Anim Cogn
DOI 10.1007/s10071-015-0877-x

ORIGINAL PAPER

The learning of action sequences through social transmission

Andrew Whalen[1] · **Daniel Cownden**[1] · **Kevin Laland**[1]

Received: 15 October 2014 / Revised: 7 April 2015 / Accepted: 11 May 2015

Abstract Previous empirical work on animal social learning has found that many species lack the ability to learn entire action sequences solely through reliance on social information. Conversely, acquiring action sequences through asocial learning can be difficult due to the large number of potential sequences arising from even a small number of base actions. In spite of this, several studies report that some primates use action sequences in the wild. We investigate how social information can be integrated with asocial learning to facilitate the learning of action sequences. We formalize this problem by examining how learners using temporal difference learning, a widely applicable model of reinforcement learning, can combine social cues with their own experiences to acquire action sequences. The learning problem is modeled as a Markov decision process. The learning of nettle processing by mountain gorillas serves as a focal example. Through simulations, we find that the social facilitation of component actions can combine with individual learning to facilitate the acquisition of action sequences. Our analysis illustrates that how even simple forms of social learning, combined with asocial learning, generate substantially faster learning of action sequences compared to asocial processes alone, and that the benefits of social information increase with the length of the action sequence and the number of base actions.

Keywords Social learning · Sequence learning · Temporal difference learning · Markov decision process · Chaining

✉ Daniel Cownden
dcownden@gmail.com

[1] School of Biology, University of St Andrews, Harold Mitchel Building, St Andrews, Fife KY16 9TH, UK

Introduction

Social learning facilitates the transmission of behaviors within animal and human populations. The ability to learn a behavior socially may spare learners the time-consuming process of asocial learning, but also helps individuals acquire otherwise difficult to learn behaviors. The drawbacks of asocial learning are particularly significant in the case of action sequences: behaviors composed of a sequence of base actions, many or all of which may be already contained in the learner's repertoire. The number of possible action sequences grows supra-exponentially with the length of the sequence, so that even for a small repertoire of known base actions and relatively short sequences, the space of all possible action sequences is vast. Despite this, many animals in the wild are able to acquire novel action sequences, which can improve their exploitation of the environment.

Action sequences are found in the resource collection techniques of a number of animals. One of the more prominent accounts of learning action sequences is the nettle preparation technique of mountain gorillas, which allows these animals to exploit a food with a number of physical and chemical defenses (Byrne and Byrne 1993). Being able to process nettles (and other foods) contributes to enabling these gorillas to thrive in a harsher environment than their lowland kin while still maintaining a comparatively resource-rich diet (Byrne 1999). Partly because nearly every gorilla in the study population uses the same food processing technique, and partly because of the complexity of the action sequences deployed, Byrne and Russon (1998) suggest that this behavior is learned socially. Due to the species' endangered classification, it is hard to test this hypothesis in the wild, although some studies have examined the acquisition of

Published online: 26 May 2015

nettle processing in captive gorillas (e.g., Tennie et al. 2008).

Action sequences, particularly involving the use of tools, have been found in other species more amenable to controlled study. Both chimpanzees (Boesch and Boesch 1982) and capuchin monkeys (Ottoni and Izar 2008) use stone hammers to crack nuts. Chimpanzees also use stems of grass and vines to fish for termites (Goodall 1964) and may modify these stems for more efficient food gathering (Sanz et al. 2009). Given the difficulty of learning these action sequences asocially, and inter-group variation of these techniques that is not readily explained by ecological differences, researchers have posited that these behaviors are, at least in part, acquired through social transmission (Byrne and Russon 1998; Whiten et al. 1999; Byrne 2003; Biro et al. 2003).

However, a number of experimental studies suggest that the role that social transmission plays is limited in some of these species (Hoppitt and Laland 2013). In particular, while it appears that many primates are able to acquire individual elements of a sequence through observational learning, there is very little evidence that these primates are able to acquire the entire sequence through observational learning alone. Whiten (1998) found some evidence that chimpanzees could learn how to use a sequence of actions to open an artificial fruit through observational learning, although even here it is not clear how much of the sequence was learned through observation. For other primates, the evidence suggests that action sequences are rarely, if ever, acquired all at once (Hoppitt and Laland 2013). Nonetheless, elements of an action sequence (individual actions) have been found to be transmitted in both captive (Stoinski and Whiten 2003) and wild orangutans (Custance et al. 2001; Call and Tomasello 1995), capuchin monkeys (Custance et al. 1999), and captive gorillas (Tennie et al. 2008), suggesting that the learning of action sequences by primates is typically achieved piecemeal over a period of time, through a combination of social and individual information. This begs the question: "How does limited social learning, coupled with asocial learning, enable the transmission of action sequences?"

To address this question, we extend a well-established machine learning and decision-making framework, temporal difference (TD) learning in Markov decision processes, to analyze how learners might tackle the problem of learning action sequences.

Markov decision processes

For many action sequences, the order in which the actions are performed is important; often early actions in a sequence change the environment so that previously ineffective actions become effective. Understanding how individuals learn action sequences thus requires a model of the environment that changes with a learner's actions. A Markov decision process (MDP) is a mathematical formalism used to represent how individuals interact with the world. In an MDP, learners navigate a series of states which represent possible configurations of the world. The learner then moves between states by performing actions. These actions change the state of the world and can reward the learner with a payoff. This is a general setting that can capture a wide range of decision problems, including the learning of action sequences.

To give a concrete example of an MDP, consider the problem of picking an apple from a tree. In the initial state of the problem, the learner is at the base of the tree with the apple out of reach above her. The learner considers three actions: climb the tree, pull the branch down, and try to grab the apple. The first two actions change the learner's state: climbing up the tree changes the learner's location, and pulling the branch down changes the apple's location. In the absence of the first two, the third action, trying to grab the apple, does not change the learner's state, but rather returns her to the initial state, having now wasted some time and energy trying to grab an out of reach apple. However, after climbing the tree or pulling the branch down, the third action, grabbing the apple, now results in a reward, the apple.

