Cultural reinforcement learning: a framework for modeling cumulative culture on a limited channel

Ben Prystawski[†], **Dilip Arumugam**[‡], **Noah D. Goodman**[†‡]
[†]Department of Psychology, Stanford University
[‡]Department of Computer Science, Stanford University
{benpry,dilip,ngoodman}@stanford.edu

Abstract

Humans' capacity for cumulative culture is remarkable: we can build up vast bodies of knowledge over generations. Communication, particularly via language, is a key component of this process. Previous work has described language as enabling posterior passing, where one Bayesian agent transmits a posterior distribution to the next. In practice, we cannot exactly copy our beliefs into the minds of others—we must communicate over the limited channel language provides. In this paper, we analyze cumulative culture as Bayesian reinforcement learning with communication over a rate-limited channel. We implement an agent that solves a crafting task and communicates to the next agent by approximating the optimal rate-distortion trade-off. Our model produces documented effects, such as the benefits of abstraction and selective social learning. It also suggests a new hypothesis: selective social learning can be harmful in tasks where initial exploration is required.

Keywords: cumulative culture; cultural transmission; rate-distortion theory; reinforcement learning

Introduction

As humans living in modern society, we have access to an enormous body of knowledge that stretches back millennia. Our most distinctive attribute as a species is our capacity to learn complex skills and accumulate knowledge over generations (Tomasello et al., 1993; Tomasello, 1999; Tennie et al., 2009; Boyd et al., 2011; Mesoudi & Thornton, 2018). This capacity relies on our unique ability for and interest in teaching, social learning, and collaboration. We can learn skills, knowledge, and ideas from other humans rather than needing to learn them directly from experience. Culture enables humans to distribute computation in a way that lets us outperform what any one individual is capable of (Smaldino & Richerson, 2013; Krafft et al., 2021). A key problem in cognitive science, then, is understanding what specific cognitive mechanisms enable cumulative culture. Studying culture experimentally can help us solve this problem.

Experimental studies of culture have generally used a transmission chain design, where participants work on a task and pass messages to each other in an iterated manner, to identify the properties that enable cumulative culture (Beppu & Griffiths, 2009; Derex et al., 2013; Tessler et al., 2021; Brinkmann et al., 2022; B. Thompson et al., 2022). Beppu & Griffiths (2009) showed that chains of participants working on a function learning task can converge to learning the true function if they can pass messages in language, but not if they can only pass example points. Their results support the idea that language enables "posterior passing", where one participant's posterior distribution over functions becomes the next participant's prior. Over generations, this is equivalent to Bayesian inference given all the evidence seen in all generations. The ability of language to grow useful knowledge is not unique to function learning: Tessler et al. (2021) showed that chains of people leaving messages for each other in language can learn rules and strategies for complex games.

Posterior passing is an elegant stylized model of cultural learning, but it assumes that language enables teachers to perfectly copy their beliefs into the minds of their learners. This assumption does not hold in general: language is often ambiguous and much of our knowledge is hard to explain. Recognition of these difficulties has motivated the framing of linguistic communication as message passing over a noisy or limited channel (e.g. Levy, 2008; Gibson et al., 2013; Futrell & Levy, 2017). Tools from information theory, such as rate-distortion theory and optimal transport, have been used to model how people transmit beliefs in language (Zaslavsky et al., 2020; Shafto et al., 2021).

Beyond *how* we learn from others, *who* we choose to learn from can be important for cultural learning. B. Thompson et al. (2022) showed that selective social learning, the tendency of people to learn disproportionately from successful others, enables populations of learners to maintain strategies that are more complex yet more effective than those they could maintain when learning from random others.

In this paper, we introduce *cultural reinforcement learning*, a framework that connects rate-distortion models of communication with posterior passing. We model cultural learning as Bayesian reinforcement learning (RL) constrained by a rate-limited communication channel. We then present a crafting task suitable for studying cultural learning and describe a specific model that solves the task using the cultural RL framework. Our model captures empirical results from the literature, including the benefits of generalization and selective social learning. It also informs a new hypothesis: in tasks with explore-exploit trade-offs, selective social learning can actually be harmful as higher scores indicate less learning.

Cultural Reinforcement Learning

Agents in transmission chains must learn to perform a task well with only a few timesteps of experience. However, they can receive knowledge from an earlier agent and pass knowledge to a future agent. We can view these chains as maximizing the performance of their members, just as an individual agent would if it had many generations worth of time. The crucial difference is that members of a chain must communicate with each other. If we view inter-generational communication as passing information over a rate-limited channel, then we can frame cultural learning as Bayesian RL constrained by rate-limited posterior passing. We term this setup cultural reinforcement learning (or CultuRL for short).

Bayesian RL

Formally, we represent a culturally learned task as a finite-horizon, episodic Markov Decision Process (MDP) (Bellman, 1957; Puterman, 1994) defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle$. Letting $[H] = \{1, 2, \dots, H\}$ denote the index set for the maximum episode duration or horizon $H \in \mathbb{N}$, the goal of reinforcement learning is to learn a non-stationary policy $\pi : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$ which maps from a state and current timestep to a distribution over actions in a way that maximizes the sum of cumulative rewards or return over the course of $K \in \mathbb{N}$ episodes or generations. We define the value function induced by executing policy π in MDP $\mathcal{M}$ starting from state $s \in \mathcal{S}$ as $V_{\mathcal{M}}^{\pi}(s) = \mathbb{E}\left[\sum\limits_{h=1}^{H} \mathcal{R}(s_h, a_h) \mid s_1 = s\right]$, where the expectation accounts for randomness in the action selections, the transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is assumed to be deterministic, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$. An optimal policy $\pi^{\star}$ is defined as achieving maximal value across all H timesteps $V_{\mathcal{M}}^{*}(s) = \max\limits_{\pi \in \{\mathcal{S} \times [H] \to \Delta(\mathcal{A})\}} V_{\mathcal{M}}^{\pi}(s)$. We write any value function without its argument to implicitly average over the initial state distribution: $V_{\mathcal{M}}^{\star} \triangleq \mathbb{E}_{s_1 \sim \beta}\left[V_{\mathcal{M}}^{\star}(s_1)\right]$.

The performance of a reinforcement learning agent is measured by its discounted regret: the difference between the optimal return and the actual return obtained by the policy $\pi^{(k)}$ used in the kth episode, summed up over all K episodes and weighted by a discount factor $\gamma \in [0, 1]$ that prioritizes near-term over short-term performance.

$$\textsc{Regret}(K, \{\pi^{(k)}\}_{k \in [K]}, \gamma, \mathcal{M}) = \sum_{k=1}^{K} \gamma^{k-1} \left(V_{\mathcal{M}}^{\star} - V_{\mathcal{M}}^{\pi^{(k)}} \right)$$

While this discounted notion of regret has appeared in prior work (Russo & Van Roy, 2022), the more traditional definition is obtained when $\gamma = 1$.

If the underlying transition function $\mathcal{T}$ and reward function $\mathcal{R}$ were fully known, the agent could simply use a planning algorithm (Bertsekas, 1995) and there would be no uncertainty in $\pi^{\star}$. The agent's uncertainty in the underlying MDP model $\mathcal{M} = (\mathcal{T}, \mathcal{R})$ drives its uncertainty in optimal behavior. The Bayesian RL setting (Bellman & Kalaba, 1959; Duff, 2002; Ghavamzadeh et al., 2015) represents the agent's initial uncertainty with a prior distribution $\mathbb{P}(\mathcal{M} \in \cdot \mid H_1)$ under an initial null history $H_1 = \emptyset$. In each episode $k \in [K]$, all the learner's knowledge about the world based on the interactions of the preceding generations is characterized by the distribution $\mathbb{P}(\mathcal{M} \in \cdot \mid H_k)$, from which the agent determines a policy $\pi^{(k)}$ to execute. The resulting experience observed $E_k \in \mathcal{E}$ yields an updated history of environment interaction $H_{k+1} = H_k \cup \{E_k\}$ and revised beliefs about the world are reflected by a posterior distribution $\mathbb{P}(\mathcal{M} \in \cdot \mid H_{k+1})$ for the next episode. Since the agent's regret is now a random variable due to uncertainty in $\mathcal{M}$, taking an expectation over the prior yields the γ-discounted Bayesian regret

$$\textsc{BayesRegret}(K, \{\pi^{(k)}\}_{k \in [K]}, \gamma) = \mathbb{E}\left[\textsc{Regret}(K, \{\pi^{(k)}\}_{k \in [K]}, \gamma, \mathcal{M})\right].$$

Members of a culture in one generation could seek to minimize the total Bayesian regret for all future generations. This leads to both theoretical and practical complications, so we posit discounted regret as the ideal objective. Indeed, for practical reasons we restrict to $\gamma = 0$ below—agents who consider only the next generation.

Rate-Limited Posterior Passing with Language

While standard Bayesian RL agents aim to accumulate knowledge and minimize Bayesian regret, CultuRL agents optimize for the same objective while transmitting knowledge between generations. Language is an expressive medium for teaching (Morgan et al., 2015), but it does not enable us to copy beliefs perfectly. Hence we posit a rate limit R restricting how much information can be transferred from one agent to the next. Agents need to decide how to best transmit their beliefs, which are determined by their own received message L_k and observed experience E_k, subject to the rate limit.

A CultuRL agent is able to interpret a message in language as a belief distribution over MDPs $\mathbb{P}(\mathcal{M} \in \cdot \mid L_k)$ where L_k denotes the language received at the start of generation $k \in [K]$. Similarly, the generation k agent has a posterior $\mathbb{P}(\mathcal{M} \in \cdot \mid L_k, E_k)$ after observing experience E_k and must determine a message L_{k+1} that will induce a useful prior over MDPs $\mathbb{P}(\mathcal{M} \in \cdot \mid L_{k+1})$ for the next agent. We can define the distortion attributed to a candidate next message $L' \in \mathcal{L}$, given an initial message $L \in \mathcal{L}$ and observed experience $E \in \mathcal{E}$ as

$$d_L(L, E, L') = \left(\mathbb{E}\left[V_{M'}^{\star} - V_{M'}^{\pi^{(k+1)}} \mid L'\right] - \mathbb{E}\left[V_{M}^{\star} - V_{M}^{\pi^{(k+1)}} \mid L, E\right] \right)_{+},$$

where the first expected regret term occurs with respect to the prior over MDPs defined by the chosen language $\mathbb{P}(M' \in \cdot \mid L')$ and the second term is taken over the agent's posterior $\mathbb{P}(M \in \cdot \mid L, E)$. $(f(x))_+$ represents $\mathrm{ReLU}(f(x))$. The resulting expected distortion encapsulates a notion of *cultural regret*, the difference in expected regret between an agent using the posterior at generation k and the prior defined by the rate-limited message L_{k+1}.

This leads to the distortion-rate function (Shannon, 1959) that agents must minimize to determine a message for the next generation:

$$\min_{\mathbb{P}(L_{k+1} \mid L_k, E_k)} \mathbb{E}\left[d(L_k, E_k, L_{k+1})\right] \text{ s.t. } \mathbb{I}(L_k, E_k; L_{k+1}) \leq R, \quad (1)$$

which quantifies the fundamental limit of lossy compression and where $\mathbb{I}$ denotes the mutual information (Cover

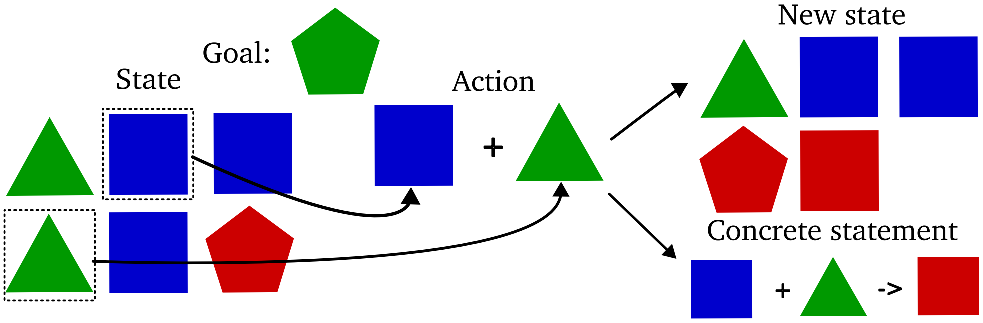

Figure 1: Visualization of the crafting task. States consist of inventories of items. Actions are ordered pairs of items to craft together which produce a new state with the output item. The agent updates its knowledge based on the observed output.

& Thomas, 2012). While the use of lossy compression in Bayesian RL has appeared in prior work (Arumugam & Van Roy, 2022), our application to inter-generational communication of beliefs is novel.

The distortion-rate function (Shannon, 1959) can be solved using the classic Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972), but this is computationally demanding. In practice, we posit that agents constrain their messages using relatively tight upper bounds on the mutual information. For instance, the entropy of the marginal distribution over messages and the maximum surprisal of any message both give bounds: $\mathbb{I}(L_k, E_k; L_{k+1}) \leq H(L_{k+1}) \leq \max -\log p(L_{k+1})$. These bounds are appropriate when the frequency distribution of messages is clear or fixed.

Cultural Transmission in a Crafting Task

We implement CultuRL in the context of a crafting game with three properties that make it suitable for cultural learning: it adheres to the MDP formulation with dynamics that are initially unknown but learnable, knowledge about the task can be expressed conveniently in language, and we can measure an individual's learning over episodes as well as cultural learning. A visualization of the task is shown in Figure 1.

At each timestep, the agent's state in $\mathcal{S}$ is characterized by an inventory of items and a goal, both of which are sampled at the start of the episode from β. All initial states contain six non-goal items, three of which are unique. The goal is an item that the agent needs to create to attain a reward. Each action available to the agent is an ordered pair of input items (i_1, i_2) selected from the inventory. The action produces one output item o. The episode ends when the agent either crafts the goal item (success with a reward of 1) or has only one non-goal item left in its inventory (failure with reward 0). All other rewards are zero. The horizon H is guaranteed to be finite as each action causes the size of the inventory to shrink. While the MDP transition function $\mathcal{T}$ defined by the mapping from input item pairs to output items is deterministic, it is unknown to the agent whereas the rewards associated with all items are known. The agent must learn $\mathcal{T}$ by trying combinations and observing what they yield, reflecting that knowledge through its posterior beliefs $\mathbb{P}(\mathcal{M} \in \cdot \mid L_k, E_k)$.

Knowledge Representation

We represent knowledge about the environment using a simple domain-specific language (DSL) that connects features of the inputs to features of the output. Each item has a color (green, red, blue) and a shape (triangle, square, pentagon). Statements are of the form "[color] [shape] + [color] [shape] → [color] [shape]" where [color] and [shape] can be any concrete color or shape, or "any" or "anything" respectively.

When an agent takes an action, it learns a concrete statement. For example, if the agent crafts a blue square and a red triangle together and produces a green pentagon, it learns the statement "blue square + red triangle → green pentagon". Statements can also be abstract, meaning they do not completely specify all of the features of the input and output items. For example, the statement "any square + any triangle → green pentagon" is abstract.

An agent transmits knowledge by sending a message consisting of a set of DSL statements. We use these DSL statements to represent knowledge transmitted in language.

Model Specification

To implement CultuRL in the crafting task, we need two components: the method that a single agent uses to compute a policy and the method to compute a message to send to the next generation. Code and data are available at `https://github.com/benpry/cultural-rl`

Solution Policy

In each generation, our CultuRL agent synthesizes a policy using a variant of Posterior Sampling for Reinforcement Learning (PSRL) (Strens, 2000; Osband et al., 2013); not only is PSRL a classic algorithm for this Bayesian RL setting, but it also admits a competitive Bayesian regret upper bound (Osband & Van Roy, 2017) that falls within a $\sqrt{H}$ factor of the best known regret lower bound (Jaksch et al., 2010). Like standard PSRL, our CultuRL agent performs Thompson sampling (W. R. Thompson, 1933; Russo et al., 2018) over the underlying MDP model. At the start of generation $k \in [K]$, the cultural agent's knowledge is defined entirely by the language L_k received, which defines beliefs about the underlying world model $\mathbb{P}(\mathcal{M} \in \cdot \mid L_k)$ (in contrast to a PSRL

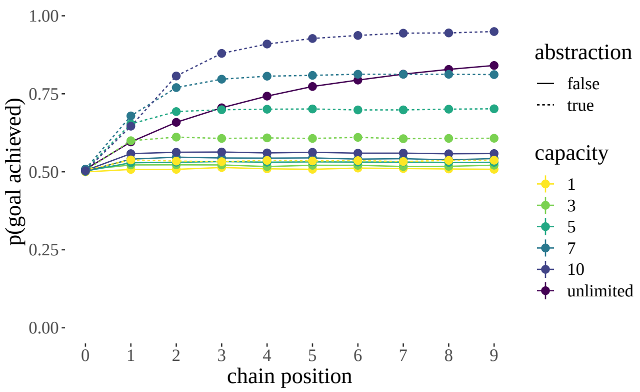

Figure 2: Probability of an agent achieving the goal by generation, estimated over 100 simulated chains per condition and game, averaged across episodes within a generation. The unlimited capacity model never has abstraction.

agent that maintains knowledge based on the full observation history H_k). The agent samples $M_k \sim \mathbb{P}(\mathcal{M} \in \cdot \mid L_k)$, which in our case is a deterministic mapping of input pairs, i_1, i_2, to ouputs, o, consistent with messages received. The agent then executes the optimal policy of the sampled MDP for the episode, $\pi^{(k)} = \pi^{\star}_{M_k}$. The problem of synthesizing $\pi^{\star}_{M_k}$ is a planning problem (Bertsekas, 1995), which we solve using breadth-first search until the agent finds a path to the goal state. If there is no path to the goal in the sampled dynamics, the agent takes a random action.

We make one change to the standard PSRL setup to make it compatible with our task. When the agent discovers an i_1, i_2, o triple that is different from what the sample dynamics predicted, it "patches" its belief, replacing the existing triple with the observed one. If the new concrete statement contradicts any statements in the knowledge base (which can happen when the received message contains an over-generalization), the agent removes the contradicted statement(s). This prevents the agent from planning action sequences that it knows to be impossible in the true environment.

Transmission Policy

After executing a policy $\pi^{(k)}$ and observing experience E_k, the agent computes a message for the next generation L_{k+1} via a distortion-rate optimization (Equation 1). This message defines the next agent's prior $\mathbb{P}(\mathcal{M} \in \cdot \mid L_{k+1})$.

As described above, the objective of an agent passing a message in CultuRL is to minimize regret for the recipient subject to the rate limit. In practice, computing an optimal channel that minimizes Equation 1 for our task is not feasible: even only considering messages with up to 10 statements gives us over 10^{29} possible messages and the convergence rate of the Blahut-Arimoto algorithm is linear in the number of possible messages (see Corollary 1 of Arimoto (1972)). Instead, we use the maximum surprisal of any message that might be sent $\max_{L_{k+1}} - \log p(L_{k+1})$ as an upper bound on the mutual information $I(L_{k+1}; L_k, E_k)$. Assuming the marginal distribution on messages, $p(L)$, is fixed and known to the agent, the rate limit can be satisfied by ensuring that all messages the agent could transmit have a surprisal less than the limit. Assuming a uniform prior over statements, the surprisal of a message depends only on the number of statements it contains. This fact means that the rate limit becomes a limit on the number of statements the agent can pass, which we call the *statement capacity*.

Since we now have a constraint on individual messages, we can approximate an optimal channel by computing the best message given an agent's posterior as the need arises. Unfortunately, computing the regret-minimizing message is still intractable. If we had a statement capacity of 10, we would need to evaluate about 10^{29} messages and choose the best one. Therefore, we use a heuristic compression algorithm to approximate an optimal message.

Our heuristic compression algorithm (losslessly) compresses multiple statements into a single statement when it can and removes the statement that is least likely to be useful when it must remove something. The algorithm starts with a list of all statements in the agent's knowledge base, including received statements and discovered concrete statements. Pseudo-code is shown in Algorithm 1.

Algorithm 1 Message compression algorithm

```
Input: Set of statements $\mathbf{L}$, statement capacity $c$
while $|\mathbf{L}| > c$ do
    $A \leftarrow$ consistent anti-unifications of statement pairs in $\mathbf{L}$
    if $|A| > 0$ then
        $S_0 \leftarrow$ least generalizing anti-unification from $A$
        $S_1, S_2 \leftarrow$ the pair of statements that $S_0$ generalizes
        $\mathbf{L} \leftarrow (\mathbf{L} \setminus \{S_1, S_2\}) \cup \{S_0\}$
    else
        $S \leftarrow$ the most redundant, most specific statement in $\mathbf{L}$
        $\mathbf{L} \leftarrow \mathbf{L} \setminus \{S\}$
    end if
end while
return $\mathbf{L}$
```

In each iteration of the main loop, the algorithm tries to turn two statements into one more general statement. The agent attempts to perform anti-unification on each pair of statements in its knowledge base (Plotkin, 1970). Anti-unification finds the more general statement that entails both statements and as little else as is possible, i.e. the *least general generalization* of the two statements. If such a statement exists, the agent checks if it contradicts anything in its knowledge base. If not, it adds the general statement to a list of candidate generalizations. The agent then chooses the candidate that is closest in specificity to the two statements it generalizes (the least general least general generalization, if you will), randomizing in the case of ties. This reduces the total number of statements by one. The agent repeats this process until either the list of statements no longer exceeds the

