Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games

Ninell Oldenburg
University of Copenhagen
niol@hum.ku.dk

Tan Zhi-Xuan
Massachusetts Institute of Technology
xuan@mit.edu

ABSTRACT

A universal feature of human societies is the adoption of systems of rules and norms in the service of cooperative ends. How can we build learning agents that do the same, so that they may flexibly cooperate with the human institutions they are embedded in? We hypothesize that agents can achieve this by assuming there exists a shared set of norms that most others comply with while pursuing their individual desires, even if they do not know the exact content of those norms. By assuming shared norms, a newly introduced agent can infer the norms of an existing population from observations of compliance and violation. Furthermore, groups of agents can converge to a shared set of norms, even if they initially diverge in their beliefs about what the norms are. This in turn enables the stability of the normative system: since agents can bootstrap common knowledge of the norms, this leads the norms to be widely adhered to, enabling new entrants to rapidly learn those norms. We formalize this framework in the context of Markov games and demonstrate its operation in a multi-agent environment via approximately Bayesian rule induction of obligative and prohibitive norms. Using our approach, agents are able to rapidly learn and sustain a variety of cooperative institutions, including resource management norms and compensation for pro-social labor, promoting collective welfare while still allowing agents to act in their own interests.

KEYWORDS

Norm Learning, Cooperative Intelligence, Norm Emergence, Bayesian Learning, Normative Systems, Markov Games

ACM Reference Format:
Ninell Oldenburg and Tan Zhi-Xuan. 2024. Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 18 pages.

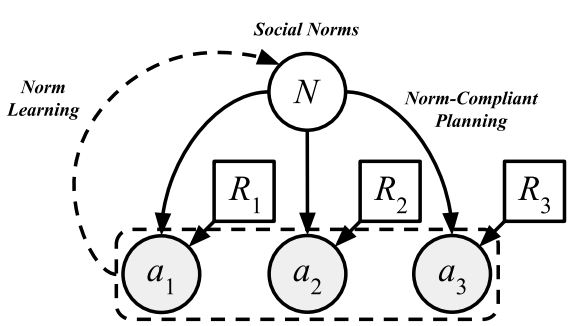
1 INTRODUCTION

For autonomous agents to integrate with society, they will have to learn to comprehend and comply with the unspoken rules and shared expectations that pervade human interaction: social norms [11]. Such norms function as a way of enforcing common standards and promoting cooperative behavior [14, 26, 30, 66], and thus serve as promising targets for aligning autonomous systems with societal interests even while they primarily serve the goals of their users [4, 64, 100]. How then can we build agents that learn and comply with norms? In answering this question, we need to contend with a paradoxical aspect of social normativity: Even though social norms function as *shared* constraints on behavior [26, 90], they are primarily learned and sustained in a *decentralized* and *emergent* manner [6, 40, 62, 68]. As such, a useful model of norm learning should capture its shared-yet-decentralized nature, enabling agents to rapidly coordinate in new contexts without central control [1, 36].

In this paper, we develop a Bayesian account of social norm learning which directly addresses this paradox. Drawing upon frameworks for joint intentionality [86, 88, 98] and shared agency [18, 90] in human cooperation, we model agents that assume *shared normativity* when learning to interact with others. In particular, each agent assumes that all other agents may be complying with a shared set of norms even while pursuing their own interests, allowing them to infer the existence of a norm from apparent violations of self-interest, and to aggregate such observations across a population of agents, updating their beliefs about which norms best explain collective behavior. Having inferred these norms, an agent can then comply with them, either because of an intrinsic motivation [55], or because it is strategically valuable [61]. Under sufficiently high levels of compliance, shared normativity is thus *self-sustaining*: Once enough agents infer a shared set of norms, they conform to them, thereby creating public knowledge that the norms are practiced, and enabling new entrants to rapidly learn the norms.

We formalize this account in the context of multi-agent sequential decision-making, introducing *norm-augmented Markov games* as a framework for studying norm-guided learning and planning. In contrast to similar frameworks based on (model-free) reinforcement learning (RL) [44, 51, 92], our approach foregrounds the *epistemic* and *cognitive* aspects of social norms, framing norm learning as an (approximately) rational Bayesian process of learning structured rules [21, 33, 65, 77, 101], rather than a reactive process of adapting one's habitual actions [7, 49]. As such, our account captures the speed and flexibility of norm learning and coordination in human children and adults [74, 75, 80], while providing a normative standard against which other approaches can be measured. We instantiate and evaluate this framework in DeepMind's Melting Pot simulator [1, 50], demonstrating that approximately Bayesian norm learning is indeed feasible in long-horizon multi-agent contexts while requiring orders of magnitude less experience than model-free norm learning from sanctions [49, 92]. We also find that shared normativity enables both the maintenance and convergence of normative systems: Common knowledge of the norms is maintained despite the "deaths" of experienced agents and the "births" of normative novices, while groups of agents can bootstrap themselves towards a stable set of norms by learning from each others' actions and experimentally complying with tentative norms.

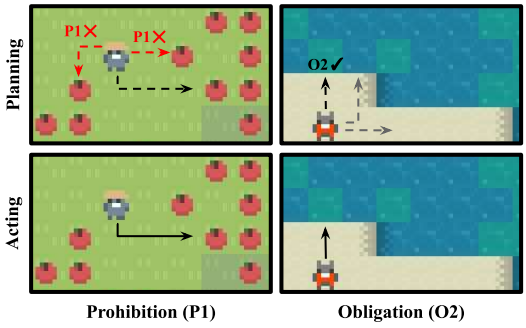
Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).



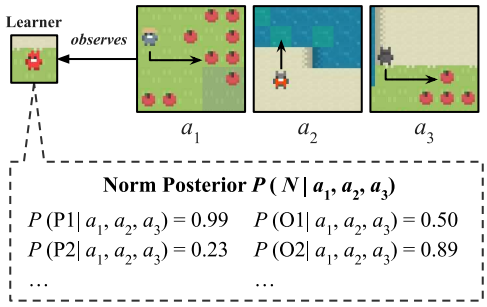
(a) Norm-Augmented Markov Game

English	Prohib. Action	Postcondition
P1: Don't empty apple orchard.	*Move(m, c)*	*Apl(c)* ∧ |*AplNear(c)*| <3
P2: Don't steal apples.	*Move(m, c)*	*Apl(c)* ∧ *Foreign(c)*
English	**Precondition**	**Postcondition**
O1: As farmer, reg. pay cleaner.	*LastPaid(m)* >30 ∧ *Look(m)* =*F*	*LastPaid(m)* =0
O2: As cleaner, clean if dirt > 30%.	*Dirt(r)* >0.3 ∧ *Look(m)* =*C*	*Cleaned(m, r)*
O3: If egalitarian, clean if dirt > 45%.	*Dirt(r)* >0.45 ∧ *Look(m)* =*E*	*Cleaned(m, r)*

(b) Social Norms



(c) Norm-Compliant Planning



(d) Norm Learning

Figure 1: Overview of our framework, summarized by **(a)** the (simplified) graphical model for a norm-augmented Markov game. Agents learn to comply with rule-based social norms **(b)**, which are factored into prohibitions (e.g. P1, P2) and obligations (e.g. O1, O2, O3). Agents comply with norms while pursuing their interests via model-based planning **(c)**, which switches between a reward-oriented mode that obeys prohibitions while collecting rewards (left), and an obligation-oriented mode that plans to satisfy a postcondition (right). Agents learn these norms via **(d)** approximate Bayesian inference, updating the probabilities of each potential norm based on observations of others' actions.

Collectively, these findings highlight the importance of integrating game-theoretic accounts of norms and conventions [6, 14, 30] with the conceptual richness and flexibility of human learning [33, 34, 71]. By uniting the structured representations of normative concepts common in normative systems research [15, 16, 19] with a Bayesian decision-theoretic approach to multi-agent coordination [5, 45], our framework models the dynamism of human normative cognition that is omitted from model-free RL approaches, while paving the way for normatively-competent autonomous systems that combine model-free and model-based learning [22, 52].

2 LEARNING AND SUSTAINING NORMATIVE SYSTEMS IN MARKOV GAMES

How can we design agents that learn and maintain systems of social norms? As motivated above, we first introduce *norm-augmented Markov games* (NMGs) as an extension of Markov games [20, 31, 57, 59, 72], providing a formal setting for norm learning and coordination in sequential decision-making (Figure 1a). As an instance of such a game, we adapt the Melting Pot simulator [1, 50] to create a multi-agent environment that affords a range of cooperative norms. We then introduce a rule-based representation of social norms (Figure 1b), which take the form of obligations or prohibitions, and show how agents can comply with these norms through model-based planning (Figure 1c). Finally, we show how agents can perform approximately Bayesian *rule induction* [33, 63], inferring norms from observations of violation and compliance with potential obligations and prohibitions (Figure 1d).

2.1 Norm-Augmented Markov Games

In Markov games [31, 57, 59, 72], a set of M agents take actions $a_i \in \mathcal{A}_i$ in a series of environment states $s \in \mathcal{S}$ over time, where $i \in \mathcal{I} := \{1, ..., M\}$ is each agent's index. Actions change the environment state via a transition function $T : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_M \to \Delta(\mathcal{S})$ that maps the current state s and joint actions $(a_1, ..., a_M)$ to a distribution over successor states s'. Each agent also has a reward function $R_i : \mathcal{S} \times \mathcal{A}_i \times \mathcal{S} \to \mathbb{R}$ that maps state transitions (s, a, s') to scalar rewards $r \in \mathbb{R}$. We also assume a shared discount factor γ. In our context, an agent's reward function can be understood as a representation of their individual desires. Markov games begin with an initial state s_1, and end after a horizon of H steps.