The MDP framework makes explicit the nature of the problem that the learner faces, the possible steps the learner can go through to achieve (or fail to achieve) a goal, and the primary reinforcement or reward that a learner receives at each step of the process. Formally, an MDP is defined by a set of states, S. For each state, $s \in S$, a learner has a set of possible actions, $a \in A_s$. For each action, a, a transition function, T, specifies (probabilistically) which state the learner will find herself in next. Similarly, a reward function, R, specifies the reward (positive or negative) that the learner receives for each action. Agents in an MDP perform sequences of actions to try and maximize their rewards. The traditional problem is to compute the optimal decision rule for an agent, given complete knowledge of the structure and rewards of the MDP.

This problem becomes far more complicated and applicable to learned behavior when learners are initially ignorant of the structure of the MDP. Learners are then tasked with an exploration–exploitation tradeoff; the more time a learner invests in discovering the structure of an MDP, the less time she will have to exploit that MDP for reward. Conversely, if a learner prematurely begins to focus on exploitation of her environment without first exploring it, she runs the risk of not discovering the most efficient means of exploitation and losing out on substantial reward. A number of multipurpose algorithms have been

developed for asocial learning in MDPs (see Sutton and Barto 1998). While many of these algorithms might require complex cognitive machinery, TD learning has emerged as a simple yet powerful model of reinforcement learning that provides an effective means to explore and exploit MDPs.

Temporal difference learning

Temporal difference (TD) learning provides a model of how animals build up associations between actions and rewards. This problem becomes difficult when actions and subsequent rewards are separated in space and time and when the relationship between reward and action is contingent upon intermediate actions. This frequently occurs when an action provides no immediate reward, but serves only to create opportunities to engage in other actions.

In the case of the apple tree example presented above, the difficulty in learning stems from the fact that pulling the branch down is not immediately rewarding, but changes the world so that reaching for the apple is now rewarding. In order to learn this sequence, learners must be able to create an association between pulling the branch down and the future reward this enables. In a TD learning framework, learners do this by creating an intermediate association between the state with the branch pulled down and a reward, allowing this state to act as a secondary reinforcer.

TD learning has received broad empirical support both at a behavioral level and at a neurological level (Seymour et al. 2004; Sutton and Barto 1990; Glimcher 2011; Dayan and Niv 2008). Additionally, despite being more than 20 years old, TD learning is still an integral part of state of the art artificial intelligence research (e.g., Mnih et al. 2015). TD learning builds on associative learning models like Rescorla–Wagner (1972) and Bush–Mosteller (1951) learning, by considering a much finer temporal resolution; unlike previous models, TD learning considers behaviors on an action by action level, instead of a trial level. This allows TD learning to explicitly model the acquisition of secondary reinforcers, which is critical for understanding how learners acquire an action sequence.

TD learning provides a theoretical grounding for the experimental observation that arbitrary action sequences are typically only acquired by animals with the use of a training technique known as chaining, where sequences are built up action by action, either starting with the final component action and working backwards (backward chaining), or starting with the initial action and working forwards (forward chaining). Associative chaining was observed as early as Thorndike's 1898 PhD thesis (Thorndike 1898) and has been most recently investigated within the experimental paradigms of "simultaneous chains" and "concurrent chains". The "simultaneous chains" paradigm was developed to investigate the formation of action sequences when only previous actions and their resulting transformations of the learner's perspective provided cues for subsequent actions. Although this paradigm was developed to test the limits of associative chaining theory (see Terrace 2005), the results of these experiments are consistent with TD learning. The "concurrent chains" paradigm was developed to investigate the relative strength of secondary reinforcers, often taking into account their temporal context (e.g., Berg and Grace 2006). This paradigm has been used to test the predictions of delay reduction theory (Fantino et al. 1993) and other theories that, like TD learning, seek to model the effect of temporal structure on instrumental learning (Grace 1994). However, delay reduction theory and related theories are trial-level models of learning and, unlike TD learning, do not make within-trial, action-level predictions of sequence acquisition.

Like Rescorla–Wagner learning, TD learning interprets association strengths as predictions of future outcomes. In TD learning, associations change on the basis of the mismatch between the predicted future reward and the rewards received. In the context of Markov decision processes, learners are faced with a series of states, s, and actions, a. In TD learning, learners predict the expected value of actions, $W(a)$, and the expected values of states, $V(s)$. To capture the effect of secondary reinforcers, TD learning treats the expected value of an action, $W(a)$, as the sum of the immediate reward produced by the action, r, and the expected value of the state the action brings the learner to $V(s)$. Moreover, the learner is able to refine these value predictions based on the difference between prior expectations and actual experience.

Specifically, if a learner was in state s_i, took action a_i, transitioned to state s_{i+1}, and received a reward r, we assume that learners update their state prediction $V(s_i)$ and their action prediction $W(a_i)$ according to the following rule,

$$V_{\text{new}}(s_i) = V_{\text{old}}(s_i) + \alpha\left(r + \gamma V(s_{i+1}) - V_{\text{old}}(s_i)\right) \quad (1)$$

$$W_{\text{new}}(a_i) = W_{\text{old}}(a_i) + \beta\left(r + \gamma V(s_{i+1}) - W_{\text{old}}(a_i)\right) \quad (2)$$

The change in predicted values is based on the difference between the old predicted value, $V_{\text{old}}(s_i)$, and the new estimate of value based on recent experience, $r + \gamma V(s_{i+1})$. This new estimate of value has two components, r which is the immediate reward, i.e., the primary reinforcement, and $\gamma V(s_{i+1})$ which is the predicted future reward of being in state s_{i+1}, i.e., the learned secondary reinforcement. The parameter γ determines how important the immediacy of reward is to the learner. If γ is close to zero, future rewards are relatively unimportant compared to immediate rewards, whereas if γ is close to one, future rewards are almost as

important as immediate rewards. The parameters α and β determine how much the learner changes her predictions on the basis of new experiences. If α and β are close to zero, then the learner only slightly changes her predictions, whereas if α and β are close to one, the learner completely changes her predictions to match recent experiences.

