Behavioral Signatures of Backward Planning in Animals

Authors: Arsham Afsardeir[1], Mehdi Keramati[2,3]*

[1] Control and Intelligence Processing Center of Excellence, School of ECE, College of Engineering, University of Tehran, P.O. Box 14395-515, Tehran, Iran.

[2] Gatsby Computational Neuroscience Unit, Sainsbury Wellcome Centre, University College London, 25 Howland Street, London W1T 4JG, UK.

[3] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square London WC1B 5EH, UK.

* Correspondence to: mehdi@gatsby.ucl.ac.uk

Abstract:

Goal-directed planning in behavioral and neural sciences is theorized to involve a prospective mental simulation that, starting from the animal's current state in the environment, expands a decision tree in a forward fashion. Backward planning in the artificial intelligence literature, however, suggests that agents expand a mental tree in a backward fashion starting from a certain goal state they have in mind. Here we show that several behavioral patterns observed in animals and humans, namely outcome-specific Pavlovian-to-instrumental transfer and differential-outcome effect, can be parsimoniously explained by backward planning. Our basic assumption is that the presentation of a cue that has been associated with a certain outcome triggers backward planning from that outcome state. On the basis of evidence pointing to forward and backward planning models, we discuss the possibility of brain using a bidirectional planning mechanism where forward and backward trees are expanded in parallel to achieve higher efficiency.

Introduction

Most real-life problems in humans and other animals involve making a sequence of choices, each of which have costs and benefits. Behavioral and neurobiological research in decision making since the cognitive revolution in the 1950s (Tolman, 1948) and particularly over the last three decades (Adams and Dickinson, 1981; Doya, 1999; Dickinson and Balleine, 2002; Daw et al., 2005) have highlighted the important role of planning as a brain decision making mechanism. Several studies have shown that humans (Balleine and O'Doherty, 2010; Glascher et al., 2010; Lee et al., 2014; Doll et al., 2015) and animals (Tolman, 1948; Dickinson and Balleine, 2002; Balleine and O'Doherty, 2010) build a cognitive map of the dynamics of their environment and then use this knowledge to foresee the potential consequences of different courses of action. Formalizing this idea, "model-based forward planning" is a class of reinforcement learning algorithms (Sutton and Barto, 1998) that has been widely used to explain this body of experimental evidence (Daw et al., 2005; Glascher et al., 2010; Keramati et al., 2011; Dezfouli and Balleine, 2012; Huys et al., 2012; Doll et al., 2015; Kurth-Nelson et al., 2016). According to forward planning models, given the individual's knowledge of environmental dynamics, a prospective mental simulation expands a decision tree that starts from the decision-maker's current state. This process reveals to the decision-maker the expected consequences of each sequence of actions, helping him to choose the option that maximizes reward.

The complicated nature of many real-life problems, however, renders forward planning insufficient, whereas humans and animals can skillfully solve them. Consider, as an example, the problem of finding the best route to a concert on the other side of the city. As the depth of a naïve forward planning that starts from your current position (analogous to a breadth-first search algorithm) increases, the number of edges (i.e., streets) of the corresponding mental decision-tree grows exponentially. Cognitive limitations like time and working memory (Otto et al., 2013) thus restrict the depth of forward planning (Huys et al., 2012) and make it impossible to take into account the rewarding value of states that are far away (i.e., the concert).

Several solutions to this so-called curse of dimensionality are suggested in the field of artificial intelligence, and behavioral and neural signatures of some of those strategies are also reported in humans and animals. These strategies include pruning (Huys et al., 2012), hierarchical reinforcement learning (Botvinick, 2008; Botvinick et al., 2009), and plan-until-habit strategies (Keramati et al., 2016). Another classical solution in artificial intelligence to tackle the curse of dimensionality in sequential decision problems is bidirectional planning (Dijkstra, 1959; Pohl, 1971; Russell and Norvig, 2009). According to this approach, when the goal-state (e.g. the concert venue, in our example) is known to the artificial agent, two simultaneous mental trees are expanded: one forward, initiating from the current state, and one backward, initiating from the goal state. If these two trees meet in the middle, a path (i.e., a sequence of choices) for reaching the goal-state from the current state is discovered. This forward-backward scheme increases the effective depth of planning, with imposing minimal cognitive costs to the system (discussed further in the next sections). Although this approach is widely used in artificial intelligence, it has remained, to the best of our knowledge, unexplored in cognitive science (Balleine and O'Doherty, 2010).