We extend Markov games with the addition of *social norms*, formalized as functions $\nu : \mathcal{H} \to \{0, 1\}$ where $\mathcal{H}_i := \bigcup_{t=1}^{H}((\mathcal{S} \times \mathcal{A}_i)^t \times \mathcal{S})$ is the set of state-action histories for agent i, and $\mathcal{H} := \bigcup_{i \in \mathcal{I}} \mathcal{H}_i$ is the set of histories across all agents. In words, a social norm ν *classifies* whether the behavioral history of an agent i complies with the norm [36], with 0 indicating compliance and 1 indicating violation. We call a set of norms N a *normative system.*

Let $\mathcal{N}$ be a set of functions defining the space of possible social norms, and $\epsilon \in [0, 1]$ be the frequency of a norm violation. We now define norm-augmented Markov games as a (subjective) Bayesian extension of a Markov game, endowing each agent i with a subjective prior $P_i^0(N)$ over which sets of norms $N \subseteq \mathcal{N}$ will be complied with by *all agents* with frequency at least $1 - \epsilon$:

$$P_i^0(\nu) = P_i\left(\frac{1}{HM}\sum_{t,j=1}^{H,M}(1 - \nu(h_j^t)) \geq 1 - \epsilon\right), \quad P_i^0(N) = \prod_{\nu \in N} P_i^0(\nu) \tag{1}$$

where h_j^t is the state-action history for agent j at step t, treated as a random variable. Each agent also has conditional prior $P_i(h_t|N)$ over agent histories h_t given norms N, and this induces a sequence of posteriors over norms $P_i^t(N) := P_i(N|h_t) \propto P_i^0(N)P_i(h_t|N)$, defining agent i's belief in norms N at each t after observing h_t. We also endow each agent i with a norm violation cost function $C_i : \mathcal{N} \to \mathbb{R}$, which defines how intrinsically motivated an agent is to comply with norms that they believe to be true. This induces a *norm-augmented reward function* $R_i' : \mathcal{H} \to \mathbb{R}$. Given a history $h = (h', s, a_i, s')$, R_i' decomposes into a base reward function R_i and the expected cost of norm violation under the posterior $P_i(N|h_t)$:

$$R_i'(h', s, a_i, s') := R_i(s, a_i, s') - \sum_{N \subseteq \mathcal{N}} P_i(N|h) \sum_{\nu \in N} C_i(\nu)\nu(h) \tag{2}$$

By adding priors over norms to a Markov game, each agent i effectively factors their beliefs about other agent's policies π_{-i} into two parts: Beliefs about whether the policies will comply with a *shared* set N of relatively simple yes/no rules, and beliefs about non-normative features of those policies. This is illustrated in Figure 1a, depicting how agents' actions a_i are influenced by both shared norms N and individual reward functions R_i. An agent's prior over norms thus serves as a *correlating device* over their beliefs about how other agents behave [5, 30]. If agents then act according to these correlated beliefs (e.g. by complying with the norms), then their *actual* policies will also be correlated [45], helping agents achieve coordination. The existence of norm violation costs C_i further aids this coordination, promoting compliance with a normative system N once it is sufficiently believed in.

When is a normative system N stable? In general, a set of beliefs about norms $\{P_i^t(N)\}_{i=1}^M$ at time t is consistent with a wide range of correlated strategies $\pi_{1:M}$, some of which may form (subjective) correlated equilibria: joint policies which no agent has an incentive to deviate from [5, 59]. Out of these equilibria, consider those where the *expected frequency* of each agent i complying with norms $N \sim P_i^t(N)$ over all future steps is at least $1 - \epsilon$, where the expectation is taken over their own belief distribution $P_i^t(N)$ over the norms. Since these are policies where agents frequently comply with the norms they (subjectively) believe might be true, we call them *subjective normative equilibria*. If all agents have converged to a single distribution $P^t(N)$ over normative systems (or δ-close to it), this forms an *(objective) normative equilibrium* (respectively, a normative δ-equilibrium). If $P^t(N)$ *concentrates* on a specific N, then N is *stable* with respect to that normative equilibrium. Note that if $P^t(N)$ is concentrated enough on some N, and agents are (norm-augmented) reward-maximizers, we can guarantee that N is stable by setting violation costs for $\nu \in N$ to be sufficiently high.

As an example setting for norm-augmented Markov games, we adapt and combine several existing games from the Melting Pot suite of environments [1, 50], including elements of Commons Harvest [43, 72], Clean Up [42], and Territory [50]. The resulting environment is shown in Figure 1. In the environment, agents can collect and eat apples to gain reward but otherwise incur a cost for actions. Apples regrow at a density-dependent rate (more surrounding apples lead to faster regrowth), but the rate decreases as the nearby river gets polluted over time. To prevent pollution, agents have to use clean-up action on the river. In addition, several regions are visually marked as each agent's territory (but do not otherwise have functional effects), and agents have one of three visually distinct appearances that demarcate *roles*. By combining these elements, we allowed for a range of social norms, including resource management norms, property norms, role-based division of labor, and compensation for pro-social labor. In our experiments, we gave experienced agents high prior beliefs in some norm set N, and learning agents a low prior belief in all (non-empty) norm sets. We present details of the environment's dynamics in the Appendix.

2.2 Representing Social Norms

NMGs are a general setting for multi-agent norm learning, but they leave unspecified the set $\mathcal{N}$ of possible norms. Drawing upon deontic and temporal logic representations of norms [2, 21, 46], we consider learning both (maintenance-based) *prohibitions* $\mathcal{N}_P$, which forbid the effects of certain actions at all times, and (achievement-based) *obligations* $\mathcal{N}_O$, which require pursuing *temporally extended* goals, giving our full set of norms $\mathcal{N} = \mathcal{N}_P \cup \mathcal{N}_O$. Several examples are shown in Figure 1b. Note that the distinction between prohibitions and obligations is not fundamental: Prohibitions can be represented as obligations via negating the effects and vice versa. Nonetheless, we found it more natural to represent maintenance norms in their prohibitive form (e.g. "Don't empty the apple orchard.") and achievement norms in their obligative form (e.g. "Pay the cleaner in the next 30 steps.").

Following prior work on norm and rule learning [33, 101], we express norms as rules in a first-order language $\mathcal{L}$, which includes predicates and functions defined over objects in a state s. Given a first-order expression ϕ, we denote its valuation in s as $s[\phi]$. Using this notation, we represent a prohibition norm $\nu_P \in \mathcal{N}_P$ as a tuple $\langle \text{Prohib}, \text{Post} \rangle$ where Prohib is a function of objects in s that returns a set of potentially prohibited actions, and the postcondition Post holds in a state s whenever the prohibited effect occurs. In other words, if an action $a_i \in s[\text{Prohib}]$ leads to state s' such that $s'[\text{Post}]$ is true, this means that the prohibition has been violated, and hence that $\nu_P(h, s, a, s') = 1$. As an example, an agent m in our environment is prohibited from moving to a cell c (Prohib = Move(m, c)) if that cell contains apples and is in foreign territory (Post = Apl(c) $\wedge$ Foreign(c)) as depicted in Figure 1b.

Next, we define obligations $\nu_O \in \mathcal{N}_O$ as tuples $\langle \text{Pre}, \text{Post}, \tau \rangle$, where Pre is a precondition formula that holds true when the obligation is triggered, and Post is a postcondition formula that holds true once the obligation has been discharged, with τ as the duration during which the obligation must be satisfied. Given a history $h = (s_{t-k}, a_{-i,t-k}, ..., s_t, a_{n,t})$ up to time step t, this means that $\nu_O(h) = 1$ (the obligation is violated) if there exists some $k \leq \tau$ such

that $s_{t-k}[\text{Pre}]$ is true but $s_{t-k+j}[\text{Post}]$ is false for all $k < j \leq \tau$. As another example, when agents are in the cleaner role (Look(m) = C) and find the river more than 30% polluted (Dirt(r) > 0.3) they are obliged to have cleaned the river (Cleaned(m, r)) within the next 20 steps ($\tau = 20$). Therefore, (Pre = (Dirt(r) > 0.3 ∧ Look(m) = C), Post = Cleaned(m, r), $\tau = 20$).

2.3 Norm-Compliant Planning

Having introduced how norms classify behavior, we now describe how agents can plan to comply with a set of norms $N = N_P \cup N_O$ where $N_P \subseteq \mathcal{N}_P$ and $N_O \subseteq \mathcal{N}_O$. Absent strategic considerations (which norms themselves are meant to solve), each agent's goal is to maximize the sum of rewards given by the norm-augmented reward function (Equation 2). We first assume that the agent is certain that a set of norms N holds, then discuss how agents can plan under uncertainty about which norms are true.

Even when an agent is certain about N, complying with all of them can be challenging due to temporally extended obligations. As such, we design our agents to use two different planning modes, namely *reward-oriented planning* and *obligation-oriented planning*. Each of these corresponds to a different sub-problem that is a *proxy* of the true game. Agents switch from the former mode to the latter mode whenever an obligation precondition is triggered.

In reward-oriented planning, agents maximize Equation 2 except with the cost of obligation norms ignored. Since prohibition norms are not temporally extended, this gives us a proxy reward $R_i^{N_P}(s, a_i, s')$ that only depends on the current state transition. We maximize this reward through model-based planning, assuming that each agent i has a reasonably accurate model $T_i(s'|s, a_i)$ of the environment and other agents. In particular, we use real-time dynamic programming (RTDP) [9] to compute estimates of the state and action value functions:

$$Q_i^{N_P}(s, a_i) = \sum_{s' \in \mathcal{S}} T_i(s'|s, a_i) \left[R_i^{N_P}(s, a_i, s') + \gamma V_i^{N_P}(s') \right] \quad (3)$$

$$V_i^{N_P}(s) = \max_{a_i \in \mathcal{A}_i} Q_i^{N_P}(s, a_i) \quad (4)$$

While reward-oriented planning accounts for prohibitions, it ignores the potential future costs that might arise due to obligations becoming active. Instead, our agents reactively switch to obligation-oriented planning once the precondition Pre of some obligation $\nu_O = \langle \text{Pre}, \text{Post}, \tau \rangle$ is met and then plan to achieve Post as a goal. This creates a (stochastic) shortest-path problem, where any state s with $s[\text{Post}] = 1$ is treated as a *terminal* state, with a modified reward function that accounts only for action costs C_{a_i}, prohibition costs $C_i(\nu_P)$ and delivers a reward of 1 when Post is satisfied:

$$R_i^{\nu_O \cup N_P}(s, a_i, s') = C_{a_i} + \sum_{\nu_P \in N_P} C_i(\nu_P)\nu_P(s, a, s') + s'[\text{Post}] \quad (5)$$

We similarly maximize this reward via model-based planning, giving us value estimates $Q_i^{\nu_O \cup N_P}$ and $V_i^{\nu_O \cup N_P}$ for each obligation ν_O that becomes active. Agents thus act by selecting actions that maximize either $Q_i^{N_P}$ or $Q_i^{\nu_O \cup N_P}$ depending on whether an obligation ν_O is active, performing several iterations of RTDP to update the value functions before acting. Obligations are satisfied in the order that their preconditions are triggered, and each agent maintains a queue of such obligations to satisfy.

Thus far, we have described how planning occurs when agents are *certain* about a set of norms N. In general, however, agents only have a *belief* $P_i^t(N)$ over norms. To handle this uncertainty, we consider two simple approaches, *thresholding* and *sampling*, which avoid the difficulty of taking expectations over all possible norm sets $N \subseteq \mathcal{N}$. When thresholding, agents comply with a norm ν whenever its probability is higher than a threshold $P_i^t(\nu \in N) \geq \theta$. When sampling, agents sample of set of norms $N' \sim P_i^t(N)$ to comply with for a certain number of steps K, before resampling. This is also known as probability matching [29] or Thompson sampling [76]. Each approach has properties that may be useful. Thresholding is low variance and may be appropriate when norm violation is sufficiently costly. In contrast, sampling can promote exploration, especially under high uncertainty about which norms are true [27].

