




A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings

Collective Intelligence
Volume 2:2: 1–14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: [10.1177/26339137231162025](https://doi.org/10.1177/26339137231162025)
journals.sagepub.com/home/col


Eugene Vinitsky 
UC Berkeley, Berkeley, CA, USA

**Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets and
Joel Z Leibo** 
Deepmind, London, UK

Abstract

Society is characterized by the presence of a variety of social norms: collective patterns of sanctioning that can prevent miscoordination and free-riding. Inspired by this, we aim to construct learning dynamics where potentially beneficial social norms can emerge. Since social norms are underpinned by sanctioning, we introduce a training regime where agents can access all sanctioning events but learning is otherwise decentralized. This setting is technologically interesting because sanctioning events may be the only available public signal in decentralized multi-agent systems where reward or policy-sharing is infeasible or undesirable. To achieve collective action in this setting, we construct an agent architecture containing a classifier module that categorizes observed behaviors as approved or disapproved, and a motivation to punish in accord with the group. We show that social norms emerge in multi-agent systems containing this agent and investigate the conditions under which this helps them achieve socially beneficial outcomes.

Keywords

Multi-agent systems, social norms, reinforcement learning

Corresponding author:

Eugene Vinitsky, UC Berkeley, 652 Sutardja Dai Hall, Berkeley, CA 94720, USA. Email: vinitsky.eugene@gmail.com



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Significance Statement

It is difficult to achieve collective action when groups of individuals have conflicting interests and complex coordination is required. Humans often resolve real-world problems of this kind using social norms, that is, collective patterns of sanctioning underpinned by a classification of behavior into approved versus disapproved categories. We study a setting where agents are collectively aware of which behaviors are actively sanctioned by the group. The reinforcement learning agent we construct has two important features. First, it tracks the group's assessment of which behaviors are approved versus disapproved by classifying which behaviors are or are not sanctioned. Second, it is motivated to sanction others in accord with the group's evaluation. We see that social norms quickly emerge in these simulations through a "bandwagon effect" wherein initially weak patterns of sanctioning become magnified through reproduction. This is an interesting platform to study the emergence of social norms *in silico* because it allows the content of the norm to emerge in parallel with the learning of enforcement and compliance behaviors.

Autonomously operating learning agents are becoming more common and this trend is likely to continue accelerating for a variety of reasons. First, cheap sensors, actuators, and high-speed wireless internet have drastically lowered the barrier to deploy an autonomous system. Second, autonomy creates the possibility of learning "on device," keeping experience local and off of any central servers. This makes it easier to comply with privacy requirements (Kairouz et al., 2019) and increases robustness by removing a single point of failure. Third, the autonomous approach is a potentially better fit for never-ending life-long learning (Platanios et al., 2020) since it does not require periodic syncing with updated centralized models. Indeed fully autonomous agents do not require any train-test separation at all, a property thought to be important for establishing open-ended autocurricula (Leibo et al., 2019; Stanley, 2019).

However, the presence of multiple interacting autonomous systems raises a host of new challenges. Autonomously operating learning agents must be robust to the presence of other learning agents in their environment (e.g. (Carroll et al., 2019; Crandall et al., 2018)). A significant issue that arises in the case of autonomous and decentralized learning agents is how to align their incentives. Working together is often difficult when agents all may prefer to maximize their own rewards at one another's expense. For instance, autonomous vehicles from multiple competing technology companies must share the road with one another and with human drivers (e.g. Liang et al., 2019). Each car (company) "wants" to "selfishly" transport its users as quickly as possible. However, road congestion emerging from poor coordination negatively affects everyone. Human users also participate in these multi-agent systems, with even more autonomy. For instance, city neighborhoods compete with each other to reshape their roadways to incentivize driving apps to route traffic to other neighborhoods (Çolak et al., 2016). Fundamentally, in collective action problems, letting agents egoistically optimize their

own reward leads to a worse outcome for everyone than if all cooperate. This problem is particularly difficult if different ways to cooperate exist and agents have divergent preferences over the outcomes. In this case, *uncoordinated* cooperation may be no better than mutual defection. In these cases, it is difficult for a consensus to emerge.

To address such social dilemmas, we take inspiration from a mechanism that human societies use to resolve some of the collective action problems they face: *social norms*—group behavior patterns that are underpinned by decentralized social sanctioning (approval and disapproval: equivalently, reward and punishment) (Balafoutas et al., 2014; Fehr and Fischbacher, 2004; Wiessner, 2005). Social norms enable cooperative behavior in a wide variety of collective action problems which otherwise would fail due to free-riding and defection. Human civilization is thick with social norms (Henrich and Muthukrishna, 2021; Tomasello and Vaish, 2013; Young, 2015). They are critical to our welfare because they discourage harmful behaviors (e.g. smoking in public places) and encourage beneficial behaviors (e.g. charitable donation and voting) (Bicchieri, 2016; Nyborg et al., 2016). Social norms are also important components in institutional solutions to small community scale natural resource management problems (Hadfield and Weingast, 2013; Ostrom, 2009) and aid large-scale collective actions like labor negotiations and democratic elections (Granovetter, 1978; Marwell and Oliver, 1993; Olson, 1965; Ostrom, 1998).

The critical assumption that will enable our agents to learn social norms by decentralized multi-agent reinforcement learning is that of *public sanctioning*. In this paradigm, there are discrete events when agent i makes their disapproval of agent j known, an event that is typically punishing to the recipient in the sense of reinforcement learning. These events are considered to be public so learning may be conditioned on knowledge of all sanctioning events from any agent to any other agent. This paradigm has several positive features. For instance, it allows for the possibility of

human participants sanctioning autonomous machines through the same ‘API’ that the machines use to sanction one another. For instance, human drivers and self-driving cars could honk at each other or leave 1-star reviews. As sanctions occur and are stored, databases of sanctioning events could enable agents to adapt to local customs like differing driving patterns between cities.

We construct an agent architecture that can use public sanctions to spark the emergence of social norms in a multi-agent reinforcement learning system. Our approach, which we call *Classifier Norm Model (CNM)*, takes inspiration from a specific account of norms by [Hadfield and Weingast \(2012\)](#). They argue that social norms divide behavior into approved and disapproved categories. That is, they are classifiers ([Hadfield and Weingast, 2014](#)). In their account norms are community-based evaluations of behavior; however, they do not necessarily envision a centralized classifier. Each agent may have its own private representation of the group’s schema for what constitutes approved behavior as long as they tend to agree in their classifications. In our model, agents view other actors in the scene and generate a prediction for whether society at large would approve or disapprove of their behavior. The other critical ingredient in the Hadfield and Weingast account of norms is decentralized collective punishment ([Hadfield and Weingast, 2013](#)). To capture this we endow CNM learning agents with an intrinsic motivation to disapprove of behaviors that their group disapproves of, as humans do ([Boehm, 2012](#); [Fehr and Fischbacher, 2004](#); [Xiao and Houser, 2005](#)).

We show that CNM magnifies emergent joint activity patterns that arise by chance in early exploratory learning. This “bandwagon” effect simultaneously pushes agents to cooperate and encourages them to cooperate in the same way as one another. Thus, it mitigates the two fundamental dilemmas within each collective action problem: the start-up and free-rider problems (terminology from ([Marwell and Oliver, 1993](#))). In two complex collective action problems, we show that groups of CNM agents acquire beneficial social norms that decrease free-riding and coordinate cooperative actions, thereby causing higher per-agent returns. Next, we consider our results in light of arbitrariness properties of real-world social norms. That is, specific norms are not always beneficial relative to counterfactual situations where other norms prevail (different ways of cooperating) ([Bicchieri, 2016](#); [Ostrom, 2009](#)). This is a key property of real-world norms and our model also captures it. Finally, we analyze the CNM agent architecture with ablation experiments to understand which architectural assumptions are key to our results.

Related work

Significant progress in multi-agent reinforcement learning has occurred over the last few years driven by rapid

innovation in a paradigm where researchers assume that even though policies must ultimately be executed in a decentralized manner (without communication at run time), they can be trained offline beforehand in a centralized fashion. This paradigm is called centralized training with decentralized execution (CTDE) ([Baker, 2020](#); [Foerster et al., 2018a](#); [Iqbal and Sha, 2019](#); [Lowe et al., 2017](#); [Sunehag et al., 2018](#); [Rashid et al., 2018](#)). Many algorithms in this class ([Baker, 2020](#); [Lowe et al., 2017](#)) take an actor-critic approach and employ a centralized critic that takes in observations from all agents to produce a single joint value. One algorithm called OPRE maintains the division between training and test phases but does not learn a centralized critic. Instead in OPRE each agent learns its own critic but all critics are conditioned on the observations of the other players. This is interpreted as information available in “hindsight” ([Vezhnevets et al., 2020](#)). Other techniques make extensive use of the centralized regime by expanding and pruning the support of policies in each rollout; this includes algorithms like PSRO ([Lanctot et al., 2017](#)) and XDO ([McAleer et al., 2021](#)).

A rather different class of models takes the approach of constraining the *kind* of information that can be communicated between agents, instead of constraining the time (training time versus test time) of its communication. These models avoid the need for explicit training and testing phases. They can be executed online and maintain full decentralization except for the specific data they need to communicate. Some researchers have studied the case where no information at all is communicated between agents. However this approach cannot usually resolve social dilemmas or coordinate on beneficial equilibria when multiple equilibria exist unless special environmental circumstances prevail ([Köster et al., 2020](#); [Leibo et al., 2017](#); [Pérolat et al., 2017](#)). A few algorithms eschew training/testing but still cannot be considered fully decentralized since they require each player to be able to access the policies of other players ([Foerster et al., 2018b](#); [Jaques et al., 2019](#)). Most algorithms in this class that can robustly find socially beneficial equilibria in collective action problems require public rewards ([Eccles et al., 2019](#); [Gemp et al., 2020](#); [Hughes et al., 2018](#); [McKee et al., 2020](#); [Peysakhovich and Lerer, 2018](#); [Wang et al., 2018](#)) or the ability to redistribute rewards amongst agents’ ([Lupu and Precup, 2020](#); [Wang et al., 2021](#)). This class of algorithms assumes that while they are learning all agents will have real-time access to one another’s rewards.

However, making reward data public is undesirable for several reasons. (A) Agent designers may want to alter reward functions without affecting the larger multi-agent system. (B) Agent designers may be prohibited from sharing their agents’ reward function on privacy grounds, for instance, if they constructed it from individual user data ([Kairouz et al., 2019](#)), or their reward functions may be

proprietary. (C) Humans may inhabit the same multi-agent system as artificial agents. This is most apparent in autonomous vehicle applications. Humans cannot publicize their instantaneous reward signals, but both human-driven and self-driven cars can honk their horn to admonish others for bad driving.

In the real world, social norms need not be beneficial. For example they may ossify inefficient economic systems or unfairly discriminate against classes of people (Akerlof, 1976; Bicchieri, 2016; Mackie, 1996). In other cases, social norms can be “silly rules” that are neither directly harmful nor helpful (Hadfield-Menell et al., 2019; Köster et al., 2022). Yet some social norms are clearly helpful, like those that discourage harmful behavior. There are two main mechanisms through which beneficial social norms function: (A) stabilizing cooperation in social dilemma situations as the sanctioning can transform the payoffs into a game with new equilibria (Kelley et al., 2003; Ullmann-Margalit, 1977-2015) and (B) equilibrium selection. Here the question is how it can be predicted which equilibrium a society will select, given that multiple equilibria exist for the social situation in question (e.g. (Lewis, 1969)). In this case, the norm is a piece of public knowledge on which individuals may condition their behavior to rationally coordinate their actions with one another (Gintis, 2010; Vanderschraaf, 1995). Naturally, these two functions are often intertwined (e.g. (Bicchieri, 2006; Hadfield and Weingast, 2012)). In this spirit, social norms have been treated in AI research as equilibria of repeated normal form games (Sen and Airiau, 1507; Shoham and Tennenholtz, 1997).

Recent work has aimed to study social norms in more complex models of human societies. One line of research has represented social norms with classifiers that label a behavior’s social approval or disapproval. For instance, (Boyd and Mathew, 2021) studied how such a classifier can interact positively with a reputation-based account of cooperation in iterated matrix games and (Köster et al., 2022) demonstrated the potential benefits of a “hand-crafted” (ie not learned) classifier on the learning dynamics of enforcement and compliance behavior in multi-agent reinforcement learning.

Multi-agent reinforcement learning with sanctions

The formal setting for multi-agent reinforcement learning with sanctions is an N -player partially observed general-sum Markov game (e.g. (Littman, 1994; Shapley, 1953)) augmented with a concept of sanctioning and a public observation function that indicates when a player has sanctioned another player and with what valence (approval or disapproval).

Definition: Markov game

At each state $s \in \mathcal{S}$ of a Markov game, each player $i \in I = \{1, \dots, N\}$ takes an action $a_i \in \mathcal{A}_i$. Players cannot perceive each state directly, but instead receive their own d -dimensional partial observation of the state $o_i \in \mathbb{R}^d$, which is determined by the observation function $\mathcal{O}: \mathcal{S} \times I \rightarrow \mathbb{R}^d$. After the players’ joint action $\vec{a} = (a_1, \dots, a_N)$, the state changes according to the stochastic transition function $\mathcal{T}: \mathcal{S} \times \mathcal{A}_1 \times \dots \times \mathcal{A}_N \rightarrow \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of discrete probability distributions over \mathcal{S} . After each transition, each player i receives a reward $r_i \in \mathbb{R}$ according to the reward function $\mathcal{R}: \mathcal{S} \times \mathcal{A}_1 \times \dots \times \mathcal{A}_N \times \mathcal{S} \times I \rightarrow \mathbb{R}$.

We extend this standard definition to include the additional concept of *sanctioning*. Sanctioning is assumed to be something that one player does to another player (it is dyadic). All players are assumed to have common knowledge of which events are sanctioning events and their valence (whether they are approval or disapproval).

Definition: Markov game with sanctions

We define a *sanctioning opportunity* as a situation where one agent can sanction another agent by taking an action that causes them a reward or punishment. The reward implications may be indirect. Sanctioning may not produce any instantaneous reward. For instance, an action may be punishing if it causes its recipient’s future rewards to be less probable or delayed. There may be many different ways for agents to cause each other reward and punishment. Not all actions that cause reward or punishment are sanctioning actions. The Markov game with sanctions model stipulates that certain specific events are sanctioning events. It assumes all the agents have common knowledge of which events are sanctioning events.

If agent i has an opportunity to sanction agent j and chooses to punish them with its next action we call this a *disapproval event*. If agent i has a sanctioning opportunity but does not choose to punish agent j with its next action we call this an *approval event*¹. Sanctioning opportunities are often situations where agent i and agent j are physically near one another, but in general they need not be. For instance, a user of a decentralized restaurant recommendation platform may leave a 1-star review to show their disapproval of a restaurant they visited several days prior.

Formally, for any given state $s \in \mathcal{S}$, let the set of sanctioning opportunities be given by $\mathcal{J}(s) \subseteq I^2$, where $(i, j) \in \mathcal{J}(s)$ whenever agent i has a sanctioning opportunity towards agent j . Note that $\mathcal{J}(s)$ may be empty if no agent has a sanctioning opportunity in state s , and at the other extreme $\mathcal{J}(s) = I^2$ when every agent can sanction every other agent (including themselves).

In this work, agents show their disapproval by emitting a zapping beam that has a punishing effect on any agent hit by

it. A sanctioning opportunity (i, j) therefore exists only if agent i is physically in range to zap agent j .

Definition: Markov game with public sanctions

A Markov game with public sanctions is a Markov game with sanctions that has been additionally augmented with a *sanctioning observation* that is shared by all players. At each state, in addition to their individual observation o_i , each player i also receives a sanctioning observation $g \in \mathcal{G}$, defined by the sanction-observation function $\mathcal{B}: \mathcal{S} \rightarrow \mathcal{G}$. This observation broadcasts information on the occurrence of sanctioning to all players.

It is natural to regard the public sanctioning observation as arising from a process of gossip whereby knowledge of who transgressed rapidly diffuses through a community. This interpretation may be useful for research that applies the Markov game with public sanctions model to study social-behavioral phenomena. On the other hand, when we think of modern technology like autonomous vehicles through this lens then we usually envision the public sanctioning observation as a kind of database to which all cars may read and write.

Let $\mathcal{C}(s, i, j)$ be the *context* of sanctioning opportunity $(i, j) \in \mathcal{J}(s)$ —the perspective of the decision-making agent leading up to its choice to approve/disapprove. In general, $\mathcal{C}(s_t, i, j) = (o_{0:t}^{(i)}, a_{0:t-1}^{(i)})$, the full history of the decision-making agent’s individual observations and actions; however, it is also possible to use less context. For instance, in the environments we study here, agents change color as a function of their recent behavior. Thus it is sufficient to choose $\mathcal{C}(s_t, i, j) = o_t^{(i)}$, the current observation of the agent with the sanctioning opportunity. For example, think about a child stealing a cookie. If when you encounter them, they still have chocolate all over their face then you need not have directly observed their transgression to disapprove of their behavior.

Finally, let $\mathcal{Z}(s, \vec{a}, i, j) \in \{0, 1\}$ be a binary indicator of whether the actions \vec{a} taken in state s resulted in a *disapproval* event (of j by i). In this work, we define $\mathcal{Z}(s, \vec{a}, i, j) = 1$ if agent i zaps agent j .

Putting everything together, we get a sanction-observation function that, at time t , returns a view of the sanctioning opportunities at time $t - 1$, the sanctioning decisions made at those opportunities, and the context for those decisions

$$\mathcal{B}(s_{t-1}, \vec{a}_{t-1}) = \{(i, j, c, z) \text{ such that} \\ (i, j) \in \mathcal{J}(s_{t-1}) \text{ and} \\ c = \mathcal{C}(s_{t-1}, i, j) \text{ and} \\ z = \mathcal{Z}(s_{t-1}, \vec{a}_{t-1}, i, j)\}$$

Note that this depends on the previous state s_{t-1} and actions taken \vec{a}_{t-1} , but it can still be represented as $\mathcal{B}(s_t)$ by augmenting the state to include prior observations or actions.

Interpretation of the definitions

To build intuition for what constitutes sanctioning, consider a human driving along the highway. We assume that humans dislike having a car horn honked at them. This attitude may only partly depend on the intrinsically aversive nature of the honking sound itself. Most of the negative experience of being honked at derives from understanding the sound’s cultural context. Drivers honk when they want to admonish other drivers for their bad behavior. Thus being honked at may be aversive through a guilt mechanism (“I am sorry I transgressed”) or through an anger/reciprocity mechanism (“how dare you say I transgressed!”). No matter the cause, the important thing is common knowledge on the part of the whole driving community that honking is meant to be admonishing.

Of course drivers do not always honk to sanction one another. For instance, they also honk to alert one another of danger. There is plenty of scope for disagreement concerning whether a given honk was intended as sanctioning or alerting. In this, sanctioning is no different from any other form of communication where ambiguity is pervasive but humans are nevertheless able to recover their partner’s intent. In the case of honking it is usually obvious from context that a given honk was intended as sanctioning. Sometimes, if worried the current context may not make their meaning clear, individuals may seek to resolve ambiguity by adding an extra “flourish” to their honk such as a rude gesture. However, for the driver who was honked at to feel punished, it is not always necessary for the driver who honked at them to have intended to sanction them. The driver who was honked at, even if it was just to alert them, may still feel punished by the interaction. The critical point is that the overall pattern of honking exerts its influence on collective driving behavior via its inducement of individuals to change how they drive.

As you drive along, any time another driver is in hearing range of your horn constitutes a *sanctioning opportunity*; you have an opportunity to honk your horn at a nearby driver and either chooses to do so or not to do so. Each time you honk the horn, this constitutes a *disapproval event* and each time-step when you do not honk is an *approval event*. The *context* of the sanctioning opportunity could be the current time at the point of sanctioning, or it could also optionally include some number of time-steps that preceded the sanctioning opportunity. While the sanction opportunities only occur if agents are within hearing distance, the sanction-observation function \mathcal{B} can be either local or global. In the local case, an agent is only aware of a sanction opportunity and its outcome if it physically observed/experienced it. In the global case, we can imagine that \mathcal{B} is streamed to a database and available to all agents. As an instantiation, one could image a dash-cam and microphone streaming every sanctioning opportunity and approval/disapproval to a database that would be accessible to all drivers and agents. This latter variant, in which all sanctioning opportunities and outcomes in an episode are

available to all agents, is the main setting we consider in this work.

Learning to classify transgression

In this work we are concerned with developing a multi-agent simulation model where social norms emerge as the system self-organizes by learning. As such, the things the agents do in their world do not have any objective normative status. The classification of whether or not a given behavior constitutes a transgression is determined entirely by whether the group has sanctioned similar behavior in the past.

Each *Classifier Norm Model (CNM)* agent has its own representation for what it thinks the group would sanction—that is, a classifier that predicts whether the group would approve or disapprove of any given behavior. We train each individual’s classifier on the public sanctioning observations provided by $\mathcal{B}(s_{0:T})$. Given a classifier Ψ_ϕ that outputs probabilities of sanctioning and assuming the set of sanctioning opportunities is of size M , we form a binary cross-entropy loss

$$\mathcal{L}_\phi = \frac{1}{M} \sum_{c, z \in \mathcal{B}} -z \log(\Psi_\phi(c)) - (1-z) \log(1 - \Psi_\phi(c))$$

and minimize it with stochastic gradient descent.

There are some potential challenges with learning this classifier. One key issue arises because the classification is learned from the stream produced by an ongoing simulation. The data distribution may not be stationary. For example, when a particular behavior becomes effectively suppressed, perhaps because it was being punished so all agents learned to stop doing it, then the classifier will no longer receive training samples of it being approved or disapproved. This shift in the data distribution violates a stationarity assumption underpinning the classifier’s training procedure and as a result, may cause *catastrophic forgetting* (McClelland et al., 1995), a phenomenon where a neural network unlearns its prior pattern of behavior. To avoid this problem, we stop the classifier from continuing to learn after some fixed number of time-steps by setting its learning rate to zero. This freezes at that point in time each agents’ representation of how context determines whether one has or has not transgressed, but it does not prevent subsequent drift in their sanctioning behavior or compliance behavior.

Learning how to enforce and comply

The core idea of the **CNM** agent is that an individual embedded in a wider group is motivated to sanction in accord with the group’s joint pattern of approval and disapproval. This shapes the group’s behavior because disapproval is punishing.

The motivation to sanction consistently with the group is created by a pseudoreward term in the agent’s reward function (ie an intrinsic motivation in the sense of (Singh et al., 2004)) that encourages each reinforcement learning agent to disapprove in contexts that their classifier assesses as likely to provoke disapproval from others in the group

$$\Omega_\phi(o_t, a_t) = \begin{cases} +\alpha & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) \geq 0.5 \\ -\beta & \text{if } a_t \text{ is disapproval} \wedge \Psi_\phi(o_t) < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

for $\alpha, \beta \in \mathbb{R}_0^+$.

A **CNM** agent learns its classifier while simultaneously learning to maximize reward augmented by this intrinsic motivation to align its sanctioning with that of its group. It learns by applying a decentralized multi-agent reinforcement learning algorithm. Achieving high intrinsic reward demands the agent learn an efficient enforcement policy that sanctions like the wider group. Achieving high extrinsic reward demands the agent learn an efficient compliance policy that avoids provoking disapproval from others.

Each agent i learns a parameterized behavior policy that is conditioned solely on the history of its own individual observations and actions and its estimate of the collective sanctioning pattern $\pi_\theta(a_t^{(i)} | o_{0:t}^{(i)}, a_{0:t-1}^{(i)}, p_t)$ where $p_t = \text{stop}(\mathbb{1}[\Psi_\phi(o_t) \geq 0.5])$ and $\text{stop}(\cdot)$ is the stop gradient operator.

Both classifier and policy consist of a convolutional backbone attached to a multi-layer perceptron (MLP). The classifier MLP directly outputs the predictions whereas the policy MLP feeds into a recurrent network (an LSTM (Hochreiter and Schmidhuber, 1997)) whose outputs are the action probabilities. The classifier network takes the prior frame to make its prediction (context length is one, see Section 2), whereas the policy takes the current frame to get an action. The classifier and policy do not share any layers in this architecture. The overall architecture, including the manner in which predictions are passed to the policy and the pseudoreward computation, is illustrated in Figure 1.

Each agent’s policy is implemented using a private neural network, with no parameter sharing between agents. Each agent’s policy parameters are independently trained to maximize the policy’s long-term γ -discounted payoff

$$V_{(\theta, \phi)}^{\vec{\pi}}(s_0) = \mathbb{E}_{\vec{\pi}_{t,T}} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, \vec{a}_{t,s_{t+1}}) + \gamma^t \Omega_\phi(o_{t-1}^{(i)}, a_{t-1}^{(i)}) \right]$$

where the pseudoreward term shapes sanctioning behavior towards coherence with the group’s pattern of approval and disapproval. We train on episodes sampled from $\vec{\pi}$. All agents control exactly one player in every episode.

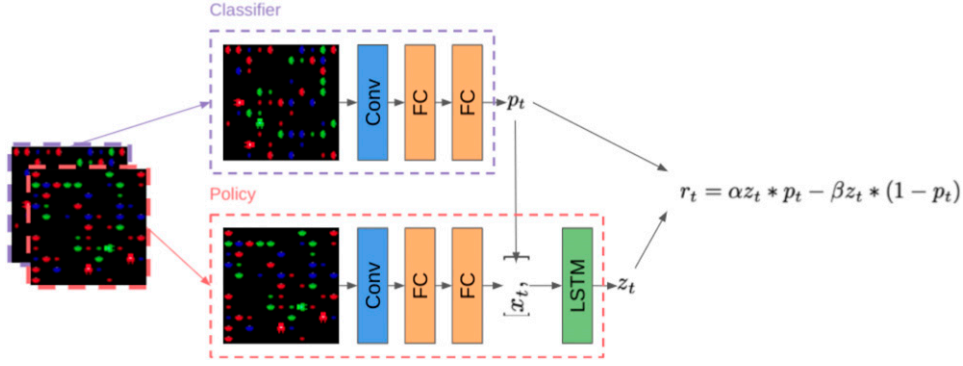


Figure 1. Visual depiction of how the classification is done and how the pseudoreward for aligning with the classifier is generated. The frame at which disapproval occurs and the frame before are stacked together; the frame before the disapproval is fed into the classifier to generate a prediction. If the agent chooses to disapprove, then a reward or penalty is generated based on whether its choice aligns with its classifier prediction.

The reinforcement learning algorithm used for each agent is A3C (Mnih et al., 2016) with a V-Trace loss for computing the advantage (Espeholt et al., 2019). To the standard A3C loss, we add a contrastive predictive coding loss (A.v.d et al., 2018) in the manner of an auxiliary objective (Jaderberg et al., 2017), which promotes discrimination between nearby timepoints via LSTM state representations. For more details, please refer to the Appendix.

Environments

We study two complex collective action problems implemented in Melting Pot (Leibo et al., 2021). The two games are depicted in Figure 2, *Allelopathic Harvest* (AH)² and *Clean Up with Startup Problem* (CSP). Both games have the flavor of bargaining problems in the sense that several different Pareto-optimal outcomes are possible but individuals’ preferences over said outcomes conflict with one another. Both games contain several different equilibria, each associated with a distinct type of “work” and superior to other uncoordinated equilibria. Thus, both games contain start-up and free-rider sub-problems (terminology from (Marwell and Oliver, 1993)). This means that in order to achieve high rewards the agents must distribute some amount of work among themselves (cooperate) and most of that work should advance the same unified goal (coordinate). Learning in both games may be decomposed loosely into two phases. First, before much learning has occurred, very few individuals work consistently toward any goal so defection is motivated by fear that too few others will contribute to successfully establish any norm (the start-up problem). In the later phase of learning, when most individuals are engaged, then the motivation to defect is greed since one can free-ride on the efforts of others (Heckathorn, 1996). Games with this kind of bargaining-like collective

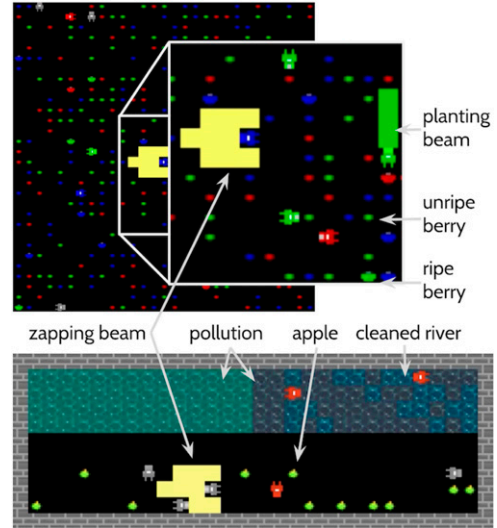


Figure 2. (Top) Allelopathic Harvest. Agents can recolor (replant) berries using one of three colored beams; a green beam is shown here. An agent’s color is given by the berry color they most recently changed a berry to be (planted) or stochastically reverts to gray upon eating a berry. They also can zap agents to punish them (yellow beam). **(Bottom)** Clean Up with Startup Problem. Agents have a cleaning beam that can be used to clean pollution on either side of the divide as well as having a zapping beam that they can use to punish agents.

action problem structure were previously studied with MARL in (Köster et al., 2020).

In *Allelopathic Harvest* (adapted from (Köster et al., 2020)), agents are presented with an environment that contains three different varieties of berry (red, green, and blue) and a fixed number of berry patches, which can be replanted to grow any color variety of berry. The growth rate of each berry variety depends linearly on the fraction that variety (color) comprises of the total. As depicted in

Figure 2, agents have three planting actions with which they can replant berries in front of themselves in their chosen color. Agents in AH have heterogeneous tastes. Specifically, half the agents receive twice as much reward from eating red berries relative to other berries and the other half have preferences of the same form except that they favor green. Agents can achieve higher return by selecting just one single color of berry to plant, but which one to pick is difficult to coordinate (start-up problem). They also always prefer to eat berries over spending time planting (free-rider problem).

In *Clean Up with Startup Problem* (adapted from (Hughes et al., 2018)), the agents need to coordinate on a specific type of pollution to clean out of two pollution types as is shown in Figure 2. The environment contains apples that the agents are rewarded for eating, but the apple spawn rate increases monotonically with the ratio between the two pollution types. If the agents clean both pollution types equally, then apples will not spawn at all. Agents thus need to coordinate on a particular pollution type to clean (start-up problem) while also incentivizing enough agents to do the work of cleaning (free-rider problem).

Both environments have a rule with an effect similar to the cookie example from Section 2. Individuals can see which kind of work (or free riding) other individuals have recently been engaged in. They change color to reflect this information. This makes it easier for agents to identify free-riders and those planting prohibited berry varieties (AH) or cleaning the wrong kind of pollution (CSP). In both environments, the agents are colored according to their most recent planting or cleaning action. For example, successful planting of a red berry (AH) or successful cleaning of red pollution (CSP) causes the agent itself to become red. Similarly, agents that eat fruit are colored gray to indicate that they have not recently planted or cleaned. Thus, gray colored agents are typically free riding.

In both environments agents can zap one another at short-range with a beam. This serves as the punishment mechanism. Importantly, in both games there are also instrumental reasons for agents to zap one another, especially to compete for berries/apples. Getting zapped once freezes the zapped agent for 25 steps and applies a mark that indicates that the agent did something that was disapproved of (similar to (Köster et al., 2022)). If a second zap is received while the agent is marked, the agent is removed for 25 steps and receives a penalty of -10 . If no zap is received for 50 steps, the mark fades. For full details on the environment please refer to Appendix Sec B.

Experiments

Existence and beneficial effects of the emergent social norms

In order to align themselves with the social norm, agents must first learn to represent it accurately. Figure 3 shows the

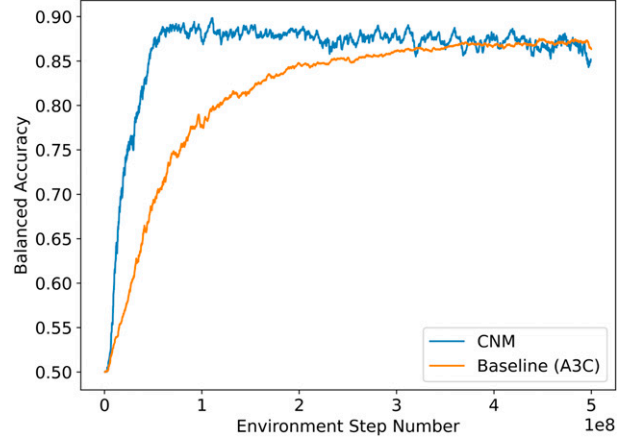


Figure 3. The classifier achieves high balanced accuracy (the average accuracy over both positive and negative samples) in predicting approval versus disapproval events.

balanced accuracy of the classifier in two cases where pseudorewards are on and one where the classifier is left on but has no influence in the environment. We observe three features. First, we are able to rapidly learn a classifier that achieves high balanced accuracy. Our ability to achieve high accuracy despite using only a single frame suggests that the initial normative behavior is something simple like “zap an agent if it might compete with you over a visible berry” or “zap agents of a particular color.” Second, we note that the pseudorewards from the classifier in turn cause the accuracy of the classifier to rapidly converge; the agents adjust their behavior to be in accord with the classifier. Finally, we freeze the classifier after $5e7$ steps, but despite this the balanced accuracy remains relatively high for the duration of training, suggesting that there is not too much drift in the norm after the freeze. Similar behavior is observed in CSP.

Next, we investigate whether the use of CNM leads to better outcomes. In AH we run 20 seeds and in CSP we run 10 seeds. In AH, the measure of success is the *monoculture fraction*, the percentage of the color that corresponds to the largest number of berry spawning sites. Figure 4(b) demonstrates that CNM increases the monoculture fraction above 50%, indicating that agents on average are converging to a single preferred color, and also increases the net agent return, indicating that the costs of norm enforcement (punishing violators) are overcome by increased berry consumption. Similarly, we observe that in CSP they are able to successfully select one of the two pollution types over the other. The inverted minimal fraction measures how imbalanced the two types of pollution are; higher inverted minimal fraction is desirable. The result is a significant consequent increase in collective return. Note that collective return, as defined here, includes the costs of being punished since these are externally imposed by other agents but does not include the pseudoreward term since it models an internal drive.

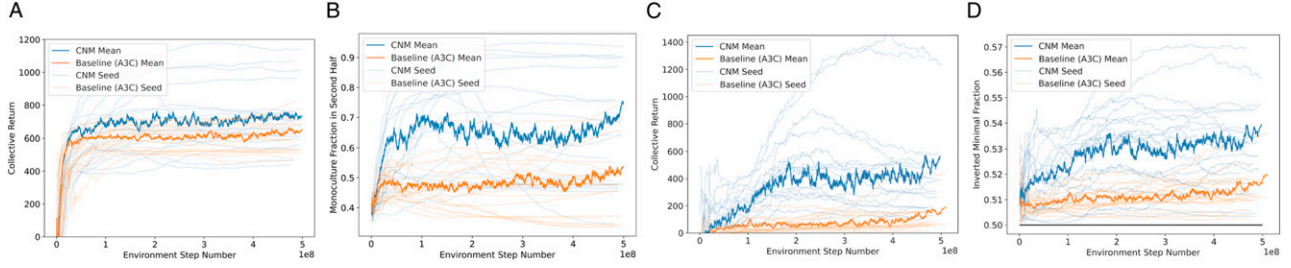


Figure 4. The effect of norms on avoiding start-up problems and overcoming free-rider problems. The thick lines represent the mean across seeds while thin, transparent lines represent individual seeds; the standard deviation is not displayed for visual clarity. (a) Collective return in *AH*. (b) Fraction of total berries constituted by the dominant berry in the second half of the episode. (c) Collective return in *CSP*. (d) Average fraction of total pollution constituted by the dominant pollution type.

Groups of **CNM** agents display a bandwagon effect, magnifying weak patterns of sanctioning in initially random exploratory behavior. They are more likely than the baseline to coordinate on a coherent joint behavior (planting a specific berry color in *AH* or cleaning a specific pollution type in *CSP*). But there is no guarantee that they will select the most beneficial equilibria available to them. This mirrors the arbitrariness of real-world social norms. For example, recall that all agents in *AH* prefer either red or green berries over blue berries (see Section 3). If agents have an early tendency to plant the undesirable blue berries and punish free-riders, the classifier will learn to approve of these behaviors and the agents will stabilize on a blue equilibrium, an outcome that none of them prefer over red or green equilibria. This is why there is so much variation in the outcomes achieved between independent runs (Figure 4). See also Figure 6 where the prevalence of blue berry centric outcomes can clearly be seen.

Finally, we confirm that the improvement in reward is not somehow occurring due to a suppression of the penalty action and a consequent decrease in penalty from zap events; rather, the total amount of punishment events actually stays the same or even increases with **CNM**. Remember, zapping can also be used instrumentally, for example, to compete over berries or apples. Figure 5 shows the average number of zaps in an episode summed over the agents for *AH* and *CSP*. Note that there is no observable amount of difference in the net amount of zapping for *AH* and zapping increases for *CSP*. Thus, improvements in collective return must be coming from changes in how zapping is used.

How does **CNM** establish social norms?

Here, we show that **CNM** increases incentives to obey social norms, that is, agents are disapproved of more for deviating from the established equilibrium. In *AH*, the equilibria are likely given by the corners of the berry fraction simplex (Figure 6). Stabilization comes from disapproval of re-planting behaviors that would push away

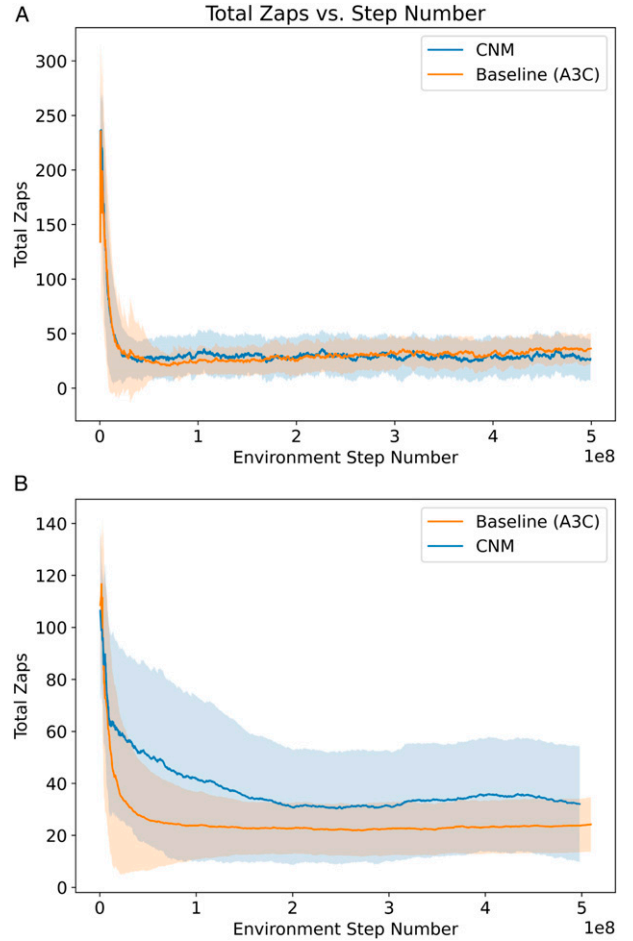


Figure 5. Effect of **CNM** on total number of zaps averaged across seeds in (a) *AH* (b) *CSP*.

from an equilibrium. We can approximately observe stability in the planting behavior by examining the evolution of the fraction of each berry color on the simplex. Figure 6 demonstrates the changes in the evolution of berry fraction during early and late phases of training. Here the center of the diagram indicates that either all agents are free-riding or

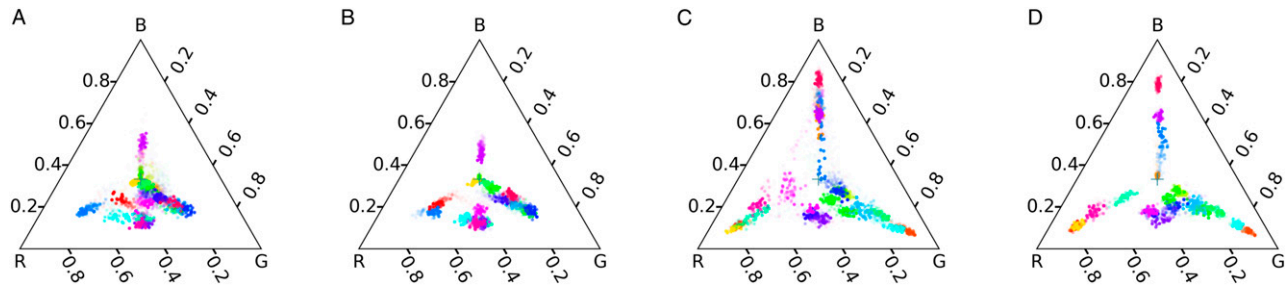


Figure 6. Evidence of stable planting behavior after 2e8 steps of training. Individual dots are samples over a run where darker dots represent later points. (a) First 2e8 steps with **CNM** off. (b) Latter 2.5e8 steps with **CNM** off. (c) First 2e8 steps with **CNM** on. (d) Latter 2.5e8 steps with **CNM** on.

that they are all canceling out one another's planting behavior (e.g. I change a red berry to blue and you change a blue berry to red so there is no net effect on the berry fractions). We observe that groups of **CNM** agents push further away from the center and towards the corners of the simplex. Furthermore, there is little change in later steps of training for the seeds that reach the simplex corners, suggesting an equilibrium. There is some small amount of drift in high blue monoculture fractions which may be occurring as the blue berries are not preferred by any agent.

The second criterion to check concerning the establishment of a social norm is that deviations from the equilibrium should be disapproved (sanctioned). We can calculate for each color $p(\text{zapped} \rightarrow \text{color})$ by Bayes' rule (details in [Appendix](#)). We then use it to investigate the sanctioning forces supporting a particular equilibrium by looking at the difference in log likelihood of being punished while working toward establishing or maintaining the equilibrium. Agents can readily perceive which equilibrium other agents in their field of view are supporting because their color shows which color berry they last planted (see Section 3). If the likelihood difference for a particular color is high it should be easy for the learning algorithm to identify that switching to that color (i.e. switching to support its corresponding equilibrium) is likely to lead to disapproval. Thus, these differences serve as a teaching signal pushing the agent towards planting one color and away from planting another.

[Figure 7](#) demonstrates this effect for two different potential switches. [Figure 7\(a\)](#) measures the difference of punishment likelihood between free-riding and planting the dominant color which we call *teaching signal 1*. If the magnitude of this signal is large and positive, it is easier for the learning algorithm to identify that switching from free-riding to planting in that color will decrease the amount that it gets punished.

[Figure 7\(b\)](#) measures the relative likelihood of getting punished when we plant the color corresponding to high monoculture versus if we were to switch to plant the second most abundant berry color which we refer to as *teaching*

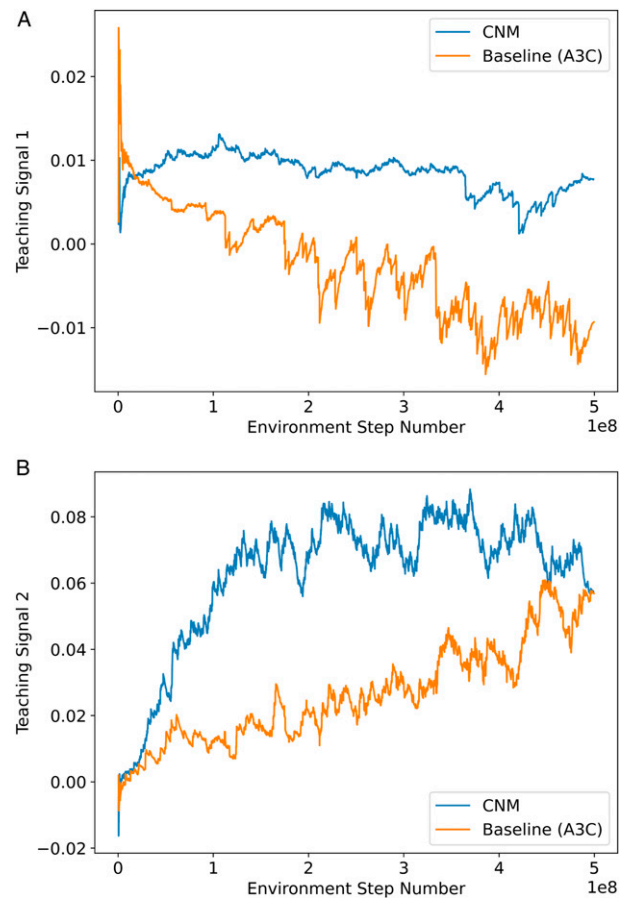


Figure 7. Measurements of the strength of disapproving sanctions applied to deviating agents in allelopathic harvest: (a) difference between zap likelihood for being the free-rider color versus being the dominant color (b) difference of zap likelihood for being the second most dominant color to being the dominant color.

signal 2. If this signal is large, it is easier for the learning algorithm to identify that sticking to the dominant color will allow it to decrease how often it gets disapproved of which in turn will help stabilize the choice of equilibrium.

Ablations on architecture components

To understand CNM better we gradually remove and alter components of the architecture to answer the following questions: (1) Is freezing the classifier necessary? (2) Is it essential to learn social norms from global sanctions or will local sanctions observed by each individual themselves suffice? (3) Is our result sensitive to the relative scale between approval and disapproval pseudorewards?

Here we study CSP as the smaller number of agents in this environment decreases environment step time and allows us to perform more rapid experimentation. We run each ablation over 10 seeds. For point (1), we allow the classifier to continue learning throughout training. For (2) we train the classifier using only the sanctioning events directly observed by each agent. Finally, for (3), we note that in all prior experiments we have scaled the pseudorewards so that the penalty for punishing discordantly with the classifier (β) is twice the reward for punishing in accord with it (α). We aim to establish whether our results are sensitive to this particular ratio.

Figure 8 demonstrates the outcome of all of these ablations; each curve is the average across 10 seeds with std. deviations removed for visual clarity. In Figure 8(a), we can see that in the absence of a frozen classifier the collective return experiences a large early spike but then decays quickly down. While we are unable to definitively establish the mechanism that forces us to freeze the classifier, there are a few plausible ones. The move away from free-riding occurs rapidly in the first $1e8$ steps of training (see Appendix Sec C). If the punishment behavior is not correspondingly suppressed quickly enough, agents performing cooperative behavior will still get punished due to exploratory noise and the classifier will consequently learn to recommend punishment of cooperative agents. Alternately, the classifier could simply experience *catastrophic forgetting* once a particular color is effectively suppressed: it’s difficult to remember how to sanction a behavior that no longer occurs. Consequently, the suppressed behavior is able to re-emerge.

In Figure 8(b), we observe that learning solely from local sanctions does improve over the baseline but does not completely match the performance of fully public sanctions. Since the agents have to infer the norm solely through agents they happen to interact with, the number of samples available for each classifier update decreases sharply which may make the subsequent learned norm noisier and harder to learn. Finally, in Figure 8(c), we set the pseudorewards to a magnitude of 0.9 for both approval and disapproval. We note that this is less than the potential reward of consuming an apple, making it feasible for an agent to zap discordantly to the recommendation of the classifier if doing so nets them

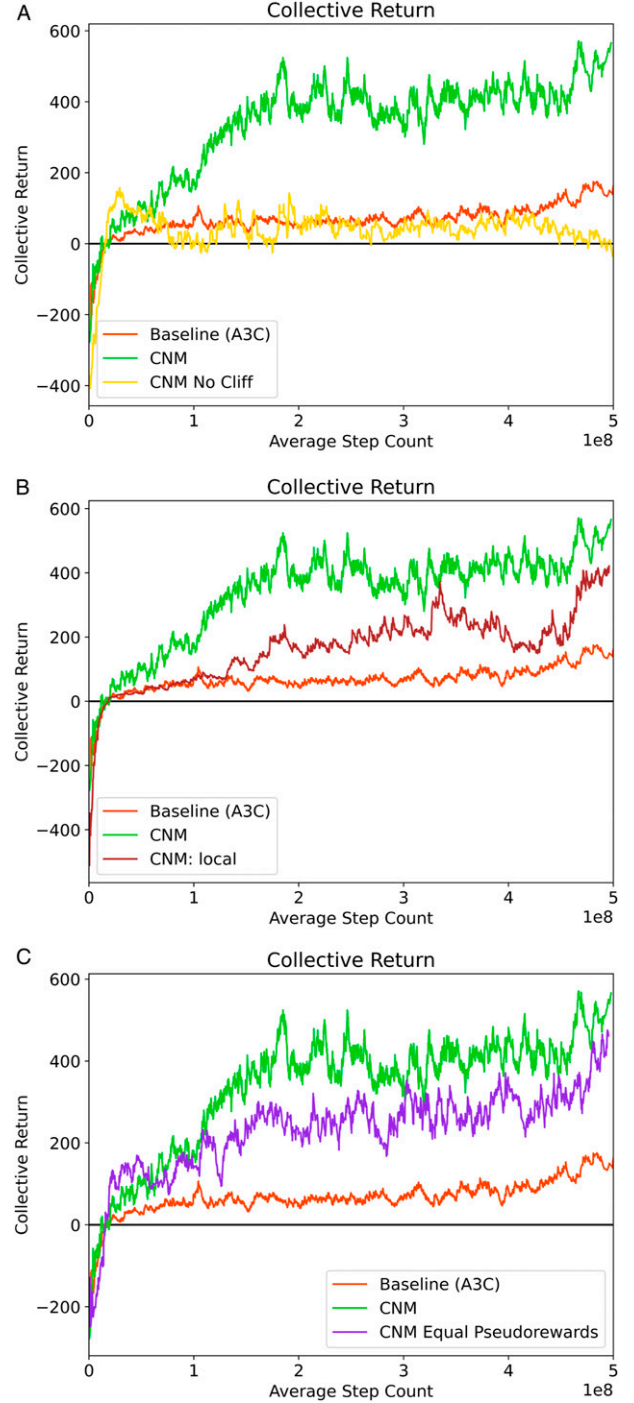


Figure 8. Ablations of key components of the agent architecture. (a) The classifier is not frozen during training. (b) The classifier is learned solely from sanctions experienced by the agent, that is, sanctions are private. (c) Effect of pseudoreward scale; both α and β are set to 0.9.

an additional apple. We see that there is a slight reduction in the collective return but there remains an improvement over the A3C baseline.

Discussion and future work

Motivated by emerging challenges in deploying multi-agent systems, we introduce and formalize a new training regime for decentralized multi-agent systems in which all sanctions are publicly observable. In contrast to centralized training methods, this approach can be trained fully online without needing access to a simulator. It also may make it easier to satisfy privacy constraints since essential proprietary data like rewards and policies do not need to be shared to achieve coordination.

We observe that in this setting decentralized agents struggle to achieve cooperative behavior in the collective action problems posed by two environments that broadly model challenges of free-riding and equilibrium selection. Inspired by social norms, which humans communities often use to overcome such dilemmas, we introduce an agent architecture **CNM** that learns to classify and enforce social norms from experience. We show that groups of **CNM** agents converge on beneficial equilibria and are better at resolving free-rider problems than agents implementing a baseline algorithm.

However, many open questions remain. The architecture used for the classifier, a convolutional network, relies on there being an identifiable visual cue that correlates with the behavior to be made normative. Thus, it is restricted in the types of norms it can identify. An extended **CNM** architecture operating on snippets of video preceding each sanctioning event may allow for different social norms to emerge.

Furthermore, while we observe the appearance of seemingly stable, beneficial norms, we do not provide a complete mechanistic explanation of how this architecture selects and stabilizes equilibria. It is possible that there exist games where this architecture would exclusively select harmful norms or deeply unfair norms. From the standpoint of using **CNM** for social science modeling, this is a feature not a bug. In the real world, for every beneficial norm enabling collective action, there are hosts of unsavory norms (but see also (Hadfield-Menell et al., 2019; Köster et al., 2022)). Moreover, we must not take for granted that social norms are always a desirable outcome for a multi-agent system. For instance, social norms impose a deadweight loss due to the effort needed to maintain them. Paying this cost may not always be worthwhile in all applications. Nevertheless, we believe that **CNM**, or a successor system, could eventually be employed fruitfully in a wide range of applications from social science modeling to real-world multi-agent systems where interfacing with human social norms is especially critical.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Eugene Vinitsky  <https://orcid.org/0000-0003-2372-4944>

Joel Z Leibo  <https://orcid.org/0000-0002-3153-916X>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Symmetrically, it is possible to define approval events to be when the agent with the opportunity takes an action to reward the other agent and disapproval events to be when it does not do so (positive sanctioning). However, we do not consider that case here. We made this choice because the bulk of the literature on sanctioning and social norms is primarily concerned with negative sanctioning (Baldwin, 1971; Bicchieri et al., 2018; Carroll et al., 2019).
2. See <https://youtu.be/la24sFmk618> and <https://youtu.be/A4zMh9359r8> for videos of example episodes of *AH* and *CSP*, respectively.

References

- Akerlof G (1976) The economics of caste and of the rat race and other woeful tales. *The Quarterly Journal of Economics* 90: 599–617.
- A.v.d O, Li Y and Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748.
- Baker B. Emergent reciprocity and team formation from randomized uncertain social preferences. In: Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, virtual, 6–12 December 2020.
- Balafoutas L, Nikiforakis N and Rockenbach B (2014) Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the United States of America* 111(45): 15924–15927.
- Baldwin DA (1971) The power of positive sanctions. *World Politics* 24(1): 19–38.
- Bicchieri C, Muldoon R and Sontuoso A (2018) Social norms. *The Stanford Encyclopedia of Philosophy*. Winter 2018. Stanford, CA: Metaphysics Research Lab, Stanford University.
- Bicchieri C (2006) *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
- Bicchieri C (2016) *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford: Oxford University Press.
- Boehm C (2012) *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Soft Skull Press.

- Boyd R and Mathew S (2021) Arbitration supports reciprocity when there are frequent perception errors. *Nature Human Behaviour* 5: 596–603.
- Carroll M, Shah R, Ho MK, et al. On the utility of learning about humans for human-ai coordination. In: Wallach HM, Larochelle H, Beygelzimer A (eds) et al. *Advances in neural information processing systems* 32: annual conference on neural information processing systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019, pp. 5175–5186.
- Çolak S, Lima A and González MC (2016) Understanding congested travel in urban areas. *Nature Communications* 7(1): 10793–10798.
- Crandall JW, Oudah M, Ishowo-Oloko F, et al. (2018) Cooperating with machines. *Nature Communications* 9(1): 233–312.
- Eccles T, Hughes E, Kramár J, et al. (2019) Learning reciprocity in complex sequential social dilemmas. arXiv Preprint arXiv 190308082.
- Espeholt L, Soyer H, Munos R et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In: *International conference on machine learning*. PMLR, Stockholmsmässan, Stockholm, Sweden, 10 July 2018, pp. 1407–1416.
- Fehr E and Fischbacher U (2004) Social norms and human cooperation. *Trends in Cognitive Sciences* 8(4): 185–190.
- Foerster J, Farquhar G, Afouras T et al. Counterfactual multi-agent policy gradients. In: *Proceedings of the AAAI conference on artificial intelligence*, New Orleans, LA, USA, 2–7 February 2018a, vol. 32.
- Foerster J, Chen RY, Al-Shedivat M et al. Learning with opponent-learning awareness. In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, Stockholm Sweden, 10–15 July 2018b, pp. 122–130.
- Gemp I, McKee KR, Everett R, et al. (2020) D3c: Reducing the price of anarchy in multi-agent learning. arXiv Preprint arXiv 201000575.
- Gintis H (2010) Social norms as choreography. *Politics, Philosophy & Economics* 9(3): 251–264.
- Granovetter M (1978) Threshold models of collective behavior. *American Journal of Sociology* 83(6): 1420–1443.
- Hadfield GK and Weingast BR (2012) What is law? a coordination model of the characteristics of legal order. *Journal of Legal Analysis* 4(2): 471–514.
- Hadfield GK and Weingast BR (2013) Law without the state: legal attributes and the coordination of decentralized collective punishment. *Journal of Law and Courts* 1(1): 3–34.
- Hadfield GK and Weingast BR (2014) Microfoundations of the rule of law. *Annual Review of Political Science* 17: 21–42.
- Hadfield-Menell D, Andrus M and Hadfield G. Legible normativity for ai alignment: the value of silly rules. In: *Proceedings of the 2019 AAAI/ACM conference on ai, ethics, and society*, pp. 115–121.
- Heckathorn DD (1996) The dynamics and dilemmas of collective action. *American Sociological Review* 61: 250–277.
- Henrich J and Muthukrishna M (2021) The origins and psychology of human cooperation. *Annual Review of Psychology* 72: 207–240.
- Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8): 1735–1780.
- Hughes E, Leibo JZ, Phillips M et al. (2018) Inequity aversion improves cooperation in intertemporal social dilemmas. In: *Advances in neural information processing systems* 31: annual conference on neural information processing systems 2018, NeurIPS 2018, Montréal, Canada, 3–8 December 2018. pp. 3330–3340.
- Iqbal S and Sha F (2019) Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. arXiv Preprint arXiv 190512127.
- Jaderberg M, Mnih V, Czarnecki WM et al (2017) Reinforcement learning with unsupervised auxiliary tasks. In: *5th international conference on learning representations, ICLR 2017*, Toulon, France, 24–26 April 2017.
- Jaques N, Lazaridou A, Hughes E et al (2019) Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: *International conference on machine learning*. PMLR, Long Beach, CA, 2019, pp. 3040–3049.
- Kairouz P, McMahan HB, Avenet B et al. (2019) Advances and open problems in federated learning. arXiv preprint arXiv:191204977.
- Kelley HH, Holmes JG, Kerr NL, et al. (2003) *An Atlas of Interpersonal Situations*. Cambridge: Cambridge University Press.
- Köster R, McKee KR, Everett R et al. (2020) Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. arXiv preprint arXiv:201009054.
- Köster R, Hadfield-Menell D, Everett R, et al. (2022) Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences of the United States of America* 119(3): e2106028118.
- Lancot M, Zambaldi V, Gruslys A et al. (2017) A unified game-theoretic approach to multiagent reinforcement learning. arXiv preprint arXiv:171100832.
- Leibo JZ, Zambaldi V, Lancot M et al. (2017) Multi-agent reinforcement learning in sequential social dilemmas. In: *Proceedings of the 16th international conference on autonomous agents and multiagent systems (AA-MAS 2017)*, Sao Paulo, Brazil, 2017.
- Leibo JZ, Dueñez-Guzman EA, Vezhnevets A et al (2021) Scalable evaluation of multi-agent reinforcement learning with melting pot. In: *International conference on machine learning*. PMLR, Online, 2021, pp. 6187–6199.
- Leibo JZ, Hughes E, Lancot M, et al. (2019) Autocurricula and the emergence of innovation from social interaction: a manifesto for multi-agent intelligence research. arXiv Preprint arXiv 190300742.
- Lewis D (1969) *Convention*. Cambridge: Harvard University Press.
- Liang X, Liu Y, Chen T et al. (2019) Federated transfer reinforcement learning for autonomous driving. arXiv preprint arXiv:191006001.

- LittmanMarkov ML. games as a framework for multi-agent reinforcement learning. In: Proceedings of the 11th international conference on machine learning (ICML), New Brunswick, USA, 1994, pp. 157–163.
- Lowe R, Wu Y, Tamar A et al (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st international conference on neural information processing systems, Long Beach, USA, 2017, pp. 6382–6393.
- Lupu A and Precup D (2020) Gifting in multi-agent reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, Auckland, NZ, 2020, pp. 789–797.
- Mackie G (1996) Ending footbinding and infibulation: a convention account. *American Sociological Review* 61: 999–1017.
- Marwell G and Oliver P (1993) *The Critical Mass in Collective Action*. Cambridge: Cambridge University Press.
- McAleer S, Lanier J, Baldi P et al. (2021) Xdo: A double oracle algorithm for extensive-form games. arXiv preprint arXiv: 210306426.
- McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102(3): 419–457.
- McKee KR, Gemp I, McWilliams B et al. (2020) Social diversity and social preferences in mixed-motive reinforcement learning. In: Proceedings of the 19th international conference on autonomous agents and multiagent systems, Auckland, New Zealand, 9–13 May 2020, pp. 869–877.
- Mnih V, Badia AP, Mirza M et al (2016) Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. PMLR, New York City, USA, 2016, pp. 1928–1937.
- Nyborg K, Anderies JM, Dannenberg A, et al. (2016) Social norms as solutions. *Science* 354(6308): 42–43.
- Olson M (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups, Second Printing with a New Preface and Appendix*. Cambridge: Harvard University Press, Vol. 124.
- Ostrom E (1998) A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American Political Science Review* 92(1): 1–22.
- Ostrom E (2009) *Understanding Institutional Diversity*. Princeton: Princeton University Press.
- Pérolat J, Leibo JZ, Zambaldi VF, et al. (2017) A multi-agent reinforcement learning model of common-pool resource appropriation. In: Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, Long Beach, CA, USA, 4–9December 2017, pp. 3643–3652.
- Peysakhovich A and Lerer A (2018) Prosocial learning agents solve generalized stag hunts better than selfish ones. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, Stockholm, Sweden, 2018, pp. 2043–2044.
- Platanios EA, Saparov A and Mitchell T (2020) Jelly bean world: A testbed for never-ending learning. *International conference on learning representations*. Ethiopia: ICLR. Addis Ababa.
- Rashid T, Samvelyan M, Schroeder C et al (2018) Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: *International conference on machine learning*. PMLR, Stockholm, Sweden, 2018, pp. 4295–4304.
- Sen S and Airiau S. Emergence of norms through social learning. In: IJCAI, Vol. 1507, pp. 1512.
- Shapley LS (1953) Stochastic games. In *Proc of the National Academy of Sciences of the United States of America*.
- Shoham Y and Tennenholtz M (1997) On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence* 94(1–2): 139–166.
- Singh S, Barto AG and Chentanez N. Intrinsically motivated reinforcement learning. In: Proceedings of the 17th international conference on neural information processing systems. Vancouver British Columbia Canada, 1 December 2004, pp. 1281–1288.
- Stanley KO (2019) Why open-endedness matters. *Artificial Life* 25(3): 232–235.
- Sunehag P, Lever G, Gruslys A et al. (2018) Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems. Stockholm Sweden, 10–15 July 2018, pp. 2085–2087.
- Tomasello M and Vaish A (2013) Origins of human cooperation and morality. *Annual Review of Psychology* 64: 231–255.
- Ullmann-Margalit E (1977-2015) *The Emergence of Norms*. Oxford: Oxford University Press.
- Vanderschraaf P (1995) Convention as correlated equilibrium. *Erkenntnis* 42(1): 65–87.
- Vezhnevets A, Wu Y, Eckstein M et al. (2020) Options as responses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In: International conference on machine learning. PMLR, Virtual, 13–18 July 2020, pp. 9733–9742.
- Wang JX, Hughes E, Fernando C et al. (2018) Evolving intrinsic motivations for altruistic behavior. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, pp. 683–692.
- Wang WZ, Beliaev M, Biyik E et al. (2021) Emergent prosociality in multi-agent games through gifting. In: 30th International joint conference on artificial intelligence (IJCAI).
- Wiessner P (2005) Norm enforcement among the ju/hoansi bushmen. *Human Nature* 16(2): 115–145.
- Xiao E and Houser D (2005) Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America* 102(20): 7398–7401.
- Young HP (2015) The evolution of social norms. *Annual Review of Economics* 7(1): 359–387.