Here, we show that in addition to forward planning, there are several behavioral and neurobiological evidence that point to the use of a backward planning mechanism in humans and other animals. Namely, our simulation results show that backward planning can explain outcome-specific Pavlovian-to-Instrumental Transfer (PIT) in both factual and counterfactual action-outcome mappings (Laurent and Balleine, 2015), as well as differential-outcome effects (Trapold, 1970; Urcuioli, 2005). We hypothesize that forward and backward planning mechanisms provide the brain with two major building blocks of a bidirectional planning system, enabling animals to mitigate the curse of dimensionality.

Although the behavioral and neurobiological bases of merging forward and backward trees remain unspecified, and the existing evidence can be equally explained by assuming separate forward or backward planning mechanisms that take control over behavior under different circumstances, here we take an integrative approach and frame our solution in a unified "bidirectional" planning mechanism. The advantage of this approach is merely being unified as well as being normatively motivated. Otherwise, as shown, behavioral evidence explained here by simulating a bidirectional planning system can be equally explained by mere backward planning.

Materials and Methods

Theory Sketch:

An animal, hypothetically, views the environment as a set of states – in each state, a set of actions is available; taking an action in a certain state results in receiving some immediate reward or punishment and will also take the animal to a new state. Given this view of the environment, according to the computational theory of reinforcement learning (RL; (Sutton and Barto, 1998), in each state the animal tries to find the action that maximizes the sum of expected rewards it will receive, or other similar objectives.

Forward planning, as one solution to this problem, assumes that animals learn the causal dynamics of their environment in terms of transition and reward functions, representing respectively the ensuing new state and immediate reward upon performing a certain action in a certain state. When at a choice point, the animal exploits this knowledge and expands a mental decision tree that starts from its current state in the world and foresees the expected consequences of different courses of actions. The simplest way to take the limitation of cognitive resources into the forward-planning account is to assume that, given cognitive resources, the depth of the tree is limited (Sutton and Barto, 1998). Whether using this naïve depth limitation or more sophisticated pruning methods, resource-limited forward planning produces a mental tree that starts from the current state and ends at several terminal states, called leaves. Each sequence of imagined actions takes the animal from the current state to one of the leaves, and results in a sequence of immediate rewards. Thus, for a limited depth *D*, this algorithm only takes into account the first *D* rewards that will be received following a certain choice, and will ignore further consequences.

One efficient solution to expand the thinking horizon of the algorithm, while imposing minimal cognitive costs, is to expand simultaneously another mental tree in a backward fashion (Pohl, 1971; Russell and Norvig, 2009). Imagine the animal is living in an environment with very sparse rewarding states. That is, at every given point of time, only one or very few states of the environment contain outcomes that are highly interesting to the animal. Knowing the position of a goal state in such environments, the animal can start from the goal state and expand a backward tree. This would require the animal to retrieve causal associations from memory in a reverse order: what state-action pair will lead to the goal state? Having retrieved those penultimate states, this process can be repeated several times in order to gradually expand the backward tree. After *D* repetitions of such a process, the backward tree reaches a depth of *D* and the leaves of such a tree represent the states that will eventually lead to the goal state after taking a sequence of *D* actions. Thus, the backward planning process attributes a high reward value to all the leaves, as well as the states inside the backward tree, since the goal-state is reachable from those states.

Now, if the forward and backward trees expand simultaneously, the forward planning process does not necessarily need to limit its horizon to only *D* steps. In fact, if any of the leaves of the expanded forward search belong to the backward tree, then the backward planning process already has some estimate of the value of those states. Thus, rather than ignoring the consequences further than *D* steps (i.e., further than its leaves), the forward process can use the values computed, for its leaves, by the backward process (**Fig. 1**).

Bidirectional planning, therefore, can increase the depth of planning from *D* to *2D* (in the best case), along the course of actions where the forward and backward trees meet. This can be seen in **Fig. 2** where the efficiency of forward, backward, and bidirectional planning strategies are compared in an environment with highly sparse rewards. This increased depth and efficiency, given an appropriate implementation of the algorithm, can have no extra time cost. For example, if the causal structure of the world is represented in an associative network and tree expansion is realized by a spreading activity algorithm (Baronchelli et al., 2013), then a neural network implementation of this system can spread neural activity in parallel (Rogers and McClelland, 2004) in forward and backward directions from start and goal locations. This will have the same time cost as spreading activity in only one direction. The biological processes allowing merging the trees at leaf level, however, remains unspecified in the spreading-activity networks literature (Baronchelli et al., 2013).