2.4 Bayesian Norm Learning

As noted in Section 2.1, a posterior $P_i^t(N)$ over norms is defined with respect to a conditional distribution $P_i(h_t|N)$ over how other agents will act, which we left unspecified. To specify this distribution, we make two assumptions: (i) Agents model each other as norm-compliant planners, and (ii) they assume that other agents act *as if* they know the norms N with certainty. Since agents have full observability and can observe the actions a_{-i} that all other agents take at state s after history h. This induces the following posterior:

$$P_i^{t+1}(N|h, s, a_{-i}) \propto \frac{\pi_{-i}(a_{-i}|h, s, N)\ P_i^t(N)}{P_i^t(a_{-i}|h, s)} \quad (6)$$

The likelihood term $\pi_{-i}(a_{-i}|h, s, N)$ is the policy over actions that results from all other agents following the procedure in Section 2.3, conditional on all of them knowing that N is true. This means that actions that comply with a norm $\nu \in N$ are more likely to occur, and violations are less likely. Note, however, that this may not be an accurate model in general. After all, other agents might *also* be uncertain about the norms, and act based on their *beliefs* about which norms are true. Nonetheless, the assumption of shared normativity greatly reduces the complexity of norm learning: Similar to how assumptions of shared agency make decentralized cooperation easier [41, 83, 86], our agents avoid having to infer everyone else's beliefs about the norms, and directly infer the norms themselves.

Even with this assumption, inferring the exact posterior over N — the *rule induction* problem [33, 63] — is still highly intractable, since it requires enumerating over all subsets $N' \subseteq \mathcal{N}$ and updating their probabilities. Instead, we compute a *mean-field approximation* [85] of the posterior, assuming that it factorizes into approximate local posteriors $\tilde{P}^t(\nu|h, s, a_{-i})$ for each potential norm $\nu \in \mathcal{N}$:

$$P_i^{t+1}(N|h, s, a_{-i}) \propto \prod_{\nu \in \mathcal{N}} \tilde{P}^{t+1}(\nu|h, s, a_{-i}) \quad (7)$$

Making this approximation allows us to update beliefs about each norm ν *independently* via Bayes rule:

$$\tilde{P}^{t+1}(\nu|h, s, a_{-i}) \propto \pi_{-i}(a_{-i}|h, s, \nu)\tilde{P}_i^t(\nu) \quad (8)$$

where $\pi_{-i}(a_{-i}|h, s, \nu)$ approximates the probability that the other agents take actions $a_{-i,t}$ under the assumption that ν is the *only* norm that is true. We compute this probability by using the Q-values described in Section 2.3, assuming that each agent j selects its action a_j according to a Boltzmann distribution:

$$\pi_{-i}(a_j|h, s, \nu) \propto \exp(Q^{\nu}(s, a_j)) \quad (9)$$

By using the Q-values derived by planning, we automatically take into account the fact that other agents are unlikely to violate norms since such actions will have lower Q-values.

3 EXPERIMENTS

Having developed our model of norm-guided learning and planning, we investigate whether it reproduces the aspects of human normativity that our theory suggests. To do, we conducted experiment studying four key phenomena: (1) *Passive Norm Learning*: We explore the speed and effectiveness of learning norms via (approximately) Bayesian updating from passive observations [21], assuming that norms are shared by other agents. (2) *Norm-Enabled Social Outcomes*: We investigate whether our model of norm learning can foster collectively beneficial coordination [14, 30], given the right set of norms [12, 28, 68]. (3) *Intergenerational Norm Transmission*: We assess if norms can be sustained across generations of agents, and under what circumstances this occurs. (4) *Norm Emergence and Convergence*: We investigate whether our assumption of shared normativity enables norms to organically emerge and stabilize through Bayesian learning without explicit sanctions.

In all of our experiments, we used a sample size of 13 runs for each condition, determined based on an effect size of 0.8 with $\alpha = 0.05$ (one-sample t-test assuming a normal distribution).

3.1 Passive Norm Learning

In our norm learning experiments, we investigated whether a single learning agent l could acquire the same norms believed in by a set of $M_{Exp} = 3$ experienced agents, each playing one of three roles $z \in Z := \{\mathsf{C}, \mathsf{F}, \mathsf{E}\}$ (read as "cleaner", "farmer" and "egalitarian"). The learning agent also had an egalitarian role ($z_{Lea} = \mathsf{E}$), and each episode lasted $H = 300$ steps. All experienced agents assigned a probability of 1 to the norms $N_{\text{active}} = \{\mathsf{P1}, \mathsf{P2}, \mathsf{O1}, \mathsf{O2}, \mathsf{O3}\}$ shown in Figure 1b. The full space $\mathcal{N}$ of 68 possible norms was generated by enumerating over a variety of predicates and numeric thresholds (see Appendix). The learning agent handled uncertainty over norms via thresholding, and had a high norm violation cost, motivating it to comply with learned norms.

Figure 2 demonstrates the versatility and efficiency of our learning model across a diverse set of norms. The majority of norms practiced by the experienced agents were rapidly and effectively internalized within 300 steps, in contrast to the *millions* of steps required by model-free norm learning approaches [49]. This was despite some norms being complied with only occasionally (e.g. the cleaner's obligation to clean the dirty river), providing a limited learning signal for observers. Interestingly, Obligation 1 (concerning the farmer's payment) was *not* learned very well, with an average confidence of 0.4 even after 300 steps. This was likely due to the fact that the obligation was hard to satisfy, requiring the farmer to run after its assigned cleaner in order to pay them apples. As a result, the learning agent often inferred that the norm was *non-existent* rather than merely hard to satisfy. To address this, more sophisticated norm learners should distinguish unintentional failure to comply with a norm from the absence of a norm. After all, many human obligations can be quite hard to fulfill, but we still take them to be in force despite the difficulty of fulfilling them.

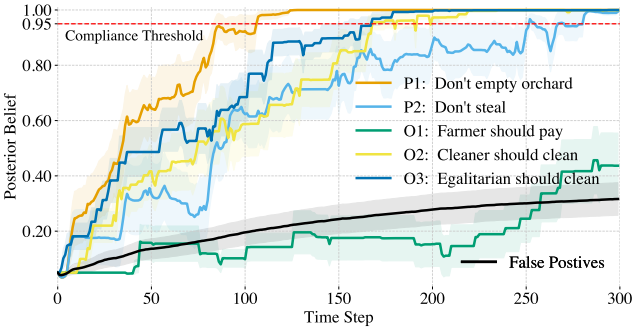


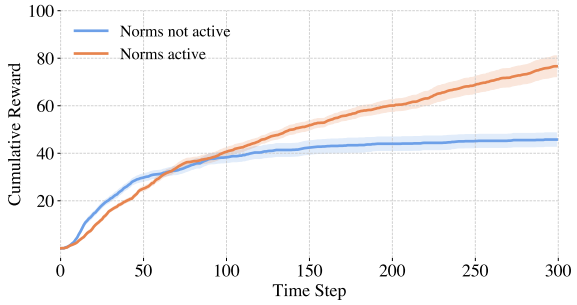
Figure 2: Passive Norm Learning. Average posterior belief of the learner in each of the five norms practiced by the experienced agents, N_{active}. Most norms were acquired within ≤ 300 steps.

$\lvert N_{\text{active}}\rvert$	$t/H = 25\%$ Pre/Rec	50% Pre/Rec	75% Pre/Rec	100% Pre/Rec
0	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
1	0.15/0.09	0.08/0.40	0.06/0.54	0.06/0.57
2	0.19/0.27	0.14/0.59	0.13/0.76	0.12/0.82
3	0.24/0.37	0.21/0.63	0.18/0.77	0.18/**0.83**
4	0.35/0.36	0.28/0.67	0.24/0.77	0.22/**0.83**
5	**0.47**/0.25	0.43/0.57	0.34/0.79	0.32/**0.83**

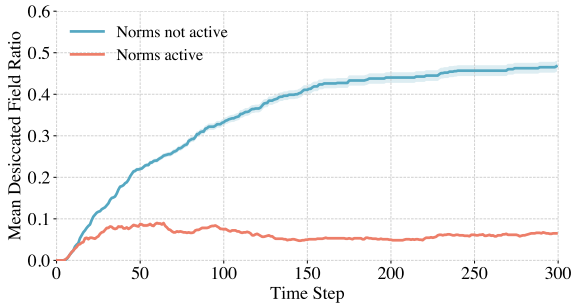
Table 1: Average Precision (Pre) and Recall (Rec) for the learner successfully acquiring ($P(v) \geq 0.95$) the active norms N_{active} after a certain % of the episode, where H denotes episode length.

Table 1 shows our model's learning curve over time for varying numbers of actively practiced norms, as reflected in the precision and recall metrics. While we observed a higher recall with more time steps, precision was low in most cases, indicating that the learner was "over-learning" norms. From qualitative inspections of each run, we found that this was due to several reasons: First, the learning agent sometimes captured the underlying dynamics of the environment. Even though norms were not explicitly followed, agents sometimes inferred them to be true if certain states of an environment were always avoided. For instance, agents learned to not be allowed to move when there was > 55% of dirt in the river which was a state that was *implicitly* avoided due to the river being cleaned on a regular basis. Such observations shed light on the phenomenon of overgeneralized norm learning in real-world contexts [65]: Similar to unintentionally failed obligations, there exist "accidentally fulfilled prohibitions", and sophisticated learners should learn to recognize and account for both.

Second, we noticed that agents often learned *logically overlapping* norms, reflecting the fact that some rules in $\mathcal{N}$ logically entailed other rules (e.g. if Dirt(r) > 0.4 is true, then Dirt(r) > 0.3 is also true). This behavior was likely due to the mean-field approximation of the posterior in Equation 7, which prevented *explaining away* among overlapping norms [96]. Future work should investigate how Bayesian norm learning can be made efficient without the mean-field approximation, perhaps via divide-and-conquer strategies [56] that first infer local posteriors over individual norms, then combine them into joint posteriors over sets of norms.



(a) Cumulative reward per set of practiced norms.



(b) Desiccated field ratio per set of practiced norms.

Figure 3: Social outcomes for two sets of norms N practiced by the experienced agents: either all norms in {P1, P2, O1, O2, O3} (*Norms active*) or none of them (*Norms not active*).

3.2 Norm-Enabled Social Outcomes

Norms are instrumental in steering pro-social behavior and cultivating cooperation within societies [26, 28, 37, 79], and past research has illustrated the efficacy of norm compliance in addressing social challenges and enabling favorable outcomes [20, 42, 54, 68, 72]. We investigated whether it was possible to achieve such outcomes with our norm learning approach, measuring social welfare in terms of the collective cumulative reward obtained by all agents [92] (hereafter *cumulative reward*) (Figure 3a) and sustainability of the environment [72] (Figure 3b), as measured by the ratio of desiccated apple fields in our environment relative to the total number of fields. While these were only two measures of desirable social outcomes (others might include equality, efficiency, etc. [72]), they nonetheless offered a useful measure of the benefit provided by norm-guided coordination. We conducted this analysis with the independent variable being the set N of norms, where $N \in \{\emptyset, \{P1, P2, O1, O2, O3\}\}$. All other experimental conditions were consistent with those described earlier: one learning agent, and three experienced ones.

Perhaps unsurprisingly, the presence of norms P1 through O3 led to a clear increase in social welfare (Figure 3). Specifically, agents tended to achieve greater collective rewards with these norms in our environment (Figure 3a). Unlike the scenario with no active norms, cumulative reward exhibited a consistent uptrend with time. Further evidence was provided by our sustainability metric: The percentage of desiccated fields (where apples do not regrow) increased rapidly in the absence of resource management norms, but remained at a low value when norms were present (Figure 3b).

Of course, not all norms are equally beneficial, and some can even be harmful [12]. As Figures A2 and A3 in the Appendix show, some of the norms we considered had a weaker effect on social outcomes. In particular, the prohibitions against eating apples in less dense areas appeared to have limited impact. This might have been because the potential sustainability gain from obeying the prohibition directly traded off against the welfare gain of eating more apples. In contrast, once the cleaning obligations were fulfilled, the environment became significantly better maintained, enabling more apples to grow, and causing cumulative reward to keep increasing.

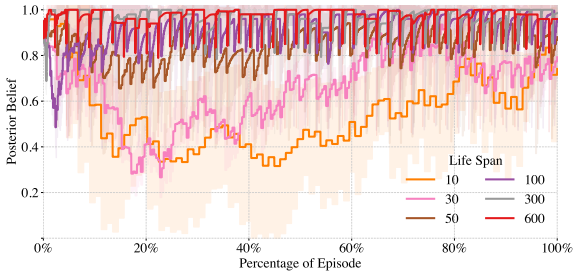
3.3 Intergenerational Norm Transmission

When agents have limited lifespans, or only join a community for fixed amounts of time, the stability of that community's norms requires the successful transmission of norms across generations [3, 32, 40, 84]. Since our account of norm learning depends on there being enough agents who *practice* a norm for common knowledge of the norm to be maintained, we predicted that norms would be sustained as long there were enough experienced agents to learn from over a reasonable period of time [10, 17, 89, 95]. For shorter lifespans, we predicted that social norms would be unstable, drifting from those originally practiced.

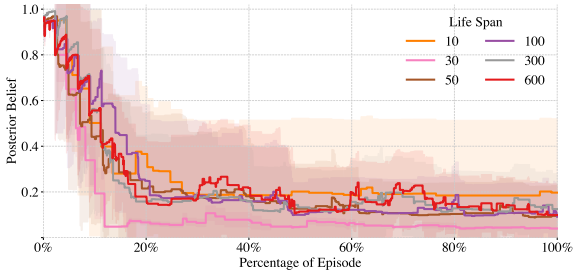
To test these hypotheses, we ran an experiment with $M = 6$ learning agents, while assigning each of the three role-associated appearances $z \in Z$ to two out of the six agents (effectively forming a parent-child pair for each role). Agents used sampling to select which norms to comply with. Each agent i had a maximum life span of L, and an initial age of $\frac{i}{M}L$. Each run was simulated for $1.5ML$ steps (i.e. three generational cycles). Upon an agent i's demise, it was replaced by an equivalent "child" agent with the same role-associated appearance z_i, but with a low uniform prior over norms $P_i(\nu) = 1 - \theta$ for all $\nu \in \mathcal{N}$. Initially, 3 experienced agents were present to seed the norms. As in earlier experiments, all agents were intrinsically motivated to comply with the learned norms.

The results of this experiment are shown in Figure 4, revealing interesting patterns of norm transmission contingent upon the life spans of agents and the types of norms. Starting with the prohibition against eating too many apples (Figure 4a), we observed striking rapidity in both its acquisition and transmission. Indeed, a life span as brief as 50 steps was sufficient for the effective transmission of these prohibition norms, aligning with the findings of our passive learning experiments (Figure 2). The amplification in learning speed can also be attributed to a larger number of agents to learn from (five instead of three), all adhering to the same prohibitive norm.

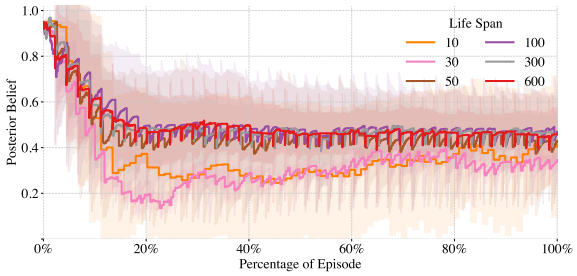
Obligation norms presented a more complex scenario (see e.g. Figure 4b). Due to the role-conditioned nature of the obligations we considered, learning opportunities were considerably more restricted compared to prohibitions. When learning the norms that applied to a particular role, an agent could only observe the two agents that were in that role, not all 6 agents. If they were the "child" agent in that role, they had only one agent to learn from (the "parent" in the same role). This resulted in a delayed learning curve for obligation norms, with agents requiring longer lifespans to learn and transmit them. Indeed, our results indicated that even an extended life span of 600 steps did not ensure that obligations were reliably transmitted.



(a) P1: Don't pick up apples, if there are too few apples around.



(b) O2: If you're a cleaner, clean the river when there is >30% dirt.



(c) Average for all five norms.

Figure 4: Intergenerational transmission of norms. Mean norm belief averaged across all six agents across generations.

Given the shorter learning time for obligations (300 steps) that we observed in Section 3.1, this inability to transmit obligations was unexpected. What we found instead was collective forgetting of all initially active obligations. Beyond the fact that obligations were less frequently observed, a possible factor was sampling: Since agents resampled the norms they would comply with every $K = 10$ steps, this added stochasticity that interfered with the long-horizon goals involved in obligation-oriented planning: If an obligation was not sampled at any step while fulfilling its goal, an agent would give up on the goal and go back to maximizing reward. This in turn created less of a learning signal for other agents.

Notwithstanding these unanticipated factors, our overall hypothesis about norm transmission was validated: Under *some* conditions, social norms are successfully maintained and transmitted (Figure 4c), and this is more likely to occur with more opportunities to learn those norms from agents who practice them. While more theoretical work is required, this suggests that in some regimes, norms are not only in a normative equilibrium but *evolutionarily stable* [82]. In contrast, when observations are scarce, norms might be abandoned [69], even by agents who used to comply with them.

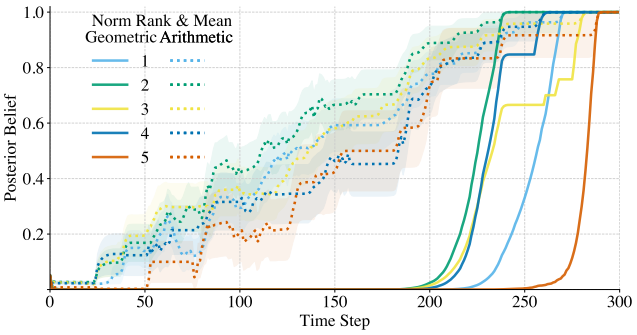


Figure 5: Norm emergence and convergence. Geometric and arithmetic mean posterior beliefs of all agents for the top 5 norms with final belief $P^{t=300}(n) \geq \theta = 0.95$.

3.4 Norm Emergence and Convergence

The emergence of norms is a complex phenomenon that has been studied both empirically [6, 78, 81, 92] and theoretically [40, 62, 91]. While prior work has studied norm emergence through adaption of policies in response to punishment [78], or learning to punish in accordance with public signals [92], our approach builds upon models of ad-hoc cooperation through joint intentionality [98], bootstrapping a shared set of norms in the same way that a team of cooperating agents can bootstrap a common goal without explicit communication [87]. Unlike the aforementioned mechanisms, this model of norm emergence does not require enforcement, as long as agents are sufficiently motivated (intrinsically or extrinsically) to comply with emergent norms.

To test whether such bootstrapping in fact occurs, we simulated a population of $M = 3$ learning agents with $z \in \{\mathsf{C}, \mathsf{E}, \mathsf{F}\}$ (every role represented once). As in our norm transmission experiments, agents select norms to comply with via sampling, effectively performing a form of *norm exploration* when uncertain about which norms hold.

Figure 5 summarizes the results of these experiments and demonstrates the emergence of shared norms among our agents. The gradual increase in the arithmetic mean over multiple runs signifies an emergent belief in the existence of certain norms, even though such beliefs may not initially be shared by all agents. This eventually results in a sharp rise in the geometric mean, demonstrating convergence and unanimous agreement among agents on specific norms. Through repeated interaction and learning, our agents coalesce around a set of norms they collectively perceive as true, illustrating how pure observational learning combined with experimental norm compliance can eventually lead to a stable set of norms.

Examining the emergent norms more closely (Table A2 in the Appendix), we found that many of the learned norms were non-functional, or "silly" norms [38]. This was likely due to intrinsic norm violation costs, which help to stabilize inefficient or even harmful norms [13]. As such, we expect that emergent norms are more likely to be beneficial when agents are more strategic about which norms they comply with. Nonetheless, recent work has shown how non-functional norms can play an epistemic role in stabilizing a normative system [38], and our model provides one mechanism for how such auxiliary norms might arise.

4 DISCUSSION AND FUTURE WORK

How should we build autonomous systems that learn to comply with the norms of their social environments? In this paper, we introduced a Bayesian framework for decentralized norm learning in long-horizon multi-agent settings, formalized as *norm-augmented Markov games* (NMGs). In these games, each agent pursues their own objectives while complying with the norms that they infer from collective behavior. In doing so, we show how agents can *rapidly acquire structured norms* by observing experienced agents, *sustain common knowledge* of these norms in a changing population, and *converge upon shared norms* even when none are initially present. These features not only explain important facets of human normativity, but may also prove crucial for autonomous agents.

Nonetheless, many conceptual and technical questions about norm learning remain unexplored by our experiments, or are not explicitly accounted for by our modeling framework. We review these questions in the following discussion, and consider how our framework could be extended to address them.

The Role of Sanctions in Learning and Sustaining Norms. Perhaps the most important difference between our experiments and previous work on decentralized norm learning [6, 78, 92] is the absence of *sanctions* — acts of punishment or approval — in motivating and sustaining norm compliance. Rather, our experiments focused on the *epistemic* aspect of norm learning, i.e., learning the *content* of normative rules, and ensuring they remain commonly known, leaving open the *motivational* aspect of norm learning. In part, this is because motivational learning is not as important for machines: We can often design them to *want* to comply with norms.

That said, nothing about the NMG formalism precludes the study of sanctions, which are important when we cannot control the motivations of other agents. Sanctioning could be introduced as a primitive action (e.g. the "zapping beams" supported by Melting Pot [1, 50]), a learned policy (e.g. taking others' apples in retaliation for theft), or a (meta)-norm that obligates punishment of agents who do not comply with norms. Sophisticated agents might then learn or plan to use sanctions against norm violators. Would-be violators — e.g. agents with no intrinsic violation costs — might learn to expect punishment for norm violation, thereby choosing to comply. In fact, by separating the epistemic and motivational aspects of norm learning, NMGs may deliver a clearer account of the role of sanctions: Sanctions provide both *evidence* that a norm exists, and *motivation* to comply with it. By adjusting their strengths, we can investigate their contributions to norm learning and stability. We could also disambiguate intrinsic motivation accounts of punishment [24] from more strategic accounts, where agents punish to signal support for existing norms [23, 36, 73], influencing the normative beliefs of other agents in ways that they deem beneficial.

Model-Free vs. Model-Based Norm Learning. Our paper takes a model-based approach to norm-oriented learning and planning, achieving orders of magnitude more sample efficiency than primarily model-free approaches [49, 92]. However, this comes at the cost of greater runtime complexity. To address these challenges, model-free and model-based approaches can be fruitfully combined. For example Bayesian learning can be accelerated via bottom-up proposals [103] and amortized inference [25, 94].

The question of model-free vs. model-based learning also relates to how norms are *represented*: Some norms — those that we tend to call *practices* or *rituals* [39] – are perhaps best represented as *habits* [70] or *skills*, which can be learned model-free [49]. Other norms are more naturally thought of as *abstract rules*, classifying behavior by its permissibility [93], while enabling public interpretability and generalization to new scenarios [37]. Our framework emphasizes the latter because we believe it is important for the conceptual richness of normative cognition: If norms lacked a symbolic, language-like structure, we would find it hard to compose, generalize, communicate, and argue about them [60].

While symbolic rules have become less popular in machine learning, recent advances in large language models (LLMs) suggest that they can made considerably more scalable. As methods like Constitutional AI [8] suggest, a norm or principle ν could be represented in natural language itself, using an LLM to *classify* whether behavior complies with the norm [64]. Alternatively, LLMs might *translate* natural language, such as legal documents, into formal specifications [19, 58], enabling grounded and reliable reasoning in a particular domain [67, 97, 102]. LLMs could thus serve as priors over norms $P_i^0(N)$ for a norm learner, or as approximate posteriors $P_i(N|u)$ over norms N given a piece of normative language u.

Symbolic norms do not eliminate the value of model-free learning, however: In general, learning how to *generate* norm-compliant behavior is harder than learning to *discriminate* it [99]. Oftentimes, it may be more efficient to learn habitual practices that comply with a norm, rather than planning them. Relatedly, the *motivation* to comply with norms might be learned via model-free mechanisms [22], even if norms are learned in a model-based manner.

Normative Reasoning and Norm Adaptation. In our experiments, agents passively learn existing norms, then comply with them once learned. Yet humans are significantly more flexible. We do not just learn norms, but also decide when it makes sense to follow them [48, 53], and adjudicate when they are fairly applied [35]. This is because we can *reason* about the function of norms [47] and how they affect each of us [79], adapting them to novel circumstances [62]. Yet given the complexity of reasoning about ideal norms, it is costly to do this all the time. Instead, if we are to build normatively competent agents, we should perhaps take a leaf from recent accounts of how humans manage this complexity [52]: defaulting to learned rules in tried-and-true situations, adjusting them in new scenarios, and forging new norms via agreement when it is feasible and worthwhile. By shifting fluidly between these levels of normative cognition, we might eventually build systems that not only *comply* with the norms of our society, but help us deliberate upon and *improve* them.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their thoughtful feedback as well as the Summer Research Fellowship for Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS) for funding parts of this project. Tan Zhi-Xuan is supported by the Open Philanthropy AI Fellowship.

SUPPLEMENTARY MATERIAL

Source code is available at *https://github.com/ninell-oldenburg/social-contracts*.

REFERENCES

[1] John P. Agapiou, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, D. J. Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z. Leibo. 2023. Melting Pot 2.0. http://arxiv.org/abs/2211.13746 arXiv:2211.13746 [cs] version: 3.

[2] Carlos E. Alchourrón. 1969. Logic of Norms and Logic of Normative Propositions. *Logique et Analyse* 12, 47 (1969), 242–268. https://www.jstor.org/stable/44083577 Publisher: Peeters Publishers.

[3] Joan Aldous and Reuben Hill. 1965. Social Cohesion, Lineage Type, and Intergenerational Transmission*. *Social Forces* 43, 4 (May 1965), 471–482. https://doi.org/10.2307/2574453

[4] Thomas Arnold and Daniel Kasenberg. 2017. Value Alignment or Misalignment: What Will Keep Systems Accountable?. In *AAAI Workshop on AI, Ethics, and Society*.

[5] Robert J. Aumann. 1987. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica* 55, 1 (1987), 1–18. https://doi.org/10.2307/1911154 Publisher: [Wiley, Econometric Society].

[6] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science* 211, 4489 (1981), 1390–1396.

[7] Alisabeth Ayars. 2016. Can model-free reinforcement learning explain deontological moral judgments? *Cognition* 150 (2016), 232–242.

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073* (2022).

[9] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72, 1 (Jan. 1995), 81–138. https://doi.org/10.1016/0004-3702(94)00011-O

[10] Vern Bengtson. 2018. *Global Aging and Challenges to Families*. Routledge.

[11] Cristina Bicchieri. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press. Google-Books-ID: 4N1FDIZvcI8C.

[12] Cristina Bicchieri and Yoshitaka Fukui. 1999. The great illusion: Ignorance, informational cascades, and the persistence of unpopular norms. In *Experience, Reality, and Scientific Explanation: Essays in Honor of Merrilee and Wesley Salmon*. Springer, 89–121.

[13] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. 2018. Social Norms. In *The Stanford Encyclopedia of Philosophy* (winter 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/social-norms/

[14] Ken Binmore and Larry Samuelson. 1994. An economist's perspective on the evolution of norms. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft* (1994), 45–63.

[15] Guido Boella and Leendert van der Torre. 2006. An architecture of a normative system: counts-as conditionals, obligations and permissions. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems (AAMAS '06)*. Association for Computing Machinery, New York, NY, USA, 229–231. https://doi.org/10.1145/1160633.1160671

[16] Guido Boella, Leendert van der Torre, and Harko Verhagen. 2006. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory* 12, 2 (Oct. 2006), 71–79. https://doi.org/10.1007/s10588-006-9537-7

[17] Bowles. 2006. Group Competition, Reproductive Leveling, and the Evolution of Human Altruism | Science. https://www.science.org/doi/abs/10.1126/science.1134829

[18] Michael E. Bratman. 2013. *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press. Google-Books-ID: jcs8BAAAQBAJ.

[19] John J Camilleri. 2017. *Contracts and Computation*. Doctoral. University of Gothenburg, Gothenburg.

[20] Phillip J. K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. 2022. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. https://doi.org/10.48550/arXiv.2208.10469 arXiv:2208.10469 [cs, econ].

[21] Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. 2016. A Bayesian Approach to Norm Identification. (2016).

[22] Fiery Cushman. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review* 17, 3 (2013), 273–292.

[23] Fiery Cushman, Arunima Sarin, and Mark Ho. 2019. Punishment as communication. *The Oxford handbook of moral psychology* (2019), 197–209.

[24] John M Darley, Kevin M Carlsmith, and Paul H Robinson. 2000. Incapacitation and just deserts as motives for punishment. *Law and human behavior* 24 (2000), 659–683.

[25] Ishita Dasgupta, Eric Schulz, Joshua B Tenenbaum, and Samuel J Gershman. 2020. A theory of learning to infer. *Psychological review* 127, 3 (2020), 412.

[26] Fred D'Agostino, Gerald Gaus, and John Thrasher. 2021. Contemporary Approaches to the Social Contract. In *The Stanford Encyclopedia of Philosophy* (winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/contractarianism-contemporary/

[27] Benjamin Eysenbach and Sergey Levine. 2019. If MaxEnt RL is the Answer, What is the Question? https://arxiv.org/abs/1910.01913v1

[28] Ernst Fehr and Ivo Schurtenberger. 2018. Normative foundations of human cooperation. *Nature Human Behaviour* 2, 7 (July 2018), 458–468. https://doi.org/10.1038/s41562-018-0385-5

[29] Wolfgang Gaissmaier and Lael J Schooler. 2008. The smart potential behind probability matching. *Cognition* 109, 3 (2008), 416–422.

[30] Herbert Gintis. 2010. Social norms as choreography. *Politics, Philosophy & Economics* 9, 3 (Aug. 2010), 251–264. https://doi.org/10.1177/1470594X09345474 Publisher: SAGE Publications.

[31] P. J. Gmytrasiewicz and P. Doshi. 2005. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research* 24 (July 2005), 49–79. https://doi.org/10.1613/jair.1579

[32] Roberto González, Belén Alvarez, Jorge Manzi, Micaela Varela, Cristián Frigolett, Andrew G. Livingstone, Winnifred Louis, Héctor Carvacho, Diego Castro, Manuel Cheyre, Marcela Cornejo, Gloria Jiménez-Moya, Carolina Rocha, and Daniel Valdenegro. 2021. The Role of Family in the Intergenerational Transmission of Collective Action. *Social Psychological and Personality Science* 12, 6 (Aug. 2021), 856–867. https://doi.org/10.1177/1948550620949378 Publisher: SAGE Publications Inc.

[33] Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. 2008. A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science* 32, 1 (2008), 108–154. https://doi.org/10.1080/03640210701802071 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1080/03640210701802071.

[34] Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. 2014. *Concepts in a probabilistic language of thought*. Technical Report. Center for Brains, Minds and Machines (CBMM).

[35] Gillian Kereldena Hadfield. 2017. *Rules for a flat world: Why humans invented law and how to reinvent it for a complex global economy*. Oxford University Press.

[36] Gillian K. Hadfield and Barry R. Weingast. 2012. What Is Law? A Coordination Model of the Characteristics of Legal Order. *Journal of Legal Analysis* 4, 2 (Dec. 2012), 471–514. https://doi.org/10.1093/jla/las008

[37] Gillian K. Hadfield and Barry R. Weingast. 2014. Microfoundations of the Rule of Law. *Annual Review of Political Science* 17, 1 (2014), 21–42. https://doi.org/10.1146/annurev-polisci-100711-135226 _eprint: https://doi.org/10.1146/annurev-polisci-100711-135226.

[38] Dylan Hadfield-Menell, McKane Andrus, and Gillian K. Hadfield. 2018. Legible Normativity for AI Alignment: The Value of Silly Rules. http://arxiv.org/abs/1811.01267 arXiv:1811.01267 [cs].

[39] Kurtis Hagen. 2010. The propriety of Confucius: A sense-of-ritual. *Asian Philosophy* 20, 1 (2010), 1–25.

[40] Robert X.D. Hawkins, Noah D. Goodman, and Robert L. Goldstone. 2019. The Emergence of Social Norms and Conventions. *Trends in Cognitive Sciences* 23, 2 (Feb. 2019), 158–169. https://doi.org/10.1016/j.tics.2018.11.003

[41] Mark K Ho, Amy Greenwald, Elizabeth M Hilliard, Stephen Brawner, and Max Kleiman-Weiner. 2016. Feature-based Joint Planning and Norm Learning in Collaborative Games. (2016).

[42] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html

[43] Marco A. Janssen, Robert Holahan, Allen Lee, and Elinor Ostrom. 2010. Lab Experiments for the Study of Social-Ecological Systems. *Science* 328, 5978 (April 2010), 613–617. https://doi.org/10.1126/science.1183532 Publisher: American Association for the Advancement of Science.

[44] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*. PMLR, 3040–3049.

[45] Ehud Kalai and Ehud Lehrer. 1995. Subjective games and equilibria. *Games and Economic Behavior* 8, 1 (1995), 123–163.

[46] Daniel Kasenberg and Matthias Scheutz. 2018. Inverse Norm Conflict Resolution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 178–183. https://doi.org/10.1145/3278721.3278775

[47] Lawrence Kohlberg and Richard H Hersh. 1977. Moral development: A review of the theory. *Theory into practice* 16, 2 (1977), 53–59.

[48] Joe Kwon, Tan Zhi-Xuan, Joshua Tenenbaum, and Sydney Levine. 2023. When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. 45, 45 (2023).