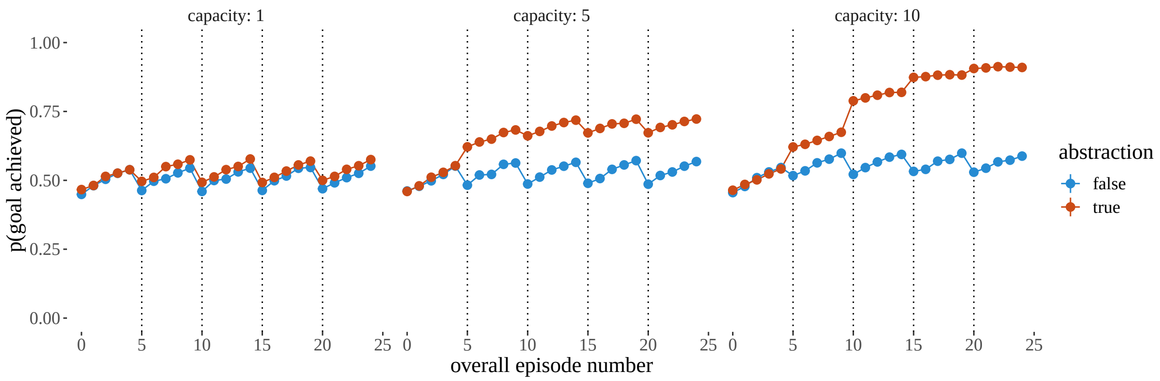

Figure 3: Model performance over all episodes by statement capacity and abstraction setting. Vertical dotted lines indicate transitions between generations. With a high statement capacity and abstraction, agents benefit from generalization.

statement capacity or no more anti-unification is possible.

If there are no valid generalizations, the agent resorts to lossy compression. It removes a statement according to two criteria. First, it prefers to remove more specific statements. The more general a statement, the more likely it is to apply when the next agent is solving the task. If multiple statements are tied for most specific, it removes the most redundant statement, where redundancy is defined as the number of other statements in the message that specify the same output item. The agent continues generalizing and removing statements until its message size meets the statement capacity.

This heuristic algorithm is not guaranteed to minimize regret, but we believe it to be a reasonable approximation for this particular game. Future work is needed to establish regret bounds for this approach or to find better and more general approximations.

Analysis of Model Behavior

In this section, we demonstrate that our model exhibits established properties of cultural transmission. We run our model for 5 episodes per generation for 10 generations. We simulate 100 chains per game on 100 randomly generated games, for a total of 10,000 simulated chains. The games are generated to have abstract structure where only one input feature matters in determining the output. In half of the games, only the colors of the input items matter in determining the output; in the other half, only the shapes matter. This ensures that generalization is possible. We also ensure the games are achievable by simulating agents with perfect knowledge starting from 50 random start states. We only keep games where the simulated omniscient agents could reach the goal all 50 times.

Our first clear finding is that higher statement capacities lead to better scores. Figure 2 shows score by generation for different statement capacities and abstraction settings. When the channel is unlimited, chains eventually reach optimal performance, but it takes longer than 10 generations.

The Benefit of Abstraction

We compare the behavior of CultuRL agents both with and without abstraction. Agents with abstraction follow the transmission policy above.

Without abstraction, agents can only remove statements until their message fits in the channel. These agents only relay direct experience with the environment. If we view the environment's dynamics as a crafting function $f(i_1, i_2) = o$ mapping from inputs to outputs, then passing concrete statements is akin to passing example points from a function in Beppu & Griffiths (2009)'s framing. In contrast, abstract messages have the language-like property of conveying general knowledge about the world—approximate posteriors.

As Figure 2 shows, chains of agents perform substantially better when they can communicate abstract knowledge than when they can only pass concrete statements. There is an interaction with statement capacity, with low-capacity chains plateauing at a low level. Remarkably, chains which can abstract and have a statement capacity of 10 reach optimal performance *faster* than chains with an unlimited channel. The limited channel creates a pressure to compress knowledge that leads to agents discovering true generalizations.

We can get a better sense of how abstraction affects chains by looking at each episode of the first five agents, as is shown in Figure 3. This plot reveals a sawtooth shape for low channel capacities or no abstraction: there is an increase in the probability of achieving the goal over the course of individual learning, followed by a drop when transmission between generations occurs. We see that the "cultural ratchet" (Tennie et al., 2009) is a graded effect: the final accuracy asymptote depends on how much knowledge is lost between generations in comparison to the extent of individual learning.

What Makes for a Useful Message?

Messages with more abstract statements led their recipients to perform better. Logistic regressions predicting goal achievement using the proportion of abstract items in the received message for channel capacities 5 and 10 both found a significant effect (5: $\beta = 0.71, p < .001$, 10: $\beta = 2.29, p < .001$).

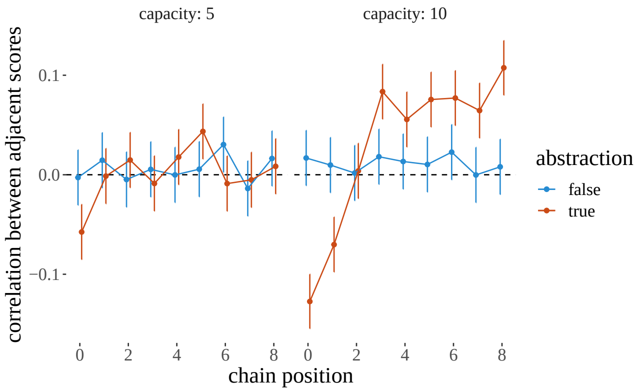

Figure 4: Correlation coefficients between adjacent chain members' scores. Error bars show 95% confidence intervals.

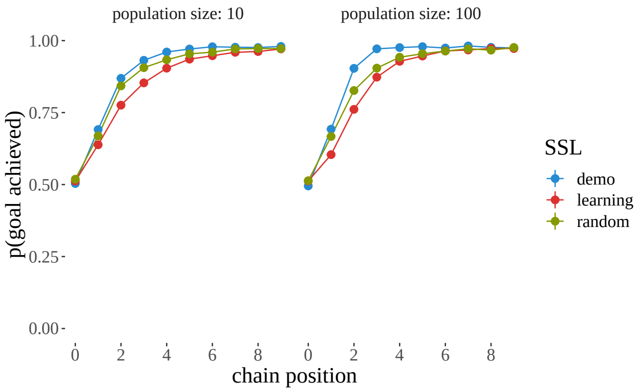

Figure 5: Mean agent scores during the learning phase by population size and selective social learning condition.

This is intuitive: abstract statements represent multiple concrete statements' worth of knowledge. The benefit is enhanced by the fact that generalizations tend to be true in our games, which were designed to have abstract structure.

The correlations between scores of adjacent chain members are shown in Figure 4. There is an interesting pattern in the correlations: scores are negatively correlated for the first two generations, but positively correlated after generation 3 for chains with high statement capacity and abstract language. This reflects the explore-exploit trade-off in the task: If an agent gets lucky and crafts all of its goals early, it learns less than if it tries more actions. Therefore, agents that succeed early in the first generation have less knowledge to pass on than agents that fail. In later generations, an agent's success is more a function of the message it received, hence success indicates that it can pass on a useful message.

Selective Social Learning

Selective social learning (SSL)—that is, learning preferentially from more successful members of the preceding generation—can be a catalyst for human culture (Henrich & Gil-White, 2001; B. Thompson et al., 2022). We explored SSL by moving messages between chains: an agent in generation i's probability of choosing the message from agent a in generation $i-1$ follows a softmax distribution on agent a's average score: $p(a) \propto e^{\frac{1}{\tau}s(a)}$, where $s(a)$ is the proportion of goals achieved. We choose a temperature of $\tau = 0.01$ to create a strong effect, and compared to random choice of message.

Figure 5 shows the mean score by chain position in populations doing selective social learning (red) or the random learning baseline (green). We find that SSL leads to *slower* learning, with a slightly stronger effect for the larger population. Recall the previous finding that success early in the game can reflect a few lucky actions instead of greater knowledge, this apparently leads to worse performance when learning from these high performers.

To test this hypothesis, we introduce a *demonstration phase* after five episodes of learning. In the demonstration phase, an agent completes 50 episodes in which it does not update its knowledge. We call the original five episodes where the agent can learn the *learning phase*. Unlike in the learning phase, performance in the demonstration phase is not confounded with the amount the agent learns. SSL based on demonstration phase scores leads the chains to learn faster than random social learning (Figure 5 blue curve), while SSL based on learning phase scores leads to slower learning.

Discussion

In this paper, we introduced the cultural reinforcement learning (CultuRL) framework, which describes cultural learning as regret minimization subject to an information rate limit between generations. We realized this model in the setting of a crafting game, using Posterior Sampling for Reinforcement Learning (PSRL) to model learning and heuristic compression for communication. Our framework is applicable more broadly, including to stochastic and continuous settings.