Even in the absence of parallel expansion, bidirectional planning is more efficient in terms of working memory usage. Since the number of visited nodes increases exponentially as the depth of tree increases, the total number of retrieved states in a tree with depth *D* is on the order of b^D, where *b* is the branching factor (i.e., number of choices in each state). Thus, the number of retrieved states for reaching a depth of *2D* is on the order of b^{2D} for a forward-only process, but b^D for a bidirectional planning process (Russell and Norvig, 2009). Therefore, bidirectional planning is an effective solution for tackling the curse of dimensionality in complex environments with spare rewards, given that the model of the environment and the reward-containing states are known to the agent.

Association Retrieval in Reverse Order

As mentioned before, one necessary mechanism for backward planning is retrieval of causal associations in reverse order. There are several lines of behavioral and neurobiological research that point to the existence of such a capability in animals.

Neurobiologically, reverse replay of hippocampal place cells (Foster and Wilson, 2006; Diba and Buzsáki, 2007; Karlsson and Frank, 2009; Carr et al., 2011) is a clear demonstration that the brain "can" retrieve experienced sequences in reverse. According to these findings, particularly during periods of relative immobility of the animal, hippocampal place cells show sequential reactivations that encode the experienced behavioral trajectories in reverse order. This reverse pattern has also been observed recently in MEG signal recorded from humans solving a sequential decision making task (Kurth-Nelson et al., 2016).

Behaviorally, animals that were passively moved through a tree-like maze from the terminal states (goal positions) back to the starting state could later use this knowledge efficiently (as compared to naïve control animals) to actively navigate from the starting state to their desired terminal state (Pritchatt and Holding, 1966). This would either directly show that animals use backward planning and use the learned associations in the same order experienced (from the end to the beginning), or if animals use forward planning, it at least demonstrates that they can exploit the experienced associations in reverse order (i.e., experienced backward, but exploited in a forward fashion). It is noteworthy that in this study (Pritchatt and Holding, 1966), forward training (i.e., moving animals through the maze passively, in a forward direction) was more effective that backward training.

Results

The theoretical argument and the experimental evidence discussed above show that bidirectional planning is an efficient and neurobiologically plausible algorithm. Here we show that bidirectional planning can explain two behavioral patterns: outcome-specific Pavlovian-to-Instrumental Transfer (PIT) and differential-outcome effect. Our simulation results show that in both cases, the backward planning mechanism is an essential component for explaining behavior.

Outcome-Specific Pavlovian-to-Instrumental Transfer

Outcome-specific PIT, although requiring specific experimental conditions, is a robust pattern that captures how cues that signal the presence or absence of certain goals influence goal-directed behavior (Kruse et al., 1983; Colwill and Rescorla, 1988; Corbit and Balleine, 2005; Balleine and Ostlund, 2007; Holmes et al., 2010). In a typical experimental scenario, different conditioned stimuli (CS) are associated with different outcomes (O) during an initial Pavlovian training phase (e.g. CS1→O1 and CS2→O2). In an upcoming instrumental phase, animals also learn that different actions (A) lead to each of those outcomes (e.g. A1→O1 and A2→O2). In a final test phase, performed in extinction, the presentation of the CS associated with an outcome is shown to enhance the instrumental response directed to the same outcome only (e.g., CS1 enhancing A1, but not A2). In other words, a CS signaling the presence of an outcome guides choices specifically toward seeking that outcome (**Fig. 3**).

A recent study (Laurent and Balleine, 2015) further showed that a CS that signals absence of an outcome, instead, motivates behaviors toward the alternative outcome. In this experiment, during the Pavlovian training phase, while CS1 predicted O1, CS1 and CS3 presented in conjunction (CS1-CS3) predicted absence of O1. Likewise, CS2 predicted O2, whereas CS2 and CS4 presented together (CS2-CS4) predicted absence of O2. Experimental results replicated the classical specific PIT patterns: CS1, but not CS3, enhanced seeking O1, and CS2, but not CS4, enhanced seeking O2. Moreover, the results revealed a new pattern: CS1-CS3 enhanced seeking the alternative outcome, O2, but not the absence-signaled outcome, O1. Vice versa, CS2-CS4 enhanced seeking O1, but not the absence-signaled outcome, O2 (**Fig. 3**).