[49] Raphael Köster, Kevin R. McKee, Richard Everett, Laura Weidinger, William S. Isaac, Edward Hughes, Edgar A. Duéñez-Guzmán, Thore Graepel, Matthew Botvinick, and Joel Z. Leibo. 2020. Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. http://arxiv.org/abs/2010.09054 arXiv:2010.09054 [cs].

[50] Joel Z. Leibo, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. http://arxiv.org/abs/2107.06857 arXiv:2107.06857 [cs].

[51] Adam Lerer and Alexander Peysakhovich. 2019. Learning Existing Social Conventions via Observationally Augmented Self-Play. http://arxiv.org/abs/1806.10071 arXiv:1806.10071 [cs].

[52] Sydney Levine, Nick Chater, Joshua Tenenbaum, and Fiery Cushman. 2023. Resource-rational contractualism: A triple theory of moral cognition. https://doi.org/10.31234/osf.io/p48t7

[53] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Joshua Tenenbaum, and Fiery Cushman. 2020. The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences* 117, 42 (2020), 26158–26169.

[54] Kemas M. Lhaksmana, Yohei Murakami, and Toru Ishida. 2018. Role-Based Modeling for Designing Agent Behavior in Self-Organizing Multi-Agent Systems. *International Journal of Software Engineering and Knowledge Engineering* 28, 01 (Jan. 2018), 79–96. https://doi.org/10.1142/S0218194018500043 Publisher: World Scientific Publishing Co.

[55] Leon Li, Bari Britvan, and Michael Tomasello. 2021. Young children conform more to norms than to preferences. *Plos one* 16, 5 (2021), e0251228.

[56] Fredrik Lindsten, Adam M Johansen, Christian A Naesseth, Bonnie Kirkpatrick, Thomas B Schön, John AD Aston, and Alexandre Bouchard-Côté. 2017. Divide-and-conquer with sequential Monte Carlo. *Journal of Computational and Graphical Statistics* 26, 2 (2017), 445–458.

[57] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, William W. Cohen and Haym Hirsh (Eds.). Morgan Kaufmann, San Francisco (CA), 157–163. https://doi.org/10.1016/B978-1-55860-335-6.50027-1

[58] Jason Xinyu Liu, Ziyi Yang, Ifrah Idrees, Sam Liang, Benjamin Schornstein, Stefanie Tellex, and Ankit Shah. 2023. Lang2LTL: Translating Natural Language Commands to Temporal Robot Task Specification. *arXiv preprint arXiv:2302.11649* (2023).

[59] Qinghua Liu, Csaba Szepesvári, and Chi Jin. 2022. Sample-Efficient Reinforcement Learning of Partially Observable Markov Games. https://doi.org/10.48550/arXiv.2206.01315 arXiv:2206.01315 [cs, stat].

[60] John Mikhail. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences* 11, 4 (2007), 143–152.

[61] Adam Morris and Fiery Cushman. 2018. A common framework for theories of norm compliance. *Social Philosophy and Policy* 35, 1 (2018), 101–127.

[62] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm emergence in multiagent systems: a viewpoint paper. *Autonomous Agents and Multi-Agent Systems* 33, 6 (Nov. 2019), 706–749. https://doi.org/10.1007/s10458-019-09422-0

[63] Stephen Muggleton. 1994. Bayesian Inductive Logic Programming. In *Machine Learning Proceedings 1994*, William W. Cohen and Haym Hirsh (Eds.). Morgan Kaufmann, San Francisco (CA), 371–379. https://doi.org/10.1016/B978-1-55860-335-6.50052-0

[64] John J. Nay. 2022. Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans. https://doi.org/10.48550/arXiv.2209.13020 arXiv:2209.13020 [cs].

[65] Shaun Nichols, Shikhar Kumar, Theresa Lopez, Alisabeth Ayars, and Hoi-Yee Chan. 2016. Rational learners and moral rules. *Mind & Language* 31, 5 (2016), 530–554.

[66] Karine Nyborg, John M. Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W. Neil Adger, Kenneth J. Arrow, Scott Barrett, Stephen Carpenter, F. Stuart Chapin, Anne-Sophie Crépin, Gretchen Daily, Paul Ehrlich, Carl Folke, Wander Jager, Nils Kautsky, Simon A. Levin, Ole Jacob Madsen, Stephen Polasky, Marten Scheffer, Brian Walker, Elke U. Weber, James Wilen, Anastasios Xepapadeas, and Aart de Zeeuw. 2016. Social norms as solutions. *Science* 354, 6308 (Oct. 2016), 42–43. https://doi.org/10.1126/science.aaf8317 Publisher: American Association for the Advancement of Science.

[67] Theo X Olausson, Alex Gu, Benjamin Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164* (2023).

[68] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

[69] Diana Panke and Ulrich Petersohn. 2012. Why international norms disappear sometimes. *European journal of international relations* 18, 4 (2012), 719–742.

[70] Wolfgang M Pauli, Jeffrey Cockburn, Eva R Pool, Omar D Perez, and John P O'Doherty. 2018. Computational approaches to habits in a model-free world. *Current Opinion in Behavioral Sciences* 20 (2018), 104–109.

[71] Steven T Piantadosi and Robert A Jacobs. 2016. Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science* 25, 1 (2016), 54–59.

[72] Julien Pérolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/2b0f658cbffd284984fb11d90254081f-Abstract.html

[73] Setayesh Radkani, Josh Tenenbaum, and Rebecca Saxe. 2022. Modeling punishment as a rational communicative social action. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 44.

[74] Hannes Rakoczy, Nina Brosche, Felix Warneken, and Michael Tomasello. 2009. Young children's understanding of the context-relativity of normative rules in conventional games. *British Journal of Developmental Psychology* 27, 2 (2009), 445–456. https://doi.org/10.1348/026151008X337752 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/026151008X337752.

[75] Hannes Rakoczy, Felix Warneken, and Michael Tomasello. 2008. The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology* 44, 3 (2008), 875–881. https://psycnet.apa.org/doiLanding?doi=10.1037%2F0012-1649.44.3.875

[76] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2018. A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning* 11, 1 (July 2018), 1–96. https://doi.org/10.1561/2200000070 Publisher: Now Publishers, Inc.

[77] Bastin Tony Roy Savarimuthu, Stephen Cranefield, Maryam A. Purvis, and Martin K. Purvis. 2013. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law* 21, 1 (March 2013), 1–46. https://doi.org/10.1007/s10506-012-9126-7

[78] Bastin Tony Roy Savarimuthu, Stephen Cranefield, Martin K. Purvis, and Maryam A. Purvis. 2009. Norm emergence in agent societies formed by dynamically changing networks. *Web Intelligence and Agent Systems: An International Journal* 7, 3 (Jan. 2009), 223–232. https://doi.org/10.3233/WIA-2009-0164 Publisher: IOS Press.

[79] T. M. Scanlon. 2000. *What We Owe to Each Other*. Harvard University Press. Google-Books-ID: 9OPsDwAAQBAJ.

[80] Marco F.H. Schmidt, Hannes Rakoczy, and Michael Tomasello. 2011. Young children attribute normativity to novel actions without pedagogy or normative language. *Developmental Science* 14, 3 (2011), 530–539. https://doi.org/10.1111/j.1467-7687.2010.01000.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2010.01000.x.

[81] Ryosuke Shibusawa and Toshiharu Sugawara. 2014. Norm Emergence via Influential Weight Propagation in Complex Networks. *2014 European Network Intelligence Conference* (Sept. 2014), 30–37. https://doi.org/10.1109/ENIC.2014.28 Conference Name: 2014 European Network Intelligence Conference (ENIC) ISBN: 9781479969142 Place: Wroclaw, Poland Publisher: IEEE.

[82] John Maynard Smith. 1982. *Evolution and the Theory of Games*. Cambridge University Press.

[83] Stephanie Stacy. 2022. *The Imagined We: Shared Bayesian Theory of Mind for Modeling Communication*. Ph.D. Dissertation. Los Angeles. https://www.proquest.com/openview/795eae7dc98cb5364a1643c68c56481d/1?pq-origsite=gscholar&cbl=18750&diss=y

[84] Kim-Pong Tam. 2015. Understanding Intergenerational Cultural Transmission Through the Role of Perceived Norms. *Journal of Cross-Cultural Psychology* 46, 10 (Nov. 2015), 1260–1266. https://doi.org/10.1177/0022022115600074 Publisher: SAGE Publications Inc.

[85] Toshiyuki Tanaka. 1998. A Theory of Mean Field Approximation. In *Advances in Neural Information Processing Systems*, Vol. 11. MIT Press. https://proceedings.neurips.cc/paper_files/paper/1998/hash/a368b0de8b91cfb3f91892fbf1ebd4b2-Abstract.html

[86] Ning Tang, Stephanie Stacy, MingLu Zhao, Gabriel Marquez, and Tao Gao. 2011. Bootstrapping an Imagined We for Cooperation. (2011).

[87] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. 2020. Bootstrapping an Imagined We for Cooperation.. In *CogSci*.

[88] Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. *Developmental science* 10, 1 (2007), 121–125.

[89] Gisela Trommsdorff. 2005. Parent–Child Relations Over the Lifespan: A Cross-Cultural Perspective. In *Parenting Beliefs, Behaviors, and Parent-Child Relations*. Psychology Press. Num Pages: 42.

[90] Raimo Tuomela. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, Stanford, Calif.

[91] Edna Ullmann-Margalit. 2015. *The emergence of norms*. OUP Oxford.

[92] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence* 2, 2 (April 2023), 26339137231162025. https://doi.org/10.1177/26339137231162025 Publisher: SAGE Publications.

[93] Georg Henrik Von Wright. 1981. On the logic of norms and actions. In *New studies in deontic logic: Norms, actions, and the foundations of ethics*. Springer, 3–35.

[94] Tongzhou Wang, Yi Wu, Dave Moore, and Stuart J Russell. 2018. Meta-learning MCMC proposals. *Advances in neural information processing systems* 31 (2018).

[95] Richard A. Watson and Eörs Szathmáry. 2016. How Can Evolution Learn? *Trends in Ecology & Evolution* 31, 2 (Feb. 2016), 147–157. https://doi.org/10.1016/j.tree.

2015.11.009 Publisher: Elsevier.

[96] Michael P Wellman and Max Henrion. 1993. Explaining'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15, 3 (1993), 287–292.

[97] Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought. http://arxiv.org/abs/2306.12672 arXiv:2306.12672 [cs].

[98] Sarah A. Wu, Rose E. Wang, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. 2021. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Topics in Cognitive Science* 13, 2 (April 2021), 414–432. https://doi.org/10.1111/tops.12525

[99] Ying Nian Wu, Ruiqi Gao, Tian Han, and Song-Chun Zhu. 2019. A tale of three probabilistic families: Discriminative, descriptive, and generative models. *Quart. Appl. Math.* 77, 2 (2019), 423–465.

[100] Tan Zhi-Xuan. 2022. What Should AI Owe To Us? Accountable and Aligned AI Systems via Contractualist AI Alignment. (2022). https://www.lesswrong.com/posts/Cty2rSMut483QgBQ2/what-should-ai-owe-to-us-accountable-and-aligned-ai-systems

[101] Tan Zhi-Xuan, Jake Brawer, and Brian Scassellati. 2019. That's Mine! Learning Ownership Relations and Norms for Robots. http://arxiv.org/abs/1812.02576 arXiv:1812.02576 [cs].

[102] Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. 2024. Pragmatic Instruction Following and Goal Assistance via Cooperative Language Guided Inverse Plan Search. In *Proceedings of the 23rd International Conference on Autonomous Agents and MultiAgent Systems*.

[103] Song-Chun Zhu, Rong Zhang, and Zhuowen Tu. 2000. Integrating bottom-up/top-down for object recognition by data driven Markov chain Monte Carlo. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, Vol. 1. IEEE, 738–745.

APPENDIX

A1 Model and Environment Parameters

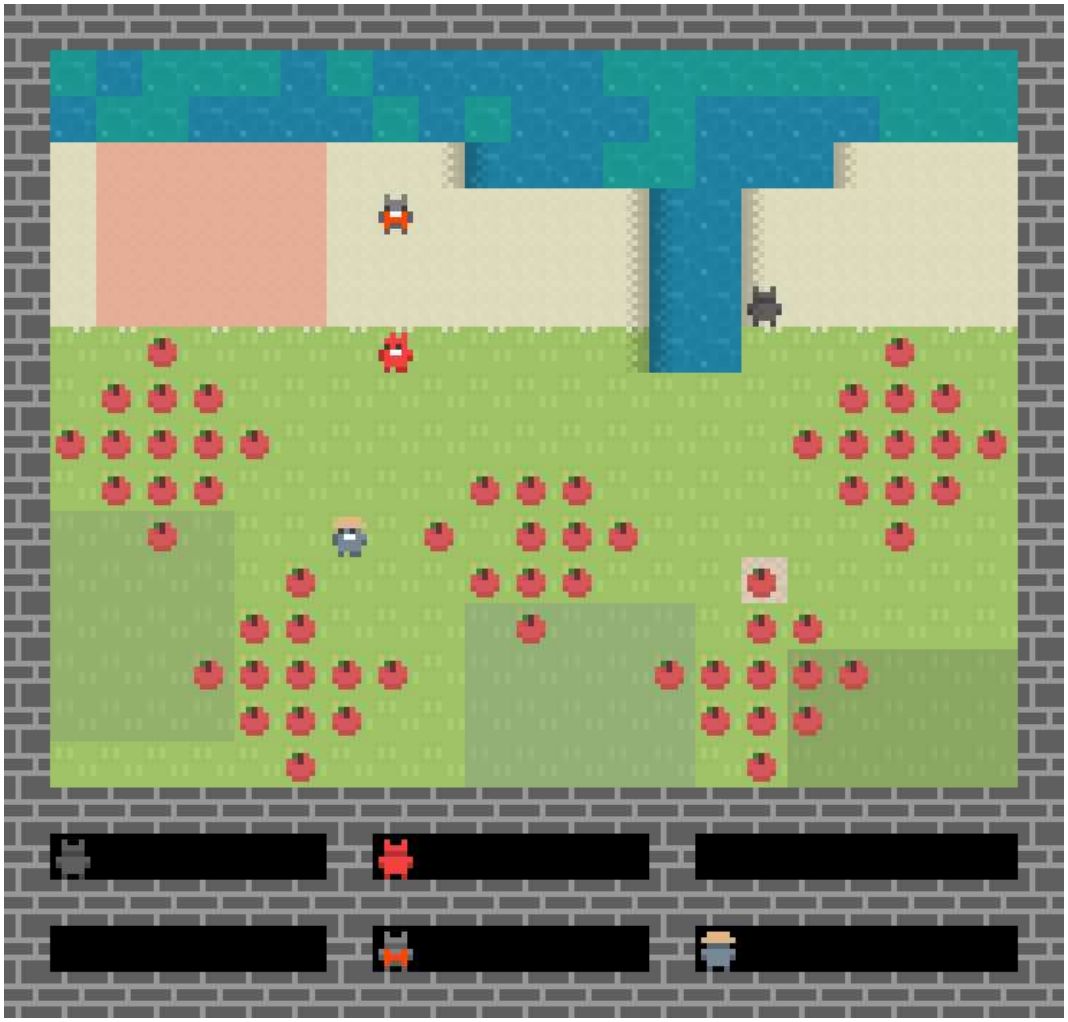


Figure A1: Example frame of our environment. The three experienced agents are depicted in grey each representing a specific role based on their appearance. The learning agent is completely colored, mimicking the “egalitarian" role based on its appearance. Each agent’s territory is visible as an agent-colored overlay. The river is located on the top of the frame with brighter parts representing the dirt. Apples are located in orchards that might, in turn, be in part on an agent’s territory. If a field is desiccated and hence unlikely but not impossible to regrow apples, it is signified by a sand-like color underneath.

Parameters for the Norm-augmented Markov Game. The individual reward for eating an apple for every agent is 1 while the action cost is 0.01. Each prohibition norm violation results in a violation cost of 1 that is applied to the value function as described in Equation 4 while, as mentioned above, the reward for the fulfillment of an obligation is 1, as specified in Equation 5. Our specific environment’s prohibition and obligation norms are specified in Figure 1b.

Parameters for the Norm-compliant Planning Algorithm. We use real-time dynamic programming (RTDP) [9] for the norm-compliant planning algorithm with a rollout depth of 20 every 2 steps and a discount factor of $\gamma = 0.9$. Agents thereby comply with the norms at threshold $\theta = 0.95$ or a sampling frequency of $K = 10$ for the intergenerational norm learning setting (Section 3.3) and $K = 15$ for the norm emergence setting (Section 3.4) to secure more robust emergence possibilities.

Parameters for the Norm-learning Algorithm. Every potential norm out of the 68 generated ones (see Table A1) was given an initial prior of $P(\nu) = 0.05$. We employed a Boltzmann temperature of 1.

Table A1: Generated list of feasible and non-trivial potential norms $\mathcal{N}$ up to a length of three parameters. (e.g. norms 24-31). The parameters are variables of the agents' observation space. The norms are feasible in the sense that the state the parameters specify is reachable and non-trivial in the sense that the state is non-tautological.

Prohibitions

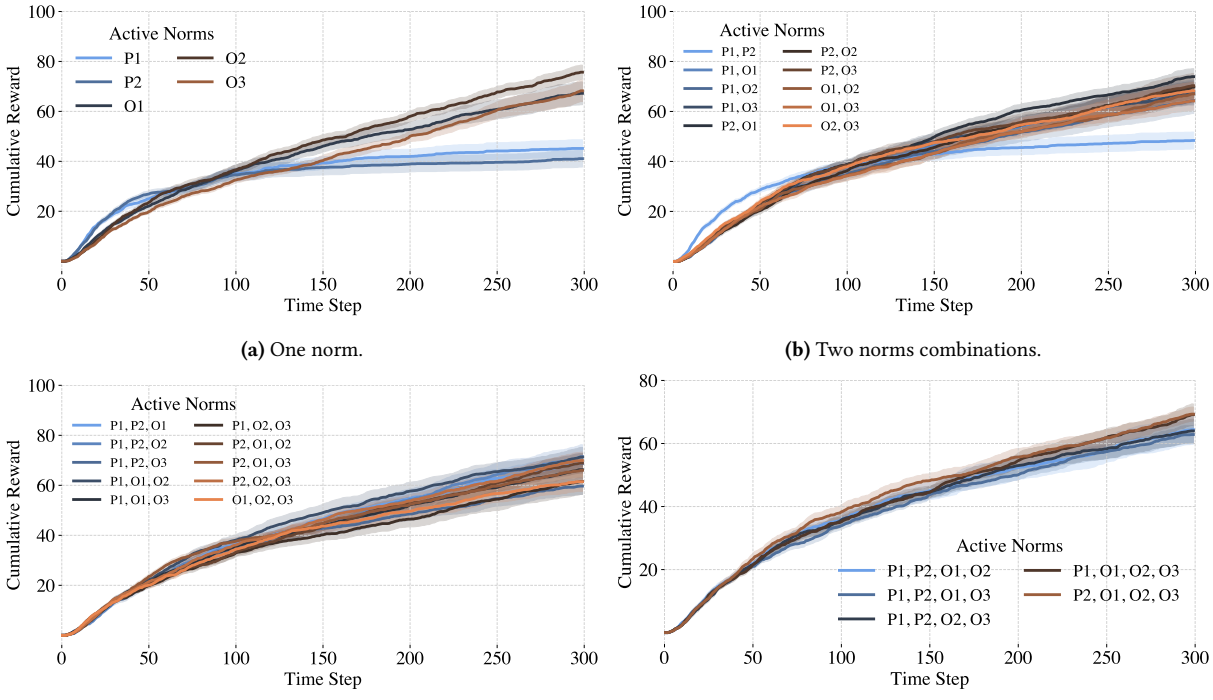
1 If a cell holds an apple → Don't move there.
2 If a cell is not your property → Don't move there.
3 If the river dirt fraction > 0.3 → Don't move at all.
4 If the river dirt fraction > 0.35 → Don't move at all.
5 If the river dirt fraction > 0.4 → Don't move at all.
6 If the river dirt fraction > 0.45 → Don't move at all.
7 If the river dirt fraction > 0.5 → Don't move at all.
8 If the river dirt fraction > 0.55 → Don't move at all.
9 If the river dirt fraction > 0.6 → Don't move at all.
10 If you're looking towards North → Don't move at all.
11 If you're looking towards East → Don't move at all.
12 If you're looking towards South → Don't move at all.
13 If you're looking towards West → Don't move at all.
14 If a cell holds an apple and it's not your property → Don't move there.
15 If a cell holds an apple and the number of apples around < 1 → Don't move there.
16 If a cell holds an apple and the number of apples around < 2 → Don't move there.
17 If a cell holds an apple and the number of apples around < 3 → Don't move there.
18 If a cell holds an apple and the number of apples around < 4 → Don't move there.
19 If a cell holds an apple and the number of apples around < 5 → Don't move there.
20 If a cell holds an apple and the number of apples around < 6 → Don't move there.
21 If a cell holds an apple and the number of apples around < 7 → Don't move there.
22 If a cell holds an apple and the number of apples around < 8 → Don't move there.
23 If a cell holds an apple and the number of apples around < 8 → Don't move there.
24 If a cell holds an apple and it's not your property and the number of apples around < 1 → Don't move there.
25 If a cell holds an apple and it's not your property and the number of apples around < 2 → Don't move there.
26 If a cell holds an apple and it's not your property and the number of apples around < 3 → Don't move there.
27 If a cell holds an apple and it's not your property and the number of apples around < 4 → Don't move there.
28 If a cell holds an apple and it's not your property and the number of apples around < 5 → Don't move there.
29 If a cell holds an apple and it's not your property and the number of apples around < 6 → Don't move there.
30 If a cell holds an apple and it's not your property and the number of apples around < 7 → Don't move there.
31 If a cell holds an apple and it's not your property and the number of apples around < 8 → Don't move there.

Obligations

32 If the river dirt fraction > 0.3 and you look like a cleaner → Clean the river.
33 If the river dirt fraction > 0.3 and you look like a farmer → Clean the river.
34 If the river dirt fraction > 0.3 and you look like an egalitarian → Clean the river.
35 If the river dirt fraction > 0.35 and you look like a cleaner → Clean the river.
36 If the river dirt fraction > 0.35 and you look like a farmer → Clean the river.
37 If the river dirt fraction > 0.35 and you look like an egalitarian → Clean the river.
38 If the river dirt fraction > 0.4 and you look like a cleaner → Clean the river.
39 If the river dirt fraction > 0.4 and you look like a farmer → Clean the river.
40 If the river dirt fraction > 0.4 and you look like an egalitarian → Clean the river.
41 If the river dirt fraction > 0.45 and you look like a cleaner → Clean the river.
42 If the river dirt fraction > 0.45 and you look like a farmer → Clean the river.
43 If the river dirt fraction > 0.45 and you look like an egalitarian → Clean the river.
44 If the river dirt fraction > 0.5 and you look like a cleaner → Clean the river.
45 If the river dirt fraction > 0.5 and you look like a farmer → Clean the river.
46 If the river dirt fraction > 0.5 and you look like an egalitarian → Clean the river.
47 If the river dirt fraction > 0.55 and you look like a cleaner → Clean the river.
48 If the river dirt fraction > 0.55 and you look like a farmer → Clean the river.
49 If the river dirt fraction > 0.55 and you look like an egalitarian → Clean the river.

If the river dirt fraction > 0.6 and you look like a cleaner → Clean the river.
If the river dirt fraction > 0.6 and you look like a farmer → Clean the river.
If the river dirt fraction > 0.6 and you look like an egalitarian → Clean the river.
If the number of steps you haven't paid anyone > 10 and you look like a cleaner → Pay someone.
If the number of steps you haven't paid anyone > 10 and you look like a farmer → Pay someone.
If the number of steps you haven't paid anyone > 10 and you look like an egalitarian → Pay someone.
If the number of steps you haven't paid anyone > 15 and you look like a cleaner → Pay someone.
If the number of steps you haven't paid anyone > 15 and you look like a farmer → Pay someone.
If the number of steps you haven't paid anyone > 15 and you look like an egalitarian → Pay someone.
If the number of steps you haven't paid anyone > 20 and you look like a cleaner → Pay someone.
If the number of steps you haven't paid anyone > 20 and you look like a farmer → Pay someone.
If the number of steps you haven't paid anyone > 20 and you look like an egalitarian → Pay someone.
If the number of steps you haven't paid anyone > 25 and you look like a cleaner → Pay someone.
If the number of steps you haven't paid anyone > 25 and you look like a farmer → Pay someone.
If the number of steps you haven't paid anyone > 25 and you look like an egalitarian → Pay someone.
If the number of steps you haven't paid anyone > 30 and you look like a cleaner → Pay someone.
If the number of steps you haven't paid anyone > 30 and you look like a farmer → Pay someone.
If the number of steps you haven't paid anyone > 30 and you look like an egalitarian → Pay someone.
If another agent violated a norms → Sanction them.

A2 Additional Results



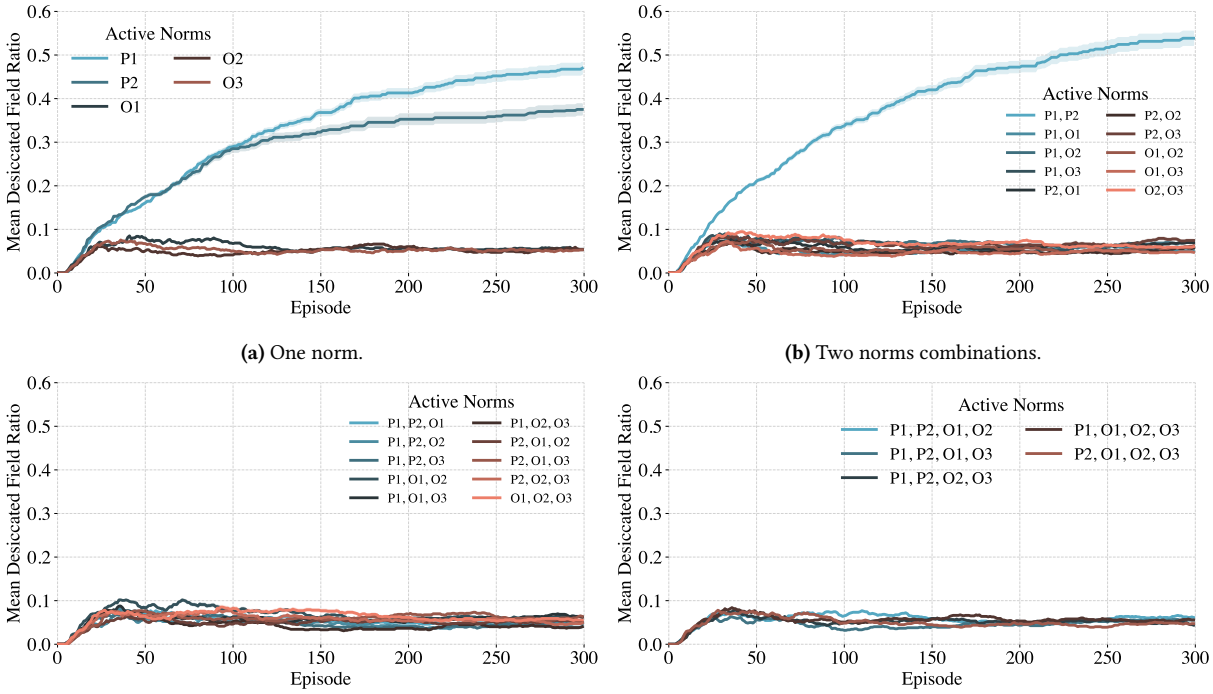
(a) One norm.

(b) Two norms combinations.

(c) Three norms combinations.

(d) Four norms combinations.

Figure A2: Cumulative reward per sets of practiced norms.



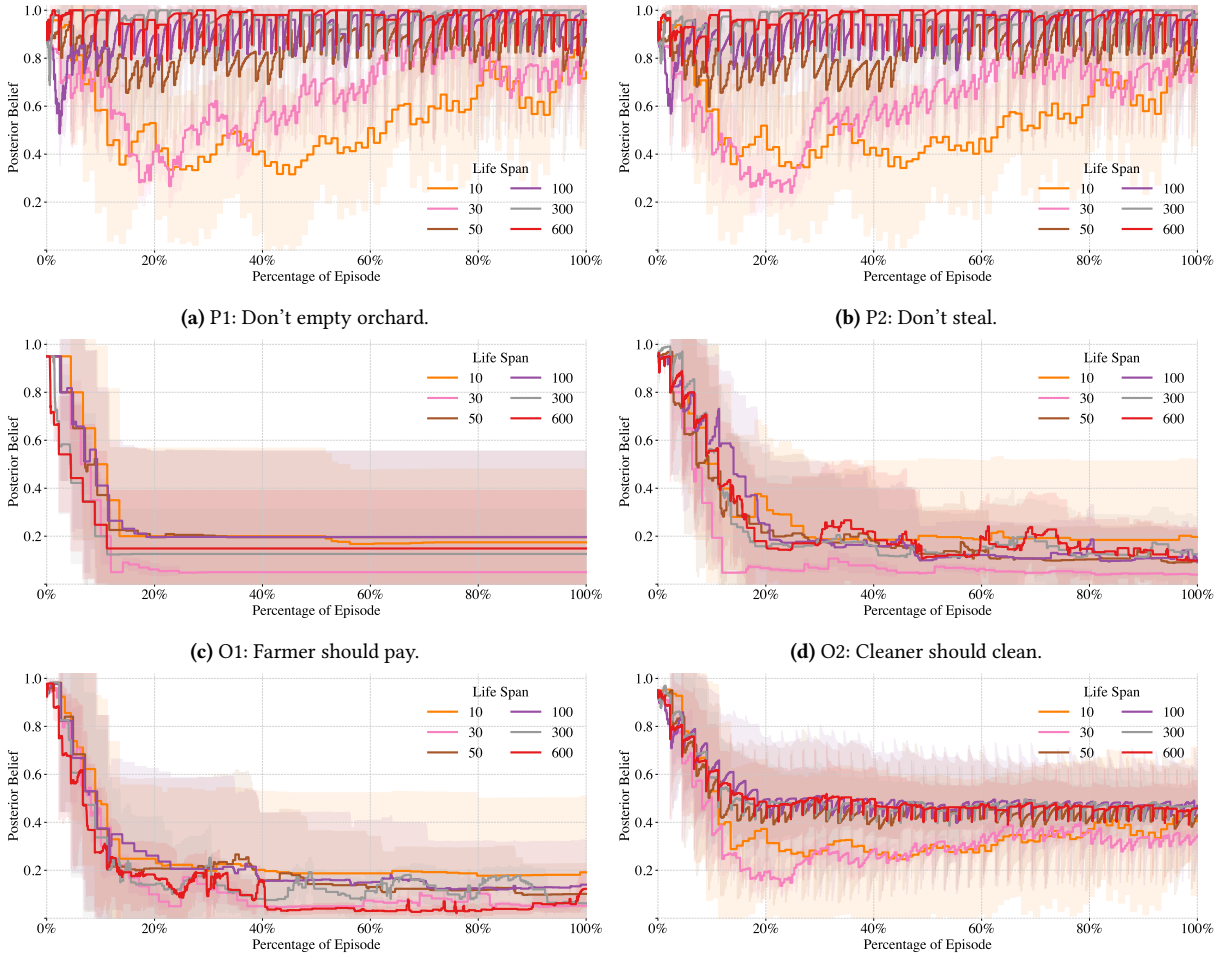
(a) One norm.

(b) Two norms combinations.

(c) Three norms combinations.

(d) Four norms combinations.

Figure A3: Desiccated field ratio per sets of practiced norms.



(a) P1: Don't empty orchard.

(b) P2: Don't steal.

(c) O1: Farmer should pay.

(d) O2: Cleaner should clean.

(e) O3: Egalitarian should clean.

(f) Average

Figure A4: Intergenerational transmission of norms. Mean norm belief averaged across all six agents across generations. Figures **(a)-(e)** shows the single norms as described in Figure1b. Figure **(f)** is the average of **(a)-(e)**.

Table A2: Top 10 spontaneously emerging norms and percentage of their occurrence, averaged over 13 runs.

	Norm	Occurs
1	If the river dirt fraction > 0.55 → Don't move.	0.61%
2	If the river dirt fraction > 0.6 → Don't move.	0.61%
3	If the river dirt fraction > 0.3 and you look like an egalitarian → Clean the river.	0.61%
4	If you're looking towards East → Don't move.	0.46%
5	If a cell holds an apple and the number of apples around < 1 → Don't move there.	0.38%
6	If it's not your property → Don't move there.	0.31%
7	If a cell holds an apple and it's not your property → Don't move there.	0.31%
8	If a cell holds an apple and the number of apples around < 1 → Don't move there.	0.31%
9	If you're looking towards West → Don't move.	0.31%
10	If the river dirt fraction > 0.55 → Don't move.	0.23%