also, however, populated by many norms that appear essentially arbitrary and without direct material consequences (5–7). For instance, in some societies, eating particular animals is unacceptable; in others, men are expected to wear pants, not skirts; in many, there are words or hand gestures that should not be used in polite company; and, in most, there are rules about how one styles one's hair or what one wears on one's head. Boyd and Richerson (8) use these examples: "wearing a tie, being kind to animals, or eating the brains of dead relatives." Following Hadfield-Menell et al. (9) we call such norms "silly rules." Silly rules abound in every human society (10). People generally do not experience these rules as silly: They treat compliance with these norms as important, and punish violations. Indeed, silly rules are often imbued with great meaning; religious traditions in most cultures are awash in silly rules. This is what makes them such a puzzle: They are meaningful and enforced but, except for effects generated by this socially- constructed salience, they have no direct or first-order impact on welfare.

A few explanations for the existence of silly rules have been explored in the literature. One explanation posits that silly rules may exist to serve as cheap signals of group membership and thus facilitate cooperation within the group (11). 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 culturally specific and causally opaque knowledge (12). 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 (13–15). 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 overimitate adults, copying—and moralizing—even apparently irrelevant aspects of adult behavior (16). No doubt these explanations account for some aspects of the phenomenon of silly rules. However, 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 for group solidarity or retained as a by-product of the cultural transmission of knowledge. The sheer abundance of silly rules seems to require an account that grants the normative moralization (5) of seemingly irrelevant actions a more fundamental role.

Significance

The fact that humans enforce and comply with norms is an important reason why humans enjoy higher levels of cooperation and welfare than other animals. Some norms are relatively easy to explain: They may prohibit obviously harmful or uncooperative actions. But many norms are not easy to explain. For example, most cultures prohibit eating certain kinds of foods, and almost all societies have rules about what constitutes appropriate clothing, language, and gestures. Using a computational model focused on learning shows that apparently pointless rules can have an indirect effect on welfare. They can help agents learn how to enforce and comply with norms in general, improving the group's ability to enforce norms that have a direct effect on welfare.

Author contributions: R.K., D.H.-M., R.E., L.W., G.K.H., and J.Z.L. designed research; R.K., R.E., L.W., and J.Z.L. performed research; R.K., D.H.-M., R.E., L.W., G.K.H., and J.Z.L. analyzed and interpreted the data, and R.K., D.H.-M., R.E., L.W., G.K.H., and J.Z.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: rkoster@deepmind.com or jzl@deepmind.com.

Published January 12, 2022.

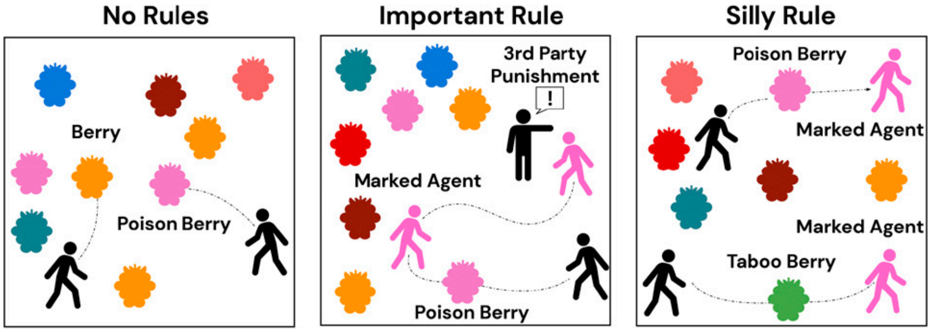

Fig. 1. Schematic overview of the experimental conditions. In the no rules condition, agents collect berries for reward. One berry color is poisonous and, after a time delay, reduces reward obtained from consumed berries. Being poisoned is invisible to all players and a hard credit assignment problem. 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 silly rule condition, the same taboo against the poisonous berry is in place, but, additionally, there is 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.

In this paper, we describe a mechanism through which silly rules can benefit a society. Our argument is based on the dynamics of learning in a group 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 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 can outweigh the dead-weight loss created by the silly rule.

Silly rules can support the emergence and stability of a beneficial normative social order (17). In a normative social order, a stable enforcement mechanism supports an equilibrium in which group behavior is patterned on a classification scheme (called a norm) that partitions behaviors into approved and disapproved (taboo) categories. Here, we employ a computational approach to investigate the effects of silly rules (Fig. 1) on how well a normative social order is learned. Our model consists of a multiagent 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. It can be interpreted as a toxic food with delayed deleterious effects. The delay introduces a difficult 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. Third-party punishment of consuming poisonous berries could greatly simplify the credit assignment problem, by providing a large negative reward closer in time. In this setting, a taboo on the poisonous berry—however it may evolve—raises individual welfare. As Boyd et al. (12) 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 (12, 15). For example, there are long-term negative health effects from consuming maize without previously soaking it in an alkaline solution—a practice that is well established in many cultures dependent on maize, but which is often not understood causally (13, 18).

The mechanisms behind social learning in humans may be multiple: a psychological propensity to conformity (19, 20), deliberate teaching practices (21), and individual trial-and-error learning in conjunction with third-party punishment of failures to follow norms (22). Our work focuses on the last of these. The requisite punishment actions can stand for anything that imposes costs on the receiving agent, including nonphysical punishments such as criticism or mockery (23). 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 generically to recognize when another agent has violated a taboo and to deliver a costly punishment to the violator.

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 berries are taboo). Thus, by varying the content of the norms, we can study the downstream effects on how the normative behaviors (enforcement and

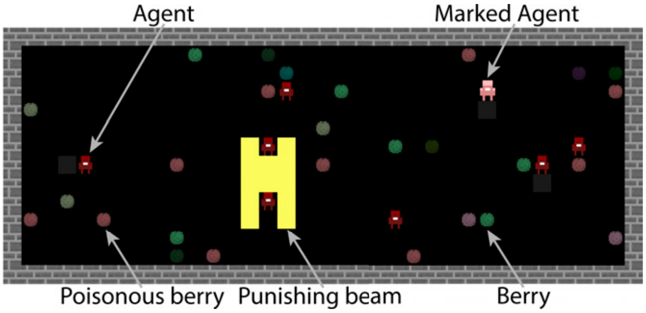

Fig. 2. Depiction of the environment. The agents inhabit a 2D world. Agents earn a reward for eating berries, which regrow probabilistically after being harvested. One type of berry is poisonous, and, if collected by an agent, it diminishes the agent's ability to gather rewards from other berries, after 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 silly rule condition, one additional, nonpoisonous, berry also triggers an agent's marking. Agents are able to punish each other using a "punishing beam," causing a loss to themselves and a larger loss to the punished agent. If a marked agent is punished, the punishing agent receives a significant reward.

compliance) are learned and hence whether and how normative social order is achieved. If a player breaks a taboo, they change color in the observations of other agents viewing their transgression. They become marked. Note that players cannot directly observe their own marking. This reflects that norm violation is in the eye of the beholder. The agent viewing the transgression, not the agent committing it, sees the marking. If a player is marked, other players can collect a reward for punishing them—modeling either intrinsic or extrinsic rewards to third-party punishment. This creates an incentive for players to learn to punish rule violations, and thus for players to learn not to violate the rules. This reflects the common situation across human history in which there is a centralized classification scheme that labels transgressive behavior, but enforcement is decentralized. For example, in medieval Iceland, the "law speaker" was available to publicly label acts as unlawful, and citizen juries at public meetings would declare violators to be "outlaws"—performing the classification function. But enforcement was completely decentralized, with no public enforcers: Citizens were obligated to shun outlaws and were authorized to take of an outlaw's property (or even life) without repercussions (24).

We first 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 set of norms in this regime includes 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 suggest an account of the ubiquity of rules that have no direct impact on well-being by showing how such rules can nonetheless raise group payoffs. They also provide a formalization of normativity in a computational setting that we think will expand the tools available for understanding both 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 (25).

Studying Social Norms with Deep RL

Norms are aggregate social phenomena. They can be studied using a microfoundational approach by investigating mechanisms by which they could emerge from the collective behavior of individuals (17, 26). A wide range of methods exist for doing this. For instance, one can study norms using stochastic evolutionary game theory to capture tipping points and norm-change phenomena (27) or use agent-based models to capture situations where diverse agents have heterogeneous preferences (28). In both classical and evolutionary game theory (29), the reduction is often to a level of description where the elementary actions are abstract choices like cooperate or defect (e.g., refs. 30–33). These choices are treated as atomic, meaning they have no substructure. However, that specific level of analysis is not special. It's possible that reducing to a more fine-grained description of behavior may bring new microfoundational mechanisms into view. For instance, it may be important to study not only what choice agents make but also how they implement that choice (34). Modern methods based on deep RL algorithms (e.g., ref. 35) make this possible (36). By using them, we can construct models of aggregate social phenomena that reduce to substantially more fine-grained descriptions of individual behavior than was possible in the past. This may prove especially important for phenomena related to learning, such as social norms. For instance, learning how to enforce a norm may be as important as learning to enforce it.

Our approach to studying the microfoundations of normativity uses multiagent deep RL, a suite of algorithmic techniques that have been applied in recent work to study spatially and temporally complex (sequential) social dilemmas (37–46). Agents in this framework are artificial neural networks, which learn behavioral policies (associating actions to states) and obtain rewards from an environment. They do not start out with representations for any concepts like "norm," "punishment," "norm violation," or "social approval." In fact, they have no prior knowledge of game rules or states at all. The actions available to them operate only at the most basic level: turn left, turn right, move one step forward, etc. They must learn how their actions, observations (raw pixels), and rewards relate to each other. They do this by learning representations that generalize between similar situations.

The environment our agents inhabit provides the raw materials of a third-party punishment regime: Agents see each other move around the environment and may observe norm violations. Also, at any time, they may take actions that reduce another player's reward (thereby punishing them). Our goal is to investigate whether learning agents, interacting in groups, can discover—on their own—the latent capacities the environment has provided. When successful, agents generate a regime where latent normative infrastructure becomes manifest: Agents enforce norms and comply with norms.

Environment Description. Agents in our experiment inhabit a two-dimensional (2D) world in which they and other objects are located at coordinates in space and only perceive locally around themselves. An agent's action space consists of basic movement (such as moving forward or rotating left and right) and using a "punishing beam" that can be directed at a nearby agent (Fig. 3). Use of the punishing beam costs the punisher 20 points and inflicts a cost of 35 points on the punished agent. A variety of "berries" of different colors are distributed randomly throughout the world. An agent receives a reward of four points 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 time steps after consumption, they reduce reward gained from consuming future berries to one point.

The environment can include a latent classification scheme (17) 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 points instead of a loss of 20 points (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 "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.

Multiagent RL. RL is a formalization of trial-and-error learning in which an agent takes actions according to a policy in response to observations it receives in an environment. The environment is organized into episodes that reset after a certain number of time steps. In particular, we use an "actor–critic" formulation, in

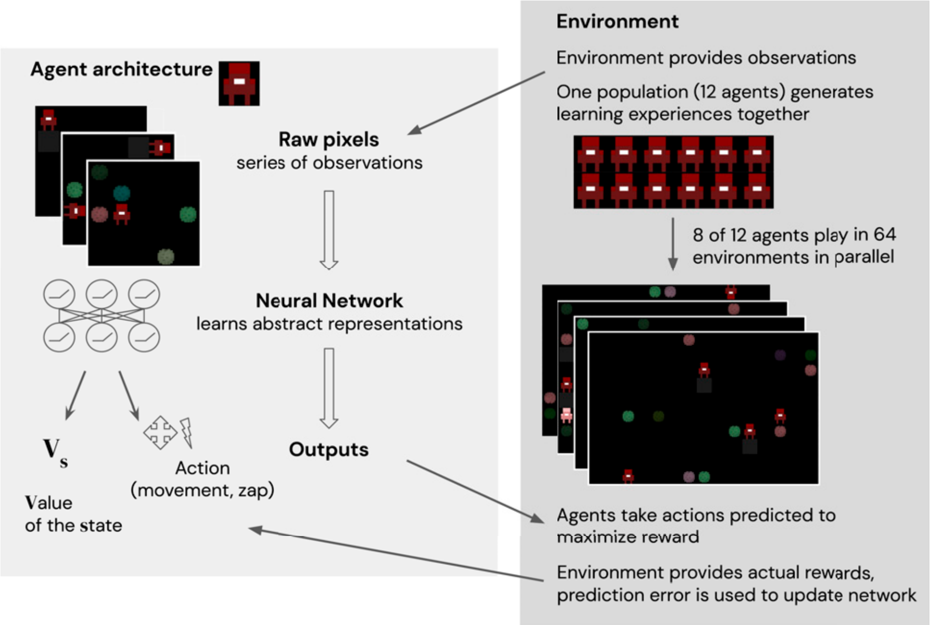
Agent architecture
Raw pixels
series of observations
Neural Network
learns abstract representations
V_s
Value
of the state
Action
(movement, zap)
Outputs
Environment
Environment provides observations
One population (12 agents) generates
learning experiences together
8 of 12 agents play in 64
environments in parallel
Agents take actions predicted to
maximize reward
Environment provides actual rewards,
prediction error is used to update network

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 neural network architecture of each agent learns abstract representations that parse the observations in a way that allows the agent to take reward-maximizing action. The output of neural network on each time step is a prediction of the value of the current state and an action (movement or zap). The expected received reward and actual received reward are compared, and the network weights are gradually adjusted to maximize long-term cumulative reward.

which the policy and the value function that estimates the value of environment states are computed in a (deep) neural network. This use of neural networks is necessary when agents observe raw pixel inputs, as this creates too many states (unique pixel patterns) for tabular RL to be tractable. The same problem would apply also to other tabular approaches such as evolutionary game theory. In deep RL, the neural network transforms the raw pixel input into actions in a series of nonlinear transformations as the agent's observation propagates through the network. The neural network weights in these transformations are gradually tuned to produce rewarding behavior. In our simulation, the agent outputs an action and a value prediction on each time step. The mismatch between the reward prediction and the actual reward achieved in the environment is used to adjust the neural network weights in order to improve the prediction and chosen action in the future.

As displayed in Fig. 3, we consider populations that consist of 12 agents. Each of these agents contains a separate neural network. The networks are identical in architecture, and are initialized with random weights. Each agent's experiences over time—including its observations, actions, and rewards—shape the weights in that agent's neural network. The weights are gradually optimized to choose actions that maximize the agent's reward over a long-term horizon, where rewards in the more distant future are discounted relative to nearer-term rewards. This setup in which each agent learns independently is analogous to noncooperative games in economic models, as opposed to multiagent setups where agents share rewards or use the same neural network weights. These agent experiences are recorded in batches from episodes in which 8 of the 12 agents are selected to play in a shared environment. Each episode consists of 1,000 time steps. On each time step, each agent receives an observation (the pixels in their individual field of view), obtains any reward received on that time step (if any), and issues an action (moving in the environment or zapping). At the start of training, agents' actions are random, but, with growing experience, they are gradually shaped to maximize reward. Importantly, while all agents learn individually, they inhabit a shared environment. Through this coexistence, they influence each other's experiences and learning. For example, one agent learning to effectively punish taboo-breaking behavior may create incentives for other agents to avoid breaking taboos.

Features of the Deep RL Framework. Agents learn continuously while being exposed to episode after episode of interactions in 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) and figuring out how to perform complex action sequences are difficult. The dynamics of how norms are learned and implemented are endogenous to the multiagent 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 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 subactions 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, 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 the enforcement behavior learned for 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, deep RL does not require us to model mistakes in behavior as random noise (e.g., refs. 33 and 47). Instead, mistakes in 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. For each episode, we obtain a metric such as the number of times unmarked agents were punished during the episode (Fig. 4*A*). We also calculate the average number of training steps that agents in the episode have experienced. We plot the average metric obtained across all populations per conditions for a given average of number of training steps for agents in the episode that generated the metric. The plots thus show how average behavior in an episode changes as agents in the episode gain more experience.

The first result to highlight is that learning to avoid the poisonous berry in this environment without the help of a social punishing mechanism is prohibitively hard: Agents in the no rules condition show no decrease in time spent poisoned (Fig. 4*D*). This suggests the credit assignment problem is too hard in this condition. By the time the negative rewards associated with eating a poisonous berry are experienced, an agent has eaten a lot of other berries and had possibly many interactions with other agents, all of which interfere with learning the correct association.

In contrast, in both normative conditions, we see that agents learn to avoid poisonous berries: Total time spent poisoned in these conditions quickly falls below the no rules level (Fig. 4*D*). Effectively, the normative scheme alleviates the credit assignment problem by introducing a negative reward for eating the poisonous berry more immediately in time after consumption. Moreover, Fig. 4*F* shows that it is worth it for a group to devote resources to establishing this normative scheme. Once initial learning is complete, collective return—the sum of total rewards earned by the group net of the costs of punishment—is higher in both the important and silly rule conditions than in the no rules condition.

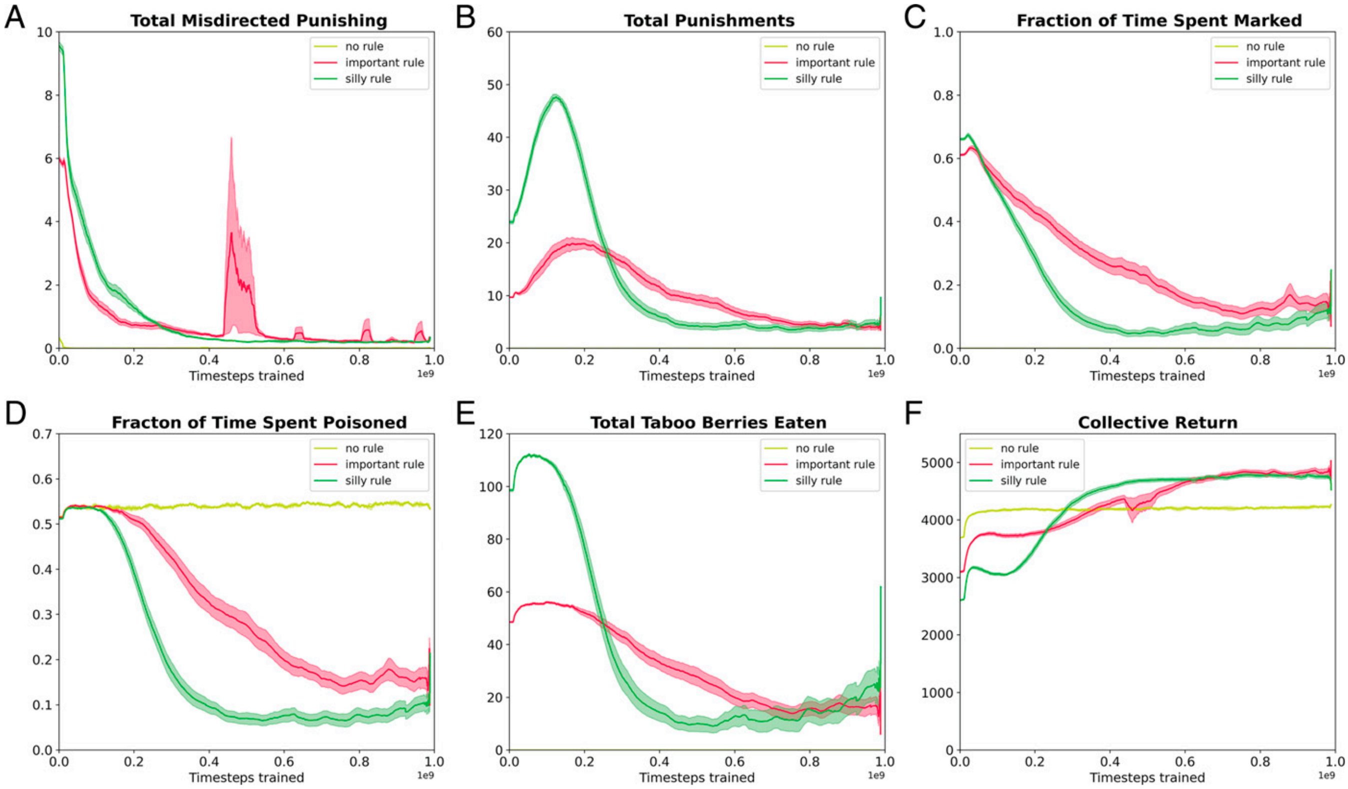

Fig. 4. We are examining per episode group-level metrics about agent populations (*y* axis) over the trajectory of learning (*x* axis in time steps). We plot the average trajectory per condition (and standard error of the mean). (*A*) Number of times unmarked agents are punished (agents that have not broken a taboo). (*B*) Number of times marked agents are punished (agents that have broken a taboo). (*C*) Fraction of time spent marked after breaking a taboo. (*D*) Fraction of time agents spent poisoned. (*E*) The number of "taboo" berries eaten (poisonous and nonpoisonous 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 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, for example, an increase in social punishment that then declines together with a decrease in number of taboo berries eaten.

But, surprisingly, for a significant period, the group's rewards are highest in the silly rule condition—that is, in the ostensibly wasteful setting in which the group devotes resources to punishing the eating of a harmless food. We test the difference in collective returns between the important rule and silly rule conditions in 10 separate time bins. There is a significant benefit of the silly rule in the fourth, fifth, and sixth time bin (fourth: $t((28)) = 3.94$, $p = 0.0005$; fifth: $t((28)) = 3.26$, $p = 0.003$; sixth: $t((28)) = 2.43$, $p = 0.022$; the fourth and fifth time bins remain significant after Bonferonni correction for 10 multiple comparisons).

The reason for the benefit of a silly rule can be seen in the learning dynamics and, in particular, the sequence with which behaviors are learned in the group as a whole. Before the group can enjoy the benefits of a rule that ties immediate negative consequences to the eating of poisonous berries, its agents need to learn to effectively punish rule violations—to recognize when a rule is violated and that punishment of a transgression will earn rewards.

As visible in Fig. 4*A*, 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. 4*F*, this rapid initial learning increases the collective return. Note that the suppression of misdirected punishing happens fastest in the no rules condition. This is expected, 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 we observe is that, in our normative conditions, the number of times marked players get successfully punished initially strongly increases before it decreases (Fig. 4*B*). We interpret the increase as an improvement in the agents' skill at enforcing the social norm, that is, being increasingly skilled at effectively punishing marked agents. We also see, as shown in Fig. 4*C*, that the amount of time agents spend marked is steadily declining. However, taken by itself, this metric does not differentiate between the case in which this decline is driven by agents becoming better at avoiding rule violation and the case in which agents get better at punishing rule breakers and thereby converting marked players back to unmarked players. As can be seen in Fig. 4*E*, 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.

This sequencing—learning to punish effectively before learning to comply with rules—expresses the value of silly rules: Silly rules help a group to more quickly learn to punish, which leads individuals to learn to comply more quickly with the important rule and thereby avoid the toxic food. Early in learning, unsurprisingly there are more taboo berries eaten in the silly rule condition than in the important rule condition (Fig. 4*E*); there are twice as many taboo berries in the environment. As agents learn to punish players who have eaten a taboo berry, this also leads to more punishment in the silly rule condition (Fig. 4*B*). But this higher rate of punishment then apparently leads agents to learn to avoid poisonous berries faster and more effectively as measured by the number of poisonous berries eaten and the time spent marked (Fig. 4 *C* and *D*); 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 total in the condition in which there are twice as many in the environment.

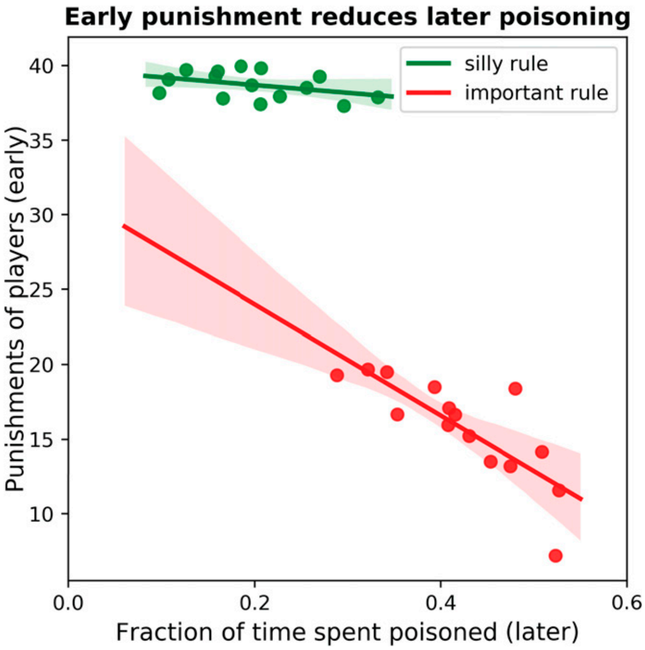

Fig. 5. Higher rates of punishment in early stages are related to less time spent poisoned later in training. Each marker is an independent population of agents. On the *y* axis, we plot how often players are punished in an episode early in training (between time steps 0 and 2e8). On the *x* axis, we plot the mean time players spend poisoned in later episodes (time steps 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 results suggest that more frequent punishment early in learning is associated with less time spent poisoned in subsequent stages of learning (Fig. 4). We directly test this hypothesis by exploiting the variance across different training runs (i.e., separate populations.) For each of the 15 populations that trained in each normative condition, Fig. 5 plots the average amount of punishment delivered in an episode during the first quintile of learning (between time steps 0 and 2e8) on the vertical axis and the total time spent poisoned during the second quintile (between time steps 2e8 and 4e8) on the horizontal axis. In both normative conditions, we find that high rates of punishment in the early stages of learning are related to low amounts of time spent poisoned in subsequent stages (important rule: $r = -0.79$, $p = 0.0004$; silly rule: $r = -0.39$, $p = 0.15$, $n = 15$). Note that the correlation is weaker in the silly rule condition, but that the magnitudes of both measures differ strongly with substantially higher early punishment and lower subsequent poisoning in the silly rule condition. However, like all large-scale observational longitudinal studies with multiple actors, this result could potentially be confounded, since all actions are entangled and interdependent—because agents react to other agents. We address this issue in the following analysis.

Probing What Agents Have Learned. 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 behavior. Psychology experiments with human participants address this issue by isolating specific mechanisms and testing these in controlled conditions, such as 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 laboratory testing to directly probe what the agents

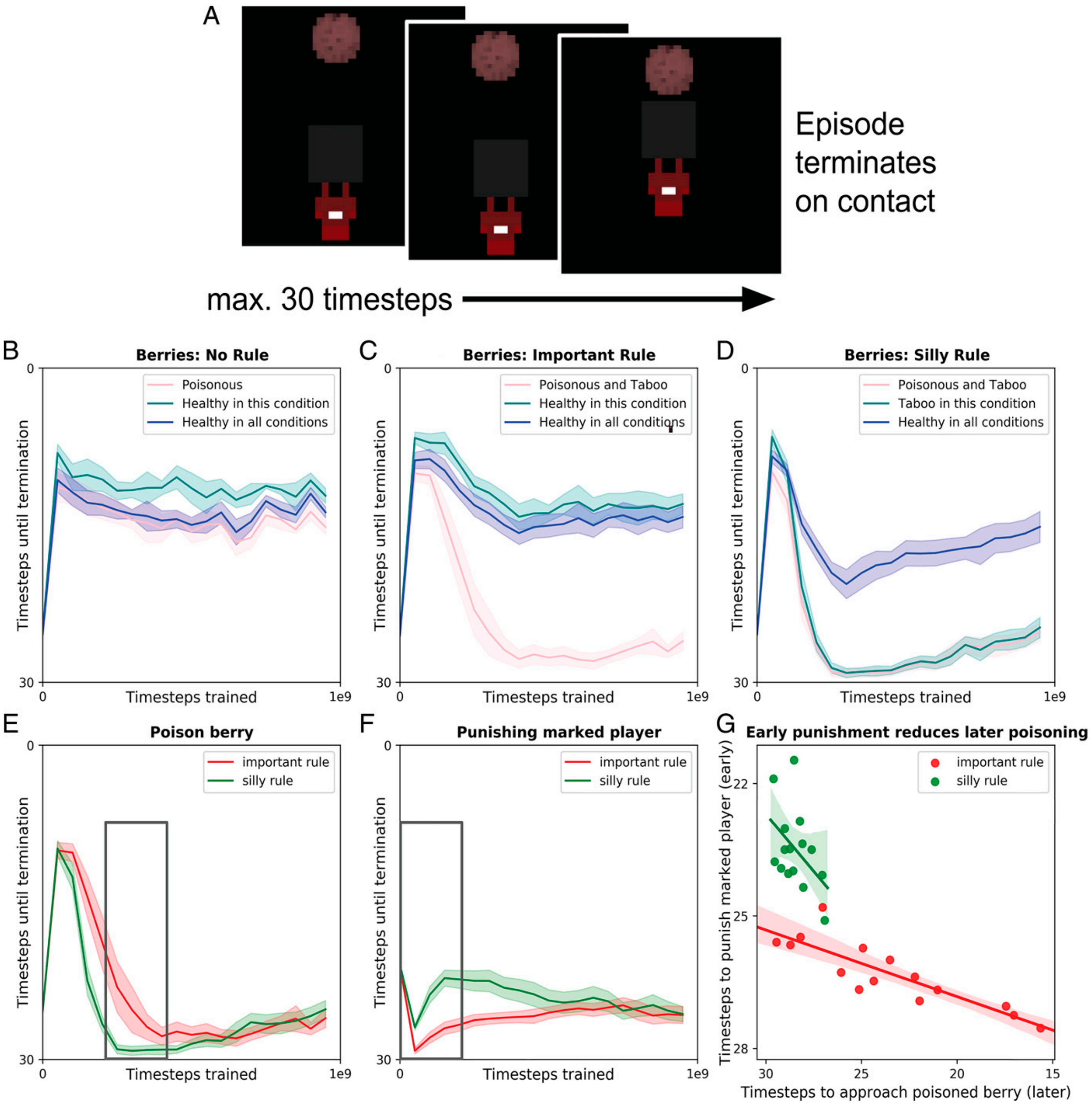

Fig. 6. Single target probes or "zero-shot generalization." (*A*) Depiction of probe. An agent is placed in an empty room with just one other object (berry or agent), and we measure how many time steps it takes before either the agent interacts with the object (eats the berry or zaps the player) or time expires (30 time steps). Note that, in all following plots, the *y* axis is reversed: The origin represents that the maximal amount of time steps passed (30), and the top represents that the minimal amount of time steps passed (zero) until the episode is terminated. (*B*–*D*) Berry types across the three 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* and *F*) Difference between important rule and silly rule for approaching the poisoned berry (same as in *C* and *D*) and punishing the marked player. Agents are faster to learn to avoid the poisoned berry and punish taboos in the 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.

have learned. As shown in Fig. 6*A*, we implement these quasi-laboratory 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 large complex environments to an empty testing environment indicates that they learned robust behavioral responses to game objects.

In each of these experiments, we measure how many time steps out of a maximum of 30 it takes for an agent to move toward different objects. Fig. 6 *B–D* displays how the number of time steps to approach different berries changes over training, when confronted with one berry at a time. Note that the vertical axis begins at 30 and declines to 0, so that a rapid approach is represented by points toward the top of the plot. These approach behaviors vary over the course of training and by the conditions the agents have learned in. As expected (compare Fig. 4*D*), agents learn to avoid the poisonous berry (pink) in the important rule and silly rule conditions but not in no rules condition. Additionally, agents learn to avoid the harmless taboo berry (teal) in the silly rule condition. Fig. 6*E* overlays the poisonous berry lines from Fig. 6 *B* and *D* and illustrates that agents learn to avoid the poisonous berry more in the silly rule condition than in the important rule condition. Similarly, Fig. 6*F* shows the results of placing an agent in an environment with a marked player and measures the time until the agent punishes the marked player. The result illustrates that agents punish marked players more if they trained in the silly rule than if they trained in the important rule condition.

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. 6*G* mirrors the results of Fig. 5, demonstrating that the single-player probes are consistent with the behavior observed in the multiagent setting. We correlate the degree to which a probed agent punishes the marked player during the early stages of training (mean of time window 0 to 2e8 steps, marked in Fig. 6*F*) with that agent's subsequent tendency to consume the poisonous berry (mean of time window 2e8 to 4e8 steps, marked in Fig. 6*E*). In both conditions, we find a negative relationship (important rule: $r = -0.86$, $p < 0.0001$; silly rule: $r = -0.46$, $p = 0.085$, $n = 15$). Again, note that the absolute magnitude of the values differs across conditions; all data points in 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 multiagent 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 rules.

Discussion

This work suggests an 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 group mates 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 credit assignment in trial-and-error learning (42, 48, 49), 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 subbehaviors 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 important rules for which ensuring compliance has first-order effects on welfare. The interpretation of our key result is that the 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 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.

Although we have used the language of physical punishment in referring to the punishment mechanism available to our agents as a "zapping beam," we could also interpret punishment as communicative: conveying the information to an agent that they have violated a social norm. The communicative aspect of punishment is also consistent with our modeling of a norm violation as being information that is available, in a foundational sense, only to third parties: It is socially constructed. Agents learn what is a norm violation (what produces marking) only from the reactions—the messages—of others. If third parties are incentivized to deliver such messages and violators develop the cognitive apparatus that causes such messages to trigger a cost, then the core features of our model are in play. A raised eyebrow, for example, can cause a person to feel shame, and avoiding the experience of shame is the avoidance of punishment—even though the punishment is, in a sense, self-inflicted (50). Such subtle punishment dynamics may underlie cultural learning, as in models that implement a taste for conformity. In our model, inflicting punishment on another is a costly act for a third party to take, and the high cost is important in our setup to ensure punishment is not overincentivized. This cost is felt most when punishing unmarked players, since the cost of punishing marked players is balanced by the intrinsic drive to punish transgression. Thus the cost to the punisher could be seen as reflecting the effect of a not explicitly modeled second-order punishment scheme that discourages the punishing of unmarked individuals.

Our model also resonates with other accounts of how multiple norms relate to one another. The generalization between norms is reminiscent of the concept of cultural "tightness" or "looseness" of societies (51, 52), which explores group-level benefits such as more effective coordinated responses to threats. This formulation describes the strength of social norms and sanctions in a society overall, across a wide range of norms. Similarly, in a simulation by Hadfield-Menell et al. (9), the presence of silly rules allows individuals to draw more accurate conclusions about whether or not rules in general are enforced in a given society. However, our model does not rely on, or speak to, competition between groups. For instance, explanations centered around in-group/out-group classification and group cohesion (14) could operate in addition to the dynamics we describe in our model.

The fact that average group payoffs are higher in simulated environments with silly rules is not, of itself, a basis for concluding that this is why silly rules evolved in human societies. However, this work illuminates what it would take to support that claim. A fully evolutionary account of silly rules requires showing that the group-level benefits which could drive group selection are unlikely to be overwhelmed by individual selection favoring individuals within the group who do not participate in the silly rule behaviors (53). In our approach, punishment and compliance are fundamentally generic behaviors: Agents learn to punish any behavior for which social punishment is rewarded. Group selection in this context would operate not at the level of individual norms but rather at the level of normative infrastructure: the practice of punishing what the group designates as punishable, whatever that might be (8). The benefit of the silly rule in our environment is that it supports the learning of the behaviors that constitute a normative infrastructure, and, once established, that normative infrastructure is capable of enforcing important norms. The individual trait at work is the inclination to punish what the group designates as punishable, and there is no distinction between the silly rule and important rule conditions in this regard. The major boundary from an evolutionary perspective is the boundary between the no rules condition and the normative conditions. This grounds our approach in the central question considered by the literature on cultural evolution of why third-party punishment of group norms has evolved in humans. We have not shown proof that punishment will evolve; we supplied our agents with an exogenously determined and hefty reward for punishing behaviors identified by an exogenous classification. We have only shown that agents can learn more quickly that punishment of certain third-party behaviors is rewarded if there are more rules to generate opportunities to practice. Therefore, this work focuses attention on how those rewards might have arisen among humans. Henrich and Boyd (20) offer one possibility: If there is second- (or higher-) order punishment of agents who fail to punish what the groups deems punishable, a small amount of a conformist tendency to punish what others punish could provide the (intrinsic) reward. This idea was explored in a deep RL model by Vinitsky et al. (54). Our simulation does not model second-order punishment, but we would conjecture that, if an agent who failed to punish another agent seen eating a taboo berry were marked in our environment, then agents would also learn to punish that behavior. Rewards for third-party punishment might also have arisen initially in the context of one individual assisting a victim in retaliating for directly harmful behaviors, and being rewarded by that individual for doing so, or from cooperative efforts to deter uncooperative actions that reduce subgroup payoffs. Boyd et al. (55) present a model in which such cooperative punishment can be evolutionarily stable if cooperative punishers can signal their presence in a group and only punish when there are enough others signaling their willingness to do so. These are possibilities to explore in further elaborations of the deep RL model.

Future work in this vein could seek to synergize deep RL models with cultural evolution models. This could be accomplished by adding an evolutionary algorithm layer using group selection (56) or endowing agents with a cognitive architecture that can dynamically generate particular norms (54). Understanding how cultural abilities can be learned in multiagent settings may play a critical role in understanding the emergence of human-level intelligence (25, 57, 58).

Materials and Methods

Multiagent RL: Model and Notation. Formally, we consider multiagent RL in partially observable general sum Markov games (59, 60). 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 $\mathcal{O} : \mathcal{S} \times \{1, \ldots, N\} \rightarrow \mathbb{R}^d$, 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 $(a^1, \ldots, a^N) \in \mathcal{A}^1 \times \ldots \times \mathcal{A}^N$, the state changes, obeying the stochastic transition function

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

$r^i : \mathcal{S} \times \mathcal{A}^1 \times \ldots \times \mathcal{A}^N \rightarrow \mathbb{R}$ 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 $\pi^i : \mathcal{O}^i \rightarrow \Delta(\mathcal{A}^i)$ (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:

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

Experiment and Conditions. Agents play a foraging task implemented as a partially observable Markov game in a 2D world (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 four points. Each berry type is persistently 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, which remains consistent for all episodes. If collected by a player, this player is "poisoned" after a delay of a fixed number of time steps (100 time steps). Poisoning reduces a player's ability to absorb nutrition: After poisoning sets in, each subsequent berry the player collects yields a reward of one point instead of four points. 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 points (the cost of punishing) and incurs the opportunity cost of time spent aiming and firing the beam instead of collecting berries, and the punished player loses a reward of 35 points.*

Each instance of the training regime for a population of agents is initialized in one of three different conditions. This is a between-subjects design: Each agent population only experiences one of these three 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 rules), the poisonous berry is taboo (important rule), and the poisonous berry and one harmless berry are taboo (silly rule). In no rules, 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 automatically 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 itself. 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 points—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 in a net loss for the group of 20 points because of the cost of using the punishment beam. Aiming punishment at a nonmarked player is costly to both players as in the no rule condition. Once punished, the marking disappears.

In 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, the classification scheme is implemented by the environment. We have not modeled the emergence of the rules themselves. Agents are incentivized to learn policies that implement the

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

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 learning agents, each comprising a separate neural network. Note that this population size is too small to accurately reflect human societies (61). We assume that population-level dynamics would generalize to larger populations, but this has to be tested empirically by future work. The environment is a 2D world of size 33 × 12 pixels, and agents observe a 15 × 15 pixel RGB window, centered on their current location. (Note that the figures in this paper are higher resolution for display purposes.) In each episode, a random subset of eight agents was drawn (without replacement) to play in that episode. Each episode lasted for 1,000 steps. For each time step *s*, each agent *i* in the population produced a policy π^i and an estimate of the value $V^i_{\pi}(s)$ computed by a neural network, implemented on a GPU. The neural networks were trained by receiving importance-weighted policy updates (62) sampled from a queue of trajectories. These trajectories were created by 64 simultaneous environments on CPUs that play the game (with eight 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.

As illustrated in Fig. 3, each agent maintained an independent neural network. This network took raw pixel inputs as observations and output an action and an estimate of the value of the current state. This value estimate was used to calculate whether the actual reward outcome was better or worse than expected. The first layers of the neural network were convolutional layers that learn spatial patterns. They were followed by fully connected layers that can learn more abstract representations of the game. Finally, a recurrent neural network that can maintain activations over multiple time steps learned temporal dependencies.

In detail, the neural network's architecture consisted of a visual encoder (2D convolutional neural net with six channels, with kernel size and stride size one) followed by a two-layer fully connected multilayer perceptron (MLP) with 64 rectified-linear unit (RELU) neurons in each layer, a long short-term memory (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 regularization of the policy logits was 0.003. We used the root mean squared propagation (RMSProp) optimizer (learning rate = 0.0004, epsilon = 1e-5, momentum = 0.0, decay = 0.99). The agent also minimized a contrastive predictive coding (CPC) loss (63) in the manner of an auxiliary objective (64).

Statistical Analysis of Observational Data. In order to assess the difference between conditions, we divide the learning time course 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 silly rule conditions in each bin. We correct the results with a Bonferonni correction for 10 multiple comparisons (10 time bins).

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

Note that all statistics are done with the data points corresponding to entire populations that each contain 12 agents. This is done because only the data of the entire populations are independent of each other (the agents within one population affect each other, and 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 time points 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 episode terminates when the agent interacts with the sprite, or after 30 time steps (timeout). Valid interactions with sprites are "eating" (upon contact) when the sprite is a berry and "zapping" with the 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 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 time points 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 time point are then averaged across all 12 agents in the population, resulting in 20 data points of each population's probe performance over the course of training (15 each for important rule and silly rule and 5 for no rules). Mirroring the observational data, we extracted the mean values for each population of early punishment (mean of the time steps 0 to 2e8) and subsequent (mean of the time steps 2e8 to 4e8) approach of the poisoned berry for the important rule and silly rule conditions. These two measures were then correlated within each condition.

Data Availability. Learning trajectories from the simulation displayed in the plots have been deposited in GitHub (https://github.com/deepmind/spurious_normativity).

ACKNOWLEDGMENTS. We thank Jane X. Wang and Eugene Vinitsky for helpful discussion.

1. P. J. 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. U.S.A.* **109**, 14824–14829 (2012).
3. M. Tomasello, A. Vaish, Origins of human cooperation and morality. *Annu. Rev. Psychol.* **64**, 231–255 (2013).
4. J. W. Buckholtz, R. Marois, The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* **15**, 655–661 (2012).
5. D. M. Fessler, C. D. 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. Biol. Sci.* **277**, 3715–3724 (2010).
7. M. C. Corballis, Laterality and myth. *Am. Psychol.* **35**, 284–295 (1980).
8. R. Boyd, P. J. Richerson, Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
9. 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* (Association for Computing Machinery, 2019), pp. 115–121.
10. G. Brennan, L. Eriksson, R. E. Goodin, N. Southwood, *Explaining Norms* (Oxford University Press, 2013).
11. R. McElreath, R. Boyd, P. Richerson, Shared norms and the evolution of ethnic markers. *Curr. Anthropol.* **44**, 122–130 (2003).
12. R. Boyd, P. J. Richerson, J. Henrich, The cultural niche: Why social learning is essential for human adaptation. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10918–10925 (2011).
13. P. Giuliano, N. Nunn, Understanding cultural persistence and change. *Rev. Econ Stud.* **88**, 1541–1581 (2020).
14. V. B. Meyer-Rochow, Food taboos: Their origins and purposes. *J. Ethnobiol. Ethnomed.* **5**, 18 (2009).
15. M. Derex, J. F. Bonnefon, R. Boyd, A. Mesoudi, Causal understanding is not necessary for the improvement of culturally evolving technology. *Nature* **529**, 484–503 (2016).
16. B. Kenward, Over-imitating preschoolers believe unnecessary actions are normative and enforce their performance by a third party. *J. Exp. Child Psychol.* **112**, 195–207 (2012).
17. G. K. Hadfield, B. R. Weingast, Microfoundations of the rule of law. *Annu. Rev. Polit. Sci.* **17**, 21–42 (2014).
18. S. H. Katz, M. L. Hediger, L. A. Valleroy, Traditional maize processing techniques in the new world. *Science* **184**, 765–773 (1974).
19. R. Bond, P. B. Smith, Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychol. Bull.* **119**, 111 (1996).
20. J. Henrich, R. Boyd, The evolution of conformist transmission and the emergence of between-group differences. *Evol. Hum. Behav.* **19**, 215–241 (1998).
21. W. Hoppitt, K. N. Laland, *Social Learning: An Introduction to Mechanisms, Methods, and Models* (Princeton University Press, 2013).
22. C. Tennie, J. Call, M. Tomasello, Ratcheting up the ratchet: On the evolution of cumulative culture. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2405–2415 (2009).
23. P. Wiessner, Norm enforcement among the Ju/′hoansi bushmen. *Hum. Nat.* **16**, 115–145 (2005).
24. G. K. Hadfield, B. R. Weingast, Law without the state: Legal attributes and the coordination of decentralized collective punishment. *J. Law Court* **1**, 3–34 (2013).
25. J. Z. 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] (2019). https://arxiv.org/abs/1903.00742 (Accessed 25 November 2021).
26. J. S. Coleman, *Foundations of Social Theory* (Harvard University Press, 1994).
27. H. P. Young, The evolution of social norms. *Economics* **7**, 359–387 (2015).
28. F. Bianchi, F. Squazzoni, Agent-based models in sociology. *Wiley Interdiscip. Rev. Comput. Stat.* **7**, 284–306 (2015).
29. J. W. Weibull, *Evolutionary Game Theory* (MIT Press, 1997).
30. R. Boyd, H. Gintis, S. Bowles, P. J. Richerson, The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3531–3535 (2003).
31. J. Henrich, Cultural group selection, coevolutionary processes and large-scale cooperation. *J. Econ. Behav. Organ.* **53**, 3–35 (2004).
32. L. C. Brooks, W. Iba, S. Sen, "Modeling the emergence and convergence of norms" in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, T. Walsh, Ed. (AAAI Press, Menlo Park, CA, 2011), pp. 97–102.

33. R. Boyd, S. Mathew, Arbitration supports reciprocity when there are frequent perception errors. *Nat. Hum. Behav.* **5**, 596–603 (2021).
34. D. Mobbs *et al.*, Promises and challenges of human computational ethology. *Neuron* **109**, 2224–2238 (2021).
35. V. Mnih *et al*., Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
36. K. R. McKee *et al.*, Deep reinforcement learning models the emergent dynamics of human cooperation. arXiv [Preprint] (2021). https://arxiv.org/abs/2103.04982 (Accessed 25 November 2021).
37. M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, J. B. Tenenbaum, "Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction" in *CogSci 2016* (Cognitive Science Society, 2016).
38. J. Z. 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*, K. Larson, M. Winikoff, S. Das, E. H. Durfee, Eds. (Association for Computing Machinery, 2017), pp. 464–473.
39. A. Lerer, A. Peysakhovich, Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv [Preprint] (2017). https://arxiv.org/abs/1707.01068 (Accessed 25 November 2021).
40. A. Peysakhovich, A. Lerer, Consequentialist conditional cooperation in social dilemmas with imperfect information. arXiv [Preprint] (2017). https://arxiv.org/abs/1710.06975 (Accessed 25 November 2021).
41. J. Perolat *et al.*, "A multi-agent reinforcement learning model of common-pool resource appropriation" in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* (Advances in Neural Information Processing Systems, 2017), pp. 3643–3652.
42. E. Hughes *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas" in *Proceedings of the 32st Conference on Neural Information Processing Systems (NIPS 2018)* (Advances in Neural Information Processing Systems, 2018), pp. 3326–3336.
43. J. Foerster *et al.*, "Learning with opponent-learning awareness" in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, M. Dastani, G. Sukthankar, E. André, S. Koenig, Eds. (International Foundation for Autonomous Agents and Multiagent Systems, 2018), pp. 122–130.
44. 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, 2018), pp. 2043–2044.
45. T. Eccles, E. Hughes, J. Kramár, S. Wheelwright, J. Z. Leibo, Learning reciprocity in complex sequential social dilemmas. arXiv [Preprint] (2019). https://arxiv.com/abs/1903.08082 (Accessed 25 November 2021).
46. B. Baker, "Emergent reciprocity and team formation from randomized uncertain social preferences" in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (Advances in Neural Information Processing Systems, 2020).
47. R. B. Myerson, Refinements of the Nash equilibrium concept. *Int. J. Game Theory* **7**, 73–80 (1978).
48. R. S. Sutton, A. G. Barto, "A temporal-difference model of classical conditioning" in *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Seattle, WA, 1987), pp. 355–378.
49. W. Dabney *et al.*, A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
50. D. M. Fessler, "Toward an understanding of the universality of second-order emotions" in *Biocultural Approaches to the Emotions*, A. L. Hinton, Ed. (Cambridge University Press, Cambridge, United Kingdom, 1999), pp. 75–116.
51. M. J. Gelfand, L. H. Nishii, J. L. Raver, On the nature and importance of cultural tightness-looseness. *J. Appl. Psychol.* **91**, 1225–1244 (2006).
52. J. R. Harrington, M. J. Gelfand, Tightness-looseness across the 50 United States. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7990–7995 (2014).
53. J. Henrich, Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses - The Tasmanian case. *Am. Antiq.* **69**, 197–214 (2004).
54. E. Vinitsky *et al.*, A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. arXiv [Preprint] (2021). https://arxiv.com/abs/2106.09012 (Accessed 25 November 2021).
55. R. Boyd, H. Gintis, S. Bowles, Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620 (2010).
56. J. X. Wang *et al.*, "Evolving intrinsic motivations for altruistic behavior" in *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2019), pp. 683–692.
57. E. Nisioti, C. Moulin-Frier, Grounding artificial intelligence in the origins of human behavior. arXiv [Preprint] (2020). https://arxiv.org/abs/2012.08564 (Accessed 25 November 2021).
58. A. Dafoe *et al*., Cooperative AI: Machines must learn to find common ground. *Nature* **593**, 33–36 (2021).
59. L. S. Shapley, Stochastic games. *Proc. Natl. Acad. Sci. U.S.A.* **39**, 1095–1100 (1953).
60. M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning" in *Proceedings of the 11th International Conference on Machine Learning (ICML)* (Morgan Kaufmann, 1994), pp. 157–163.
61. L. C. Aiello, R. I. Dunbar, Neocortex size, group size, and the evolution of language. *Curr. Anthropol.* **34**, 184–193 (1993).
62. L. Espeholt *et al*., Impala: Scalable distributed deep-RL with importance weighted actor-learner architectures. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.01561 (Accessed 25 November 2021).
63. A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding. arXiv [Preprint] (2018). https://arxiv.org/abs/1807.03748 (Accessed 25 November 2021).
64. M. Jaderberg *et al.*, Reinforcement learning with unsupervised auxiliary tasks. arXiv [Preprint] (2016). https://arxiv.org/abs/1611.05397 (Accessed 25 November 2021).