Our simulation results show that only backward (**Fig. 4C**) and bidirectional (**Fig. 4D**), but not forward (**Fig. 4B**), planning systems can explain these patterns, pointing to the essential role of backward planning. These results are based on the critical assumption that Pavlovian cues signal the availability or absence of outcomes and thus, trigger expansion of a backward decision-tree from the appropriate goal state. We simulate 120 agents in an artificial environment (**Fig. 4A**) that captures the experimental design in rats (Laurent and Balleine, 2015). In order to obtain O1, for example, the agent first needs to approach lever 1 (AL1) and press it (PL1) so that the corresponding food pellet becomes available at the magazine, and then it should approach the magazine (AM) and consume the outcome (CO1). In this environment, the animal only consumes either of the two outcomes that is highly rewarding. Other responses (i.e., AL, PL, AM), in contrast, have a small *punishment*, capturing the energetic cost of performing actions (see the subsection “Formal model and simulation details” below for details).

In the absence of any Pavlovian cue (i.e., during the instrumental phase), as traditionally assumed in many RL models, a forward planning process evaluates the rewarding consequences of choices by expanding a decision-tree from the agent’s current state. This control condition (i.e., absence of Pavlovian cues) gives rise to a certain rate of pressing each of the two levers, hereafter called the baseline rate.

The critical assumption we make for explaining PIT is that presentation of a positive Pavlovian cue (e.g., CS1) works as a reminder that enables the recall of the goal and thus, triggers the expansion of a backward tree, in parallel with the forward-tree that is always expanded from the current state. We further assume, however, that the backward search starts most of the time (two-thirds, in our simulations) from the relevant goal-state (i.e., from O1, when CS1 is presented), but sometimes (i.e. one-third of the time) from the irrelevant goal state (i.e., O2, when CS1 is presented). This assumed stochasticity could be due to memory retrieval noise, or intrinsic noise in neural activity.

For the bidirectional planning simulations (**Fig. 4D**), we assume a depth of four for both forward and backward components. For the forward-only (**Fig. 4B**) and backward-only (**Fig. 4C**) simulations, however, we assume a depth of zero for the backward and the forward components, respectively, and a depth of four for the other component.

In keeping with the assumption that CS1 triggers backward search, we assume that CS1-CS3 also triggers backward search in the same fashion (i.e., same stochasticity, same depth, etc.). However, since CS3 also signals absence of O1, the backward search process assigns a rewarding value of zero to that goal state. Thus, backward planning back-propagates zero value from that state. Similarly, in the case of presenting only CS3, a rewarding value of zero is attributed to the relevant goal state, but the backward planning process is not triggered due to absence of CS1. Therefore, the zero value of the goal-state affects decisions only when that state falls within the forward-search depth limit.

Simulation results show that in the presence of backward planning (**Fig. 4C, D**) CS1 presentation enhances responses to O1 (i.e., compared to baseline). This is because backward-planning triggered from the goal-state (G1) increases the chance of taking the rewarding value of O1 into account when the agent's current state is still distant from that goal. CS1 presentation could also, with a small probability, trigger backward search from the irrelevant goal-state (i.e., G2). This results in a marginal increase of responding for O2, compared to the baseline. This effect, though, was not statistically significant in rats (Laurent and Balleine, 2015).

Presentation of CS3 or CS4 alone, in our backward-only (**Fig. 4C**) and bidirectional (**Fig. 4D**) simulations, also has only a marginal effect on lever-press rates. Since CS3 and CS4 are never associated with the food outcomes during the Pavlovian phase, we assume their presentation does not evoke backward planning. In the absence of backward planning, the rewarding values of O1 and O2 are rarely taken into account by only forward planning when the current state is distant from them. Resetting them to zero upon presentation of CS3 or CS4, therefore, has only a small effect on choice.

In our simulations of backward-only (**Fig. 4C**) and bidirectional (**Fig. 4D**) planning, presentation of CS1-CS4 significantly increases responding toward O1. This is because CS1 triggers backward-planning from G1 most of the time and thus motivates behavior toward that goal. In the rare case that backward-planning initiates from G2 instead, the presence of CS4 signals the absence of O2 and thus, a rewarding value of zero propagates back from G2. In such cases, the rewarding value of both levers remain unaffected by their relevant outcomes and therefore, the agent remains indifferent toward the levers.

Last but not least, presentation of CS1-CS3 marginally inhibits seeking O1, but significantly guides choice toward O2 (**Fig. 4C, D**). Replicating this experimental fact by our model is the result of both the backward planning mechanism as well as the stochastic nature of choosing the root of the backward tree. That is, when CS1 triggers backward planning from G1 (most of the time), a value of zero propagates back, resulting in indifference toward either of the levers. When CS1 occasionally triggers backward planning from G2, however, a positive value propagates back that directs behavior toward O2. Therefore, explaining rats' behavior in response to compound Pavlovian CS (both congruent and incongruent), but not the basic specific PIT pattern, requires the introduced assumption on the *stochasticity* in the starting-point selection of backward tree-expansion. Note that this stochasticity assumption is only necessary for explaining the rats' behavior in response to congruent-compound CSs, but not the basic specific PIT nor the incongruent-compound-CSs pattern. In sum, our simulation results show that the backward component of the bidirectional planning model is essential for explaining the PIT patterns.