Despite its apparent simplicity, this update rule allows chains of associations to be built up over repeated trials. These associations are formed by using the value of states as secondary reinforcers, allowing actions that do not provide an immediate reward to become associated with future reward.

We can see this process at work in our apple example. One solution is to pull the branch down and then reach for the apple. If the learner is initially ignorant of the value of each state, being in a state with the branch pulled down is not associated with reward. This means that the first time the learner pulls the branch down, this action will not be associated with reward. After the learner grabs the apple, both the value of grabbing the apple and the value of having the branch pulled down are updated and become associated with reward. This means that the next time the learner pulls the branch down, they are placed in a state which they now know to be rewarding, and so only then associate pulling the branch down with reward.

As can be seen from this example, value associations can only propagate back through an action sequence one step at a time. This means that at a minimum, the number of successful trials needed to propagate the value of ultimate rewards of a sequence to initial choices is equal to the length of that sequence. This is a consequence of only updating value predictions about the immediately preceding action and state after experiencing a reward and state transition. However, it is also possible to update the value of an additional preceding state and action. This is captured in TD learning by updating not only the estimated values of $V(s_i)$ and $W(a_i)$, as in Eqs. 1 and 2, but also the estimated values of $V(s_{i-1})$, and $W(a_{i-1})$:

$$V_{\text{new}}(s_{i-1}) = V_{\text{old}}(s_{i-1}) + \alpha\gamma\left(V_{\text{new}}(s_i) - V_{\text{old}}(s_i)\right) \quad (3)$$

$$W_{\text{new}}(a_{i-1}) = W_{\text{old}}(a_{i-1}) + \alpha\gamma\left(W_{\text{new}}(a_i) - W_{\text{old}}(a_i)\right) \quad (4)$$

It is straightforward to update value estimates for an arbitrary number of previous states and actions in this manner. The number of preceding states and actions updated determines how fast new experiences influence far preceding actions. Those states and actions which are eligible for updating based on the rewards at a given time and in a given state are referred to as the "eligibility trace" in TD learning and is interpreted as a short-term memory of what the learner has done. These eligibility traces have received neurological and behavioral support in operant learning contexts (Pan et al. 2005).

TD learning produces predictions of the value of states and actions based on experiences, but does not prescribe which actions a learner should perform. A simple method to select an action, known as a greedy rule, is to choose the action with the highest predicted value. However, if learners implement a greedy rule, they may not explore the MDP sufficiently and may miss a high-payoff solution. Thus, a good decision rule needs to incorporate some degree of exploration, balanced against the risk of wasting time searching for better solutions when a good solution has already been found. We used a softmax decision rule to balance exploration and exploitation. This rule has proven effective in engineering contexts (Sutton and Barto 1998) and in predicting the choices of humans and animals (Racey et al. 2011). In a softmax decision rule, the probability of choosing an action, a, is proportional to $\exp(W(a)/\tau)$, where τ parameterizes how exploratory the learner is. The softmax decision rule becomes the greedy rule as τ goes to 0 and becomes random action selection as τ goes to infinity. Under this decision rule, only the relative value, not the absolute value, associated with an action determines its probability of being selected.

How a learner estimates the value of unknown states and actions will also influence how exploratory she is. Pessimistic estimates discourage exploration, whereas optimistic estimates encourage exploration (Sutton and Barto 1998). To enhance early exploration, when a learner has never performed and action before, for the purposes of choosing an action, the learner uses the mean estimated value of performed actions as a proxy for the value of the unperformed action.

TD learning, paired with a softmax decision rule, fully specifies how a learner navigates an MDP. These rules depend upon four parameters, α, β, γ, and τ. The ideal parameterization for a learner depends on the specific MDP the learner faces.

Social information in temporal difference learning

To the best of our knowledge, previous work on MDPs has focused on asocial learning, with the exception of multi-agent engineering contexts where agents share information in biologically unfeasible ways (e.g., Tan 1993). However, many animals are known to also use some form of social information to make decisions (Laland and Galef 2009; Zentall and Galef 1988; Heyes and Galef 1996; Hoppitt and Laland 2013). There are many ways of incorporating social information into a reinforcement learning paradigm (see for example Heyes 1994). One of the basic findings of social learning research is that seeing an action demonstrated will make it more likely for an animal to perform that action. Fierce debate has been waged about the

underlying mechanisms responsible for this empirical observation (Heyes 1994; Hoppitt and Laland 2013). Here, we ignore this debate and simply posit that learners are more likely to perform demonstrated actions. Specifically, we integrate social information into TD learning by assuming that when a learner would assign equal probability to engaging in the actions available in a given state (e.g., when a learner arrives at a state for the first time) instead of choosing between the actions randomly, they choose an action with probability proportional to the number of demonstrators who they have seen perform that action in that state.

This is a simplified implementation of social information use, but it provides a reasonable starting point for incorporating the behavioral-level effects of social information into TD learning. This use of social information is consistent with a broad variety of social learning mechanisms, ranging from local enhancement to motor imitation. In practice, our model of social information use likely underestimates the impact of social learning (animals, unlike these learners, use social information even when knowledgeable about the environment), but it nonetheless suffices to illustrate the value of social information when learning action sequences. Other forms of social information transfer, e.g., interactions with the partially processed products of experienced conspecifics, might also be investigated within an MDP and TD learning framework, but we leave this for the discussion.

In this model, we assume that learners already know how to perform all of the base actions and that they are able to discriminate between all states and all actions. We also assume that learners treat all actions independently, precluding generalization between similar actions in different states. As presented in this model, the associations developed for one action have no impact on the associations developed for any other action. This assumption is likely unrealistic, but is sufficient to investigate the impact of social information on sequence learning. How learners parse their environment and their actions into functional units is itself an area of intense study, known as the "parsing problem" (Byrne 2003; Heyes 2009; Brass and Heyes 2005). The experimental work of Reid et al. (2001) demonstrated how the presence of a non-instrumental cue in the environment can shape the functional units of behavior with profound consequences on the effects of reinforcement. Thus, more realistic models of animal learning will need to combine the problem of learning to perceive the environment with learning which actions to choose. Indeed, a current frontier in artificial intelligence research concerns extending the TD learning paradigm to rich, high-dimensional perceptual spaces (e.g., Mnih et al. 2015).

To demonstrate the applicability of this modeling framework to learning problems that animals might face, we next turn to the specific problem of nettle processing in gorillas and construct an MDP to represent it.

Nettle processing

Byrne and Russon (1998) provide an account of how gorillas process nettles to make them more palatable to eat and digest. Nettles are covered in a layer of small spines, which make nettles painful to eat, especially when these spines brush against the gorilla's sensitive lips. In order to eat nettles with a minimum of suffering, the gorillas employ a sequence of actions that allow them to avoid the majority of the spines and encase the rest of the spines in relatively spine-free leaves.

Much of the interest in gorilla nettle processing is due to the hierarchical nature of the action sequence. Although it is possible to model hierarchical processes as MDPs, these models are substantially more complicated than non-hierarchical processes. To understand how social information can be used to aid in the acquisition of action sequences, we consider a simplified, non-hierarchical model of nettle processing.

We simplify this process into a series of four procedural steps, in which the gorilla: (1) gathers leaves from a plant and removes the petioles; (2) removes debris from the bundle of leaves; (3) uses a handful of leaves to wrap the rest of the bundle in; and (4) eats the bundle of leaves, leaving their hands free to collect a new bundle of leaves. This setup assumes that each of the actions (1–4) is already contained in the learner's repertoire, and we do not model how the actions in each step are initially learned. Learning (1) may be particularly challenging, as this action is composed of two lower-level actions, stripping leaves from the stem, and tearing off the petioles of the stripped leaves. Additionally both lower-level actions may need to be repeated before the gorilla has enough partially prepared leaves in her hand.

The sequential dependencies of the actions in this sequence motivate our choice to use it as a focal example. Once the gorilla has wrapped the leaves into a bundle, she cannot remove debris. Likewise, once the gorilla has eaten the bundle, no further modification to the leaves can be made.

The resulting states, and the actions linking the states together, are given in Fig. 1. We include in this diagram an "alternative action" which represents the option of non-nettle foraging. This problem has three potential "branches" of actions. First the gorilla could eat the bundle of leaves raw. Second, the gorilla could fold the leaves over and then eat them. Or third, the gorilla could pick out debris from the leaves, either then eat them or fold the leaves over, and then eat them. After each eating action, the

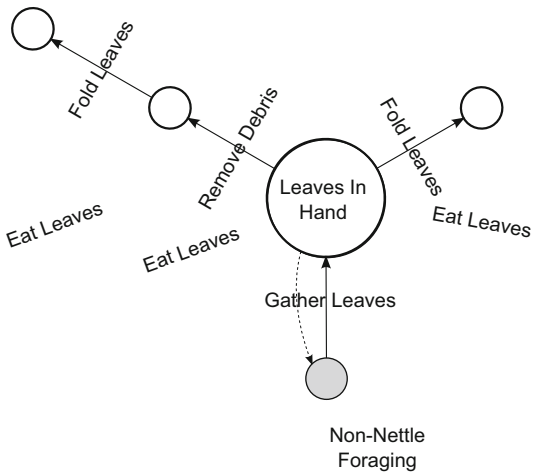

Fig. 1 A graphical representation of gorilla nettle preparation. *Arrows* between nodes represent possible actions and the transitions between states. *Dotted lines* represent eating actions

gorilla enters into the same state (shown in gray), where she is then able to gather and process more leaves. However, once the leaves are folded, we assume that the gorillas do not unfold them and then remove debris, which consequently means that the order in which these action happens influences the types of reinforcement the individuals receive.

We assume that the state where the leaves are folded and the debris have been removed is experienced as distinct from the state where the leaves are folded and the debris have not been removed. The discriminability of these states can arise either from actual immediate perceptual differences, or from the short-term memory of the gorilla. We include the option of abstaining from processing nettles and foraging elsewhere to aid understanding of why gorillas might process nettles.

Learning to process nettles, like other food processing behaviors, is difficult because feedback is only provided at the final step when the gorilla actually eats the bundle of leaves. The gorilla must choose which actions to perform based solely on past knowledge of the outcomes of these actions. Moreover performing some actions may preclude the gorilla from performing others. This means that short action sequences, like failing to remove the debris or folding the leaves, may be easier to learn, but less rewarding. We use a series of simulations to explore how social and asocial learners might solve this problem.

Simulations on the nettle task

We examine how each learner explored and exploited the nettle MDP over a series of 50 trials. Each trial consists of a sequence of actions that returned the learner to the initial state shown in gray in Fig. 1. We examine the payoff of each learner, and the number of times each learner performed the canonical action sequence, i.e., steps 1–4 above. Because the actions each learner performs are stochastic, we measure these values for 100,000 learners.

In the case of social learners, the performance of each learner depends on from whom they learn. To create a pool of experienced demonstrators, we artificially construct a population of 50 learners who explored the MDP simultaneously. At each time step, a single learner is selected from the population at random and performs an action. When a learner completes 50 trials, they are removed from the population and are replaced by a novice learner. Our goal was not to model a realistic demographic process, but to provide an environment with experienced social learners. Because this population is initially naïve, the performance of new social learners improves as the number of competent demonstrators in the population increases. Pilot simulations showed that after a turnover of 5000 learners, performance was no longer increasing. Because of this, we allow for a turnover of 10,000 learners in each simulation before measuring average payoff and number of canonical sequence performed.

The performance of learners in an MDP depends crucially on the structure and rewards of the MDP and the parameters of the learning algorithm.

We examine learners who explore the MDP characterization of the nettle task given in Fig. 1. We assume initially that any action that does not involve eating gives a reward of 0, consuming a correctly processed nettle gives a reward of 10, and the remaining two eating actions give a reward of 1; subsequently, we varied the reward of the alternative action from 2 to 10. We find similar qualitative results to those presented here for a broad number of payoff values.

For simplicity, we assume that learners have no knowledge of the task and so set $V(s)$ and $W(a)$ to 0, and because the MDP is deterministic, we set $\beta = 1$. We truncate the learner's eligibility trace to $n = 2$ initially, although we later vary this parameter. To provide a fair comparison between the performance of social and asocial learners, we use the values of α, γ, and τ that maximize an asocial learner's average payoff on a given MDP. Because the value of the alternative action changes the optimal learning parameters, the learning parameters are optimized separately for each set of payoffs in the nettle processing task.

Using the optimal values allows us to examine the benefit that social information may provide an ideal asocial learner. If we find that social learners perform better than asocial learners who are not optimized to a task, it cannot be determined whether this difference is due to the presence of social information, or to a parameterization that favors social learning. By focusing on the case where

asocial learners are optimal, we are able to examine whether social learners can perform better than the best, and hence any, asocial learner, regardless of the parameterization. Other learning parameters will optimize the performance of social learners; however, to provide a clear comparison, in our simulations social learners use the same learning parameters as asocial learners.

We used a radial basis function learning algorithm to optimize the performance of asocial learners (Buhmann 2000). In this algorithm, we sample a large number of values of α, γ, and τ and estimate the performance of asocial learners who use these sampled learning parameters. We then use these sampled values to interpolate the performance of learners for unsampled parameter values. This interpolation was done by averaging the value of nearby sampled points weighted by the inverse of their cubed Euclidean distance. This interpolation provides an efficient way to approximate the topography of the learner's payoffs as a function of α, γ, and τ, even when a learner's performance is highly stochastic. This process was then repeated, with new samples being drawn from ever smaller radii around the current estimated optima, until the process converged on an optimal value. Convergence was assessed by examining whether the difference between the largest estimated payoff and the median estimated payoff was <0.01 for the 5000 learners in the current radius.

We also ran a second set of simulations to examine how the performance of social and asocial learners depended on the parameters chosen. Given the large number of parameters, evaluating all possible permutations of parameter values would lead to a combinatorial explosion. Because of this, we focused on a single MDP and examined how the performance of learners changed when varying one parameter at a time. Our baseline MDP was a version of the nettle task when the alternative payoff had a reward of four, and learners used the optimal learning parameters, $\alpha = 0.18$, $\gamma = 0.99$, and $\tau = 0.47$.

We explored the following ranges of parameter values. We varied the eligibility trace of each learner from 1 to 5, the number of trials from 10 to 100 in increments of 10, and the size of the population of social learners from 10 to 100 in increments of 10. We examined 20 values of α evenly distributed between 0.1 and 0.9, 20 values of γ between 0.01 and 0.99, and 20 values of τ between 0.1 and 2.

All simulations were hand coded and run in Python 2.7 (python.org).

Nettle task results

Overall, we find that the use of social information increases learners' average payoff and leads to faster acquisition of the canonical action sequence compared to the optimal asocial learner.

We examine the average payoffs of the asocial and social learners and the frequency with which they performed the canonical action sequence (Fig. 2). We find that when the alternative actions provided a low reward, social learners on average receive a higher payoff than asocial learners (Fig. 2a). This difference is a result of the social learners exploiting the canonical action sequence more frequently than asocial learners (Fig. 2b). However, when the alternative action provides a sufficiently high reward, the exploration effort required to learn the canonical sequences outweighs the benefit of discovering it. In this case, we find that both the social and asocial learners do not exploit the canonical actions sequence, but instead learn to use the alternative action.

We find that the difference in performance between social and asocial learners is primarily driven by the early acquisition of the canonical action sequence by social learners. As an illustration of this, we chart the probability that a social or an asocial learner will perform the canonical action sequence over the course of their life (in Fig. 2c). We find that at all times, social learners are more likely to perform the canonical actions sequence than asocial learners. On a social learner's first trial, they have a roughly 60 % chance of performing the canonical action sequence. This chance then drops for subsequent trials as the learner explores other options, before rising again as the social learner learns to exclusively exploit the canonical action sequence.

We examine a range of parameter values and find that social learners consistently outperformed asocial learners. Results of this parameter exploration are shown in Fig. 3. When we vary the number of trials a learner performs, we find that decreasing the number of trials decreases performance of both asocial and social learners, but that social learners continue to outperform asocial learners (Fig. 3a). As the number of learning trials increases, the difference between asocial and social learners first increases and then decreases, but even at 100 learning trials there is a persistent difference between social and asocial learners (Fig. 3a). When we vary the number individuals in a population, we find that the performance of social learners is less in small population, but that their performance quickly plateaus for medium-sized populations (>20 learners), and the performance of social learners always remains above that of asocial learners (Fig. 3b).

When we vary the length of the learners' eligibility trace (Fig. 3c), we find that asocial learners' performance decreases, but even for a short eligibility trace (length 1), the performance of social learners is left largely unchanged. With longer eligibility traces, asocial performance

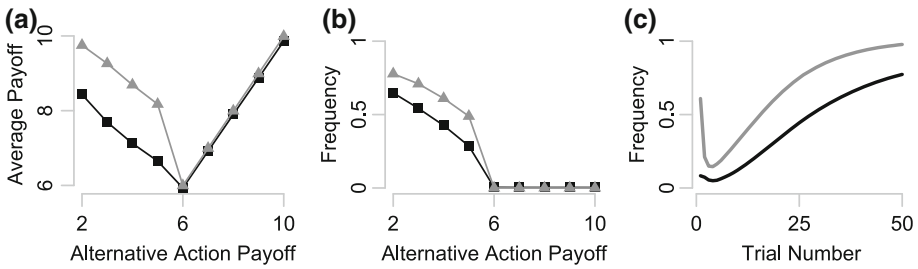

Fig. 2 Learners' performance on the nettle MDP. *Black lines* represent asocial learners; *gray lines* represent social learners. All measurements are averaged over 100,000 learners. **a** Learners' average per-trial payoff, **b** the frequency at which learners performed the canonical sequence averaged over all trials, and **c** the frequency at which learners performed the canonical sequence on each of their 50 trials, on the nettle MDP with alternative action payoff of 4

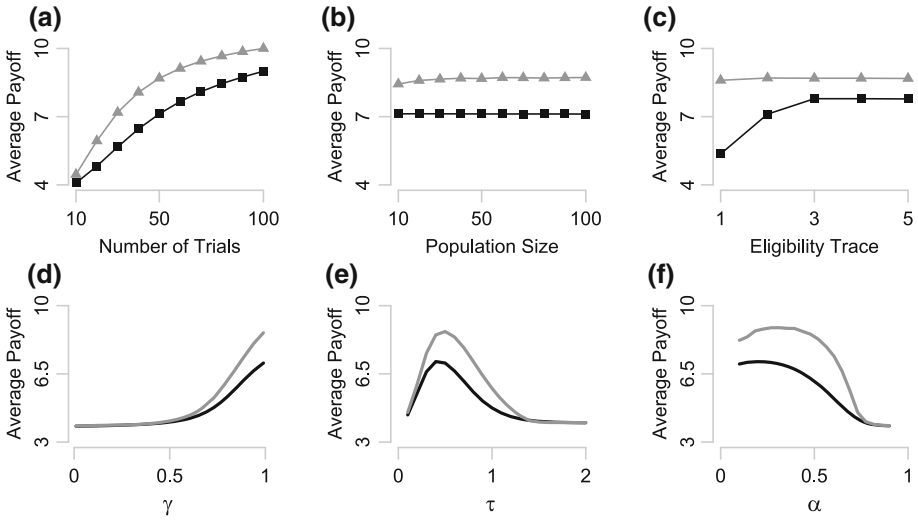

Fig. 3 Learners' average performance on the nettle MDP with an alternative action payoff of 4, when varying **a** number of trials, **b** population size, **c** the eligibility trace, **d** γ, **e** τ, and **f** α. *Black lines* represent asocial learners; *gray lines* represent social learners

improves, but remains below the performance of social learners.

We also systematically vary α, γ, and τ. The results are given in Fig. 3d–f. We find that as learning parameters deviate from the optimum, the performance of asocial learners decreases. However, for all parameter values, social learners perform better than or as well as asocial learners. Social and asocial learners tend to perform similarly, and poorly, when the learning parameters were ill-suited to the MDP. As can be seen in Fig. 3d–f, the optimal values of τ and α for social learners differ from those of asocial learners.

Broad and deep tasks

The findings on the nettle task suggest that incorporating social information into a general purpose asocial learning algorithm can substantially improve learning outcomes. This is particularly the case when the alternative action provides a lower payoff. When the alternative action provides a high payoff, both social and asocial learners perform that alternative action, and the difference in their performance shrinks. This suggests that when the task learners face is relatively simple, social information provides little benefit. In order to investigate when social information provides a benefit to learners in a more general setting, we move away from our cartoon of the nettle stripping problem to investigate two sets of simple stylized MDPs: broad tasks, and deep tasks.

The difficulty of learning an action sequence depends primarily on two things, the length of the action sequence, and the number of base actions available to the learner. We examine how each of these factors might impact the benefit of social information by first considering a set of MDPs where learners are faced with an increasing number of possible actions and then by examining a set of MDPs where learners are faced with increasingly long sequences of actions.

Simulations on broad and deep tasks

Both the breadth and depth tasks are built up from a basic binary decision task. In the binary decision task, learners

are placed at a central node and face two options. Each option leads to a different state, from which the only action available returns them to the home state. The initial transition to each state has a payoff of 0. The transition back to the home state from one of the states has a payoff of 1, the other has a payoff of 2. This creates a simple learning problem for asocial learners. To solve this problem, learners simply need to explore both branches and associate one branch with the higher reward.

In the case of broad tasks, we increase the number of available actions the learner has in the home state, but keep the number of actions in each sequence the same. We set the reward of one of the possible choices to 2 and all other choices to 1. This creates a challenging task as there are many options to be explored. We vary the number of alternative choices available between 1 and 9.

In the case of deep tasks, we increase the number of actions it takes for the learner to receive a reward and return to the home state. We vary the number of intermediate actions between 1 and 9. The difficulty in this problem is not exploration of the task—there are only two choices—but rather, creating a chain of associations between the final reward and the initial choice.

In all other respects, the details of these simulations are identical to those conducted on the nettle processing task. Here, we examine the average lifetime payoff of each learner, which also provides the frequency with which the learner performs the canonical action sequence; a learner obtains a payoff of 2 when performing the canonical action sequence, and a payoff of 1 in all other cases. All measures are averaged over 100,000 learners.

Broad and deep task results

Once again we find that social information consistently provides a benefit to learners. The results of these simulations are given in Fig. 4. Social information provides a greater benefit on broad tasks with a large number of options as compared to few. Social information also provides a greater benefit on deep tasks where the length of the actions sequence was long as compared to short. We find that the performance of social learners does not substantially change with the difficulty of the task; in most tasks, social learners perform at ceiling. In contrast, the performance of asocial learners decreases as the difficulty of the learning problem increases. This drives the difference in performance between social and asocial learners.

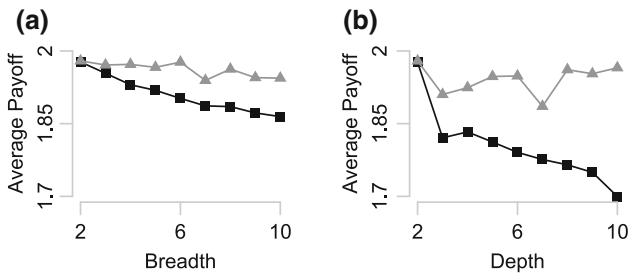

Fig. 4 Learners' performance on the **a** breadth and **b** depth MDPs. *Black lines* represent asocial learners, *gray lines* represent social learners. All measurements are averaged over 100,000 learners

Discussion

In this paper, we used Markov decision processes to model the learning of action sequences. We presented a simple and empirically supported model of individual learning, TD learning, and demonstrated the benefit that social information can provide within this learning framework. We find that the addition of social information allows learners to find and exploit high-payoff behaviors more effectively than through asocial learning alone. Further analyses suggest that social information is particularly helpful both when learners are presented with a choice of many possible actions, and when learning a long sequence of actions to receive a reward: two cases that are challenging for asocial learners.

These results highlight how social information can be integrated into an effective asocial learning algorithm to provide a benefit. In this study, we modeled social learning as performing the same action as a randomly selected demonstrator when the learner is unsure of which action to choose. There are many ways of incorporating social information into TD learning. This particular implementation provides an empirically supported yet minimal way of modeling how social information can influence a learner's behaviors. This use of social information likely underestimates the impact social information on actual learning processes. Nonetheless, we observe a consistent advantage to the use of social information.

Early theoretical work on social learning focused on a dichotomy between costly individual learning and cheap social learning (Boyd and Richerson 1985). One of the longstanding messages of these evolutionary models of social learning is that in order for social learning to be beneficial, learners must be selective in whom they copy (Rogers 1988; Henrich and Boyd 1998). In this framework, we show how a learner can simultaneously use social information and individual learning to solve a complex problem. Even though we assume that learners do not preferentially select their demonstrator, social information still provides a benefit to these learners. These results agree with previous work in suggesting that even copying a random demonstrator can be effective when it is combined with asocial learning (Enquist et al. 2007).

In this analysis, we have used a simple model of social learning: Individuals defer to social information when

unsure what to do, restricting their copying to individuals in the same state. In reality, many animals also exploit information provided by other animals occupying different states, through various forms of observational learning (Hoppitt and Laland 2013), which may further enhance the utility of social learning. Nonetheless, this minimal use of social information is consistent with a number of different mechanisms of social information transfer, such as local and stimulus enhancement, response facilitation, and emulation (Zentall and Galef 1988; Heyes 1994; Hoppitt and Laland 2008). For instance, Hoppitt et al. (2007) found evidence of a response facilitation effect on the rate at which domestic fowl initiated bouts of preening. The rate at which chickens initiated bouts of preening was more strongly related to the number of birds already preening in the same aviary than it was to the number of birds preening in an adjacent, visually obscured aviary, thus ruling out the possibility that any plausible external cues could be wholly responsible for the behavioral synchrony. For social animals in particular, it is highly plausible that the facilitatory effects of other animal's actions will frequently suffice to push them to take the same options when in the same state. Our model shows how response facilitation not only generates behavioral homogeneity across individuals, but also may accelerate sequence learning.

Although we implement social learning as a form of copying, there are other potential social influences on learning. One notable example is the case of Israeli rats learning to strip pinecones. Terkel (1996) found that juvenile rats learn to strip pinecones in an efficient, spiral manner if presented with pinecones that had been partially processed in this way. However, juveniles who received unprocessed pinecones did not learn to use the spiral method to strip pinecones. The social influence on learning in this case is not a direct form of copying, but is instead mediated by interactions with the partially processed products of an experienced learner. These same processes may be at work in other animals as well. While we do not explicitly model this form of social influence, this process can be readily understood within a TD learning framework. One of the difficulties in learning, highlighted in the deep task, was forming associations between early actions and later rewards. The same difficulty is present here; the effort required to remove the first segments of a pinecone outweighs the nutritional reward, however, this is not the case for later segments provided the spiral method is used. Partially processed pinecones then provide the learner with an opportunity to learn and build associations in the later stages of the learning problem. These associations can then serve as secondary reinforcers when learning on un-processed pinecones.

In conclusion, this paper offers insights into how social information may contribute to learning action sequences and suggests that social information may be beneficial even when learners are only able to copy single-component actions of a longer sequence. Seen another way, in the context of learning action sequences, social information only needs to point the learner in the right direction; individual learning can handle the rest.

Acknowledgments The authors thank Richard Byrne for his very helpful comments. DC thanks Magnus Enquist and Johan Lind for countless inspiring conversations and in particular for pointing out the importance of sequence learning, and the potential of Markov decision processes for modeling such learning problems. This research was supported by a grant from The John Templeton Foundation.

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical standard This article does not contain any studies with human participants or animals performed by any of the authors.

References

Berg ME, Grace RC (2006) Initial-link duration and acquisition of preference in concurrent chains. Learn Behav 34(1):50–60

Biro D, Inoue-Nakamura N, Tonooka R, Yamakoshi G, Sousa C, Matsuzawa T (2003) Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments. Anim Cogn 6(4):213–223

Boesch C, Boesch H (1982) Optimisation of nut-cracking with natural hammers by wild chimpanzees. Behaviour 88(3):265–286

Boyd R, Richerson PJ (1985) Culture and the evolutionary process. University of Chicago Press, Chicago

Brass M, Heyes C (2005) Imitation: is cognitive neuroscience solving the correspondence problem? Trends Cogn Sci 9(10):489–495

Buhmann MD (2000) Radial basis functions. Acta Numer 2000(9):1–38

Bush R, Mosteller F (1951) A mathematical model for simple learning. Psychol Rev 58(5):313

Byrne RW (1999) Cognition in great ape ecology. Skill-learning ability opens up foraging opportunities. Symp Zool Soc Lond 72:333–350

Byrne RW (2003) Imitation as behaviour parsing. Philos Trans R Soc B 358(1431):529–536

Byrne RW, Byrne JM (1993) Complex leaf-gathering skills of mountain gorillas (*Gorilla g. beringei*): variability and standardization. Am J Primatol 31(4):241–261

Byrne RW, Russon AE (1998) Learning by imitation: a hierarchical approach. Behav Brain Sci 21(5):667–684

Call J, Tomasello M (1995) Use of social information in the problem solving of orangutans (*Pongo pygmaeus*) and human children (*Homo sapiens*). J Comp Psychol 109(3):308

Custance D, Whiten A, Fredman T (1999) Social learning of an artificial fruit task in capuchin monkeys (*Cebus apella*). J Comp Psychol 113(1):13

Custance D, Whiten A, Sambrook T, Galdikas B (2001) Testing for social learning in the" artificial fruit" processing of wildborn orangutans (*Pongo pygmaeus*), Tanjung Puting, Indonesia. Anim Cogn 4(3–4):305–313

Dayan P, Niv Y (2008) Reinforcement learning: the good, the bad and the ugly. Curr Opin Neurobiol 18(2):185–196

Enquist M, Eriksson K, Ghirlanda S (2007) Critical social learning: a solution to Rogers's paradox of nonadaptive culture. Am Anthropol 109(4):727–734

Eriksson K, Enquist M, Ghirlanda S (2007) Critical points in current theory of conformist social learning. J Evol Psychol 5(1):67–87
Fantino E, Preston RA, Dunn R (1993) Delay reduction: current status. J Exp Anal Behav 60(1):159–169
Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci USA 108(15):647–654
Goodall J (1964) Tool-using and aimed throwing in a community of free-living chimpanzees. Nature 201:1264
Grace RC (1994) A contextual model of concurrent-chains choice. J Exp Anal Behav 61(1):113–129
Henrich J, Boyd R (1998) The evolution of conformist transmission and the emergence of between-group differences. Evol Hum Behav 19(4):215–241
Heyes CM (1994) Social learning in animals: categories and mechanisms. Biol Rev 69(2):207–231
Heyes CM (2009) Evolution, development and intentional control of imitation. Philos Trans R Soc B 364(1528):2293–2298
Heyes CM, Galef BG (1996) Social learning in animals: the roots of culture. Elsevier, Amsterdam
Hoppitt W, Laland KN (2008) Social processes influencing learning in animals: a review of the evidence. Adv Stud Behav 38:105–165
Hoppitt W, Laland K (2013) Social learning: an introduction to mechanisms, methods, and models. Princeton University Press, Princeton
Hoppitt W, Blackburn L, Laland KN (2007) Response facilitation in the domestic fowl. Anim Behav 73(2):229–238
Laland KN, Galef BG (2009) The question of animal culture. Harvard University Press, Cambridge
Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
Ottoni EB, Izar P (2008) Capuchin monkey tool use: overview and implications. Evol Anthropl Issues, News Rev 17(4):171–178
Pan WX, Schmidt R, Wickens JR, Hyland BI (2005) Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci 25(26):6235–6242
Racey D, Young ME, Garlick D, Pham JNM, Blaisdell AP (2011) Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. Learn Beahv 39(3):245–258
Reid AK, Chadwick CZ, Dunham M, Miller A (2001) The development of functional response units: the role of demarcating stimuli. J Expt Anal Behav 76(3):303–320
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF (eds) Classical conditioning: current research and theory. Appleton-Century-Crofts, New York, pp 64–99
Rogers AR (1988) Does Biology Constrain Culture? Am Anthropol 90(4):819–831
Sanz C, Call J, Morgan D (2009) Design complexity in termite-fishing tools of chimpanzees (*Pan troglodytes*). Biol Lett 5(3):293–296
Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS (2004) Temporal difference models describe higher-order learning in humans. Nature 429(6992):664–667
Stoinski TS, Whiten A (2003) Social learning by orangutans (*Pongo abelii* and *Pongo pygmaeus*) in a simulated food-processing task. J Comp Psychol 117(3):272
Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. In: Gabriel M, Moore J (eds) Learning and computational neuroscience: foundations of adaptive networks. MIT Press, Cambridge, pp 497-537
Sutton RS, Barto AG (1998) Reinforcement learning: AN introduction, vol 1. Cambridge Univ Press, Cambridge
Tan M (1993) Multi-agent reinforcement learning: independent vs. cooperative agents. In: Proceedings of the tenth international conference on machine learning, pp 330–337
Tennie C, Hedwig D, Call J, Tomasello M (2008) An experimental study of nettle feeding in captive gorillas. Am J Primatol 70(6):584–593
Terkel J (1996) Cultural transmission of feeding behavior. In: Heyes CM, Galef BG (eds) Social learning in animals: the roots of culture. Elsevier, Amsterdam, p 17
Terrace HS (2005) The simultaneous chain: a new approach to serial learning. Trends Cogn Sci 9(4):202–210
Thorndike EL (1898) Animal intelligence: an experimental study of the associative processes in animals. Psychol Monogr Gen Appl 2(4):1–109
Whiten A (1998) Imitation of the sequential structure of actions by chimpanzees (*Pan troglodytes*). J Comp Psychol 112(3):270
Whiten A, Goodall J, McGrew WC, Nishida T, Reynolds V, Sugiyama Y, Tutin CE, Wrangham RW, Boesch C (1999) Cultures in chimpanzees. Nature 399(6737):682–685
Zentall TR, Galef BG Jr (1988) Social learning: psychological and biological perspectives. Lawrence Erlbaum Associates, Hillsdale