CultuRL produces established effects in the literature on cultural transmission. CultuRL also makes a new prediction about selective social learning—in tasks with explore-exploit trade-offs, SSL can slow cultural learning because higher scores can reflect less exploration.

The finding that learning can occur faster with a limited channel than an unlimited one provides an interesting answer to the problem of why people generalize. Models of concept learning (e.g. Goodman et al., 2008) often induce generalization via simplicity priors: they assume that people assign lower prior probability to longer rules. In CultuRL, agents' preference for simplicity is grounded in communicative cost.

A major limitation of this work is the need for a hand-designed DSL. As an alternative we hypothesize that natural language, and particularly the generic construction, is a flexible means to convey abstract knowledge (Chopra et al., 2019). Future work could therefore apply language models to summarize accumulated knowledge in natural language and map between beliefs and language pragmatically, such as in the Rational Speech Acts framework (Frank & Goodman, 2012).

References

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Arumugam, D., & Van Roy, B. (2022). Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning. *Advances in Neural Information Processing Systems*, *35*.

Bellman, R. (1957). A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 679–684.

Bellman, R., & Kalaba, R. (1959). On Adaptive Control Processes. *IRE Transactions on Automatic Control*, *4*(2), 1–9.

Beppu, A., & Griffiths, T. (2009). Iterated learning and the cultural ratchet. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31).

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*. Athena Scientific.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(supplement_2), 10918–10925.

Brinkmann, L., Gezerli, D., Kleist, K., Müller, T. F., Rahwan, I., & Pescetelli, N. (2022). Hybrid social learning in human-algorithm cultural transmission. *Philosophical Transactions of the Royal Society A*, *380*(2227), 20200426.

Chopra, S., Tessler, M. H., & Goodman, N. D. (2019). The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 41, pp. 226–232).

Cover, T. M., & Thomas, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

Derex, M., Beugin, M.-P., Godelle, B., & Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, *503*(7476), 389–391.

Duff, M. O. (2002). *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. University of Massachusetts Amherst.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 688–698).

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian Reinforcement Learning: A Survey. *Foundations and Trends® in Machine Learning*, *8*(5-6), 359–483.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, *22*(3), 165–196.

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-Optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, *11*(4).

Krafft, P. M., Shmueli, E., Griffiths, T. L., Tenenbaum, J. B., & Pentland, A. (2021). Bayesian collective learning emerges from heuristic social learning. *Cognition*, *212*, 104469.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 234–243).

Mesoudi, A., & Thornton, A. (2018). What is cumulative cultural evolution? *Proceedings of the Royal Society B*, *285*(1880), 20180712.

Morgan, T. J., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M., … others (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature communications*, *6*(1), 6029.

Osband, I., Russo, D., & Van Roy, B. (2013). (More) Efficient Reinforcement Learning via Posterior Sampling. *Advances in Neural Information Processing Systems*, *26*, 3003–3011.

Osband, I., & Van Roy, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning* (pp. 2701–2710).

Plotkin, G. D. (1970). A note on inductive generalization. *Machine intelligence*, *5*, 153–163.

Puterman, M. L. (1994). *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, Inc.

Russo, D., & Van Roy, B. (2022). Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, *11*(1), 1–96.

Shafto, P., Wang, J., & Wang, P. (2021). Cooperative communication as belief transport. *Trends in Cognitive Sciences*, *25*(10), 826–828.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec., March 1959*, *4*, 142–163.

Smaldino, P. E., & Richerson, P. J. (2013). Human cumulative cultural evolution as a form of distributed computation. In *Handbook of human computation* (pp. 979–992). Springer.

Strens, M. J. (2000). A Bayesian Framework for Reinforcement Learning. In *Proceedings of the seventeenth international conference on machine learning* (pp. 943–950).

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2405–2415.

Tessler, M. H., Madeano, J., Tsividis, P. A., Harper, B., Goodman, N. D., & Tenenbaum, J. B. (2021). Learning to solve complex tasks by growing knowledge culturally across generations. *arXiv preprint arXiv:2107.13377*.

Thompson, B., van Opheusden, B., Sumers, T., & Griffiths, T. (2022). Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, *376*(6588), 95–98.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*(3), 495–511.

Zaslavsky, N., Hu, J., & Levy, R. P. (2020). A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*.