As shown in simulation results (**Fig. 4C, D**), our model predicts a stronger factual response to the incongruent, compared to the counterfactual response to the congruent, compound CSs. This is a limitation of our theory, since the experimental results (Laurent and Balleine, 2015) do not show a statistically significant difference between the two, or even suggest the opposite trend.

Differential Outcomes Effect

Several studies have provided compelling evidence that choice accuracy increases when each of the different stimuli predicts a distinctive, as opposed to common, reward (Trapold, 1970; Urcuioli, 2005). In a typical experiment, animals learn to choose one of two options in the presence of different stimuli. For one group, called the differential-outcome group, each of the two reinforced responses yields one of the two different outcomes (Fig. 5A). For example, performing action A1 in the presence of CS1 yields outcome O1, whereas performing A2 in the presence of CS2 yields O2. For a second group, called the nondifferential-outcome group, both reinforced responses yield, in a random way, one of the two outcomes (Fig. 5B). That is, the CSs still signal the correct choice, but are not predictive of the outcome to be received upon choosing that correct action. Experimental results show a significantly higher choice accuracy in the differential, compared to the nondifferential group (Trapold, 1970; Urcuioli, 2005).

Using an argument similar to that for specific PIT, forward-backward planning can account for the differential outcomes effect. We assume that in the differential-outcomes group, given the fact that CS1 always leads to O1 upon taking the correct choice (i.e., A1), a Pavlovian association is formed between CS1 and O1. Similarly, a Pavlovian association is established between CS2 and O2. Therefore, presentation of a CS triggers backward planning from the relevant goal state and propagates the rewarding value of that state back to the correct choice.

In the non-differential case, however, CS1 is associated with both O1 and O2 with equal probability. Therefore, we assume that presentation of CS1 triggers backward planning from one of the two relevant goal states, with equal probability. However, since each outcome is delivered only half of the time upon performing the correct action, only half of the rewarding value of that outcome propagates back to the correct choice at the starting state. Therefore, the probability of choosing the correct choice is significantly less, compared to the differential case where the full value of the outcome is back-propagated. Note than in the nondifferential case (Fig. 5B), whether or not the presentation of CS1 (or CS2) also trigger backward planning in the lower (or upper) tree does not affect the values of actions in the upper (or lower) tree, which is relevant to the animal's current state.

Fig. 5C shows the simulation results, using the exact parameter values used for simulating specific PIT. Results show that the backward-planning component is essential for explaining the differential outcomes effect.

Formal model and simulation details:

A deterministic Markov Decision Process (MDP) is defined by a 5-tuple $(S, A, T_{\cdot}(.,.), R_{\cdot}(.,.), \gamma)$, where S is a finite set of states, A is a finite set of actions, and $T(s, a)$ and $R(s, a)$ are the transition and reward functions returning, respectively, the state and reward ensued after taking action a in state s. Finally $\gamma \in [0,1]$ is the discount factor. The goal, in our case, is to choose a stochastic policy π that maximizes the expected discounted sum over a potentially infinite time-horizon (Sutton and Barto, 1998):

$$\langle \sum_{t=0}^{\infty} \gamma^{t} R(s_t, a_t) \rangle_{\pi}$$

by choosing actions $a_t = \pi(s_t)$. To approximate this value for each possible action, the Bellman equation can be used in a recursive way for *D* consecutive depths, equivalent to expanding a forward decision tree with depth *D*, to compute the value of each action a_t:

$$Q_{forward}^{D}(s,a) = R(s,a) + \gamma V_{forward}^{D-1}(s')$$

Where $V_{forward}^{D-1}(s') = \max_{a'} Q_{forward}^{D-1}(s',a')$ and $s' = T(s,a)$. When *D* is reduced to zero (i.e., after reaching depth *D*) the value of actions at the terminal states can be set to zero ($Q_{forward}^{D=0}(s',a') = 0$), equivalent to ignoring the consequences of state-action pair (s,a) that are further than *D* steps. Alternatively, if the terminal state s' belongs to the backward tree, the values at the terminal states can use the values computed by the backward tree:

$$if\ s' \in \mathbb{S}_{backward}: \qquad Q_{forward}^{D=0}(s',a') = Q_{backward}(s',a')$$

Where $\mathbb{S}_{backward}$ is the set of states in the backward tree. To construct this tree, the set is initialized to $\mathbb{S}_{backward} = \{s_G\}$, where s_G is a goal state where backward planning begins. Also, the backward value of all actions at this state are initialized to zero: $Q_{backward}(s_G,a) = 0$.

The backward tree is then expanded in *D* iterative steps. At each step, any state that can reach, with only one action, any state in $\mathbb{S}_{backward}$ will be added to the set:

$$\forall\ s,a,s': \qquad if\ s' \in \mathbb{S}_{backward}\ \ and\ \ s' = T(s,a), \qquad then\ add\ s\ to\ \mathbb{S}_{backward}$$

For each such states, s, the $Q_{backward}(s,a)$ will be computed as:

$$Q_{backward}(s,a) = R(s,a) + \gamma V_{backward}(s')$$

Where $s' = T(s,a)$ and $V_{backward}(s') = \max_{a'} Q_{backward}(s',a')$

Having computed the forward values for all available actions, a *soft-max* rule can be used for action-selection:

$$p(a_t = a|s_t) \propto e^{\beta Q_{forward}^{D}(s_t,a)}$$

where β is the rate of exploration. Generalizing this bidirectional model-based evaluation to stochastic MDPs is straightforward.

To replicate the experimental data from (Laurent and Balleine, 2015), 120 simulated agents were simulated, each for 10^4 trials. The values of the free parameters of the model were $\gamma = 0.9$ and $\beta = 0.45$. See Fig. S2 for the sensitivity of simulation results to free parameters of the model. The payoff for different actions was as follows: $R(.,nul) = -1$, $R(.,AL) = -5$, $R(.,AM) = -5$, $R(.,PL) = -10$, $R(.,CO) = 1000$.

Discussion

Although backward planning seems to be very pertinent to the kinds of environments humans and other animals live in, it has so far evaded the cognitive science literature. This is likely partly due to the fact that direct behavioral manifestation of this strategy is difficult to capture in experimental tasks. Observing the subjects' thought process, for instance by decoding neural activity, is one alternative approach to investigate backward planning more closely. In this respect, one could predict sharp wave ripple-associated hippocampal place cell activity in reverse order from the goal position at the time the animal is at a choice point and shows vicarious trial-and-error (VTE) behavior (Tolman, 1939; Wikenheiser and Redish, 2015).

In this paper, we showed that outcome-specific PIT and differential-outcome effect can be parsimoniously understood as indirect behavioral signatures of backward planning. The main mechanism in the algorithm that explains rats' behavioral patterns (Urcuioli, 2005; Laurent and Balleine, 2015) is the CSs evoking a representation of the goal states and thus, triggering backward planning.

While backward planning is essential in our explanation of PIT and differential-outcome effect, other behavioral evidences are classically interpreted as signatures of forward planning. For example, vicarious trial-and-error (VTE) behavior in rats (Tolman, 1939), defined by head movement from one choice to another at a choice point, implies thinking forward about the future (Redish, 2016). Other behavioral studies show that during prospective evaluation of a sequence of choices, humans prune a sequence (i.e., curtailed any further evaluation) once they expect a large loss to be encountered on the sequence (Huys et al., 2012). Moreover, time pressure is shown to impose a limit on the depth of planning (Keramati et al., 2016). Explaining these decision tree-pruning patterns is inherently based on the assumption that planning occurs in a forward fashion.

If the brain uses both forward and backward mechanisms for planning, one possibility is that those mechanisms are separately used in different contexts, mediated by a biologically unspecified arbitration mechanism. An alternative possibility is that a unified bidirectional planning system employs the two mechanisms is parallel, and is equipped with a (yet biologically unspecified) tree-merging mechanism. In this paper, we used this second alternative for the simulations. However, we must emphasize that there is no evidence, to the best of our knowledge, to preferentially support any of the two hypotheses. The advantage of the bidirectional approach is merely being unified as well as being normatively motivated.

Our backward planning explanation of specific PIT can be viewed as a formal modeling of the associative-cybernetic theory proposed before (Balleine and Ostlund, 2007; Balleine and O'Doherty, 2010). According to this theory, the instrumental phase of the PIT experiments engenders two different associations: the response predicts the outcome (response-outcome; R-O), and the outcome itself acts as a stimulus that predicts the next response (outcome-response; O-R). During the test session, the Pavlovian CS activates the outcome representation and thus, through the O-R associations, selectively enhances the relevant response. Our model proposes that the computational mechanism by which the outcome representation leads to the relevant response is the expansion of a decision tree from that outcome in a backward fashion. There is indeed evidence suggesting that R-O associations can be harnessed in a backward manner to guide action selection (de Wit et al., 2009).

Several studies have shown that the specific PIT pattern remains intact after outcome devaluation (Rescorla, 1994; Holland, 2004; Corbit et al., 2007). That is, devaluating the outcome by for example associating it with a bitter substance does not reduce the power of the Pavlovian CS in increasing motivation for seeking that outcome. Our model cannot explain this observation; if devaluation simply decreases the value of a goal state, and CS presentation triggers a backward propagation of that value, then devaluation would reduce the effect of the CS on instrumental choice.

While we propose that the Pavlovian CS acts as a "reminding" signal that triggers backward planning, a previous computational theory (Cartoni et al., 2013) suggests that the CS works as an "availability" signal. According to this Bayesian account, the CS cues that there is a higher probability of getting the reward associated with the instrumental action and thus, motivates pursuit of that action. Although this conceptualization explains basic specific PIT, it does not account for counterfactual PIT in response to compound congruent CSs. In this model, there is no reason for the presence of CS3 to make the presence of the hidden cause for O2 more likely. In other words, this model in its present form does not feature competition between hidden causes and therefore, is unable to account for counterfactual PIT.

While our model offers a normatively-motivated account for outcome-specific PIT, it leaves general PIT unexplained. In general PIT, the Pavlovian CS enhances all appetitive instrumental responses, even if they are associated with a different outcome (Dickinson and Balleine, 2002; Holland, 2004; Corbit and Balleine, 2005). It is, however, noteworthy that general and specific forms of PIT are shown to be governed by distinct neural processes (Corbit et al., 2001; Corbit and Balleine, 2005, 2011) and thus, explaining both of them with a single computational mechanism is not likely to be a sensible approach.

Consistent with our proposal that specific PIT stems from Pavlovian cues affecting the planning strategy, the basolateral amygdala (BLA), as one of the critical areas involved in specific PIT, projects to several areas involved in instrumental decision making: the orbitofrontal cortex, the core and shell of the nucleus accumbens (Alheid, 2003), and mediodorsal thalamus (Reardon and Mitrofanis, 2000). Lesions of these areas have indeed been shown to affect specific PIT (Ostlund and Balleine, 2005, 2007, 2008). This connectivity could be the neural substrate for the Pavlovian associations triggering the model-based planning process.

The efficiency of bidirectional planning, as discussed before, relies on starting the expansion of backward tree from “appropriate” goal-states. In this paper, we simply assumed that backward planning starts from the goal states signaled by the Pavlovian CSs, but an important open question is how to choose such appropriate states. Saliency and accessibility from the current state seem to be two key requirements. Also, whereas we used a simple breadth-first algorithm that expands the trees evenly along all directions, one could think of pruning both forward and backward trees to avoid imposing unnecessary cognitive costs. Last but not least, the idea of bidirectional planning can be generalized to multi-directional planning where several decision trees expand from different starting points; i.e., not only from the current and goal states, but also from intermediate states like important landmarks, or bottleneck states (Botvinick et al., 2009) in the environment.

In sum, we propose that several lines of evidence, namely hippocampal reverse replay patterns (Foster and Wilson, 2006), rats exploiting associations in reverse order (Pritchatt and Holding, 1966), specific PIT (Laurent and Balleine, 2015), and differential-outcome effect (Trapold, 1970; Urcuioli, 2005) suggest that animals use a backward model-based planning algorithm to estimate the value of choices. We hope this proposal motivates future research to investigate this possibility in more tailored experimental conditions.

If higher efficiency is the reason that animals use bidirectional, compared to unidirectional planning, then the critical characteristic of a behavioral task that would require bidirectional planning is a high branching factor at both the starting state and the goal state. In such cases, the total number of nodes of the expanded decision tree grows rapidly as a function of depth and thus, one cognitively feasible solution to such problems would be to use bidirectional planning. This principle can be used for designing tasks that show more direct signatures of bidirectional planning. Similarly, having a high branching factor in the forward direction (e.g., in a tree, with the root as the starting state and a leaf as the goal state) would motivate using backward-only planning, since it will the most efficient algorithm in terms of working memory load.

Conflict of Interest Statement

The authors declare that no competing interests exist.

Author Contributions

M.K. conceived the study. A.A. performed the simulations. A.A. and M.K. discussed the results. M.K. wrote the manuscript.

Data Accessibility Statement

Data are available from the corresponding author on request.

Acknowledgments

M.K. was supported by the Gatsby Charitable Foundation. We thank Peter Dayan, Zeb Kurth-Nelson and Amir Dezfouli for suggesting significant improvements to the manuscript.

References:

Adams CD, Dickinson A (1981) Instrumental responding following reinforcer devaluation. Q J Exp Psychol 33:109–121.

Alheid GF (2003) Extended amygdala and basal forebrain. Ann N Y Acad Sci 985:185–205.

Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35:48–69.

Balleine BW, Ostlund SB (2007) Still at the choice-point: action selection and initiation in instrumental conditioning. Ann N Y Acad Sci 1104:147–171.

Baronchelli A, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N, Christiansen MH (2013) Networks in Cognitive Science. Trends Cogn Sci 17:348–360.

Botvinick MM (2008) Hierarchical models of behavior and prefrontal function. Trends Cogn Sci 12:201–208.

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113:262–280.

Carr MF, Jadhav SP, Frank LM (2011) Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nat Neurosci 14:147–153.

Cartoni E, Puglisi-Allegra S, Baldassarre G (2013) The three principles of action: a Pavlovian-instrumental transfer hypothesis. Front Behav Neurosci 7:153.

Colwill RM, Rescorla RA (1988) The role of response-reinforcer associations increases throughout extended instrumental training. Anim Learn Behav 16:105–111.

Corbit LH, Balleine BW (2005) Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. J Neurosci 25:962–970.

Corbit LH, Balleine BW (2011) The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. J Neurosci 31:11786–11794.

Corbit LH, Janak PH, Balleine BW (2007) General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the ventral tegmental area. Eur J Neurosci 26:3141–3149 Available at: http://www.ncbi.nlm.nih.gov.gate2.inist.fr/pubmed/18005062.

Corbit LH, Muir JL, Balleine BW (2001) The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. J

Neurosci 21:3251–3260.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.

de Wit S, Ostlund SB, Balleine BW, Dickinson A (2009) Resolution of conflict between goal-directed actions: Outcome encoding and neural control processes. J Exp Psychol Anim Behav Process 35:382–393.

Dezfouli A, Balleine BW (2012) Habits, action sequences and reinforcement learning. Eur J Neurosci 35:1036–1051.

Diba K, Buzsáki G (2007) Forward and reverse hippocampal place-cell sequences during ripples. Nat Neurosci 10:1241–1242.

Dickinson A, Balleine BW (2002) The role of learning in motivation. In: Volume 3 of Steven's Handbook of Experimental Psychology: Learning, Motivation, and Emotion, 3rd ed. (Gallistel CR, ed), pp 497–533. New York: Wiley.

Dijkstra EW (1959) A Note on Two Problems in Connexion with Graphs. Numer Math:269–271.

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015) Model-based choices involve prospective neural activity. Nat Neurosci 18:767–772.

Doya K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural networks 12:961–974.

Foster DJ, Wilson MA (2006) Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440:680–683.

Glascher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585–595.

Holland PC (2004) Relations between Pavlovian-instrumental transfer and reinforcer devaluation. J Exp Psychol Anim Behav Process 30:104–117.

Holmes NM, Marchand AR, Coutureau E (2010) Pavlovian to instrumental transfer: A neurobehavioural perspective. Neurosci Biobehav Rev 34:1277–1295.

Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP (2012) Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS

Comput Biol 8:e1002410.

Karlsson MP, Frank LM (2009) Awake replay of remote experiences in the hippocampus. Nat Neurosci 12:913–918.

Keramati M, Dezfouli A, Piray P (2011) Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. PLoS Comput Biol 7:e1002055.

Keramati M, Smittenaar P, Dolan RJ, Dayan P (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. Proc Natl Acad Sci U S A 113:12868–12873 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5111694 [Accessed July 24, 2017].

Kruse JM, Overmier JB, Konz WA, Rokke E (1983) Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. Learn Motiv 14:165–181.

Kurth-Nelson Z, Economides M, Dolan RJ, Dayan P (2016) Fast Sequences of Non-spatial State Representations in Humans. Neuron 91:194–204.

Laurent V, Balleine BW (2015) Factual and Counterfactual Action-Outcome Mappings Control Choice between Goal-Directed Actions in Rats. Curr Biol 25:1074–1079.

Lee SW, Shimojo S, O'Doherty JP (2014) Neural computations underlying arbitration between model-based and model-free learning. Neuron 81:687–699.

Ostlund SB, Balleine BW (2005) Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. J Neurosci 25:7763–7770.

Ostlund SB, Balleine BW (2007) Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J Neurosci 27:4819–4825.

Ostlund SB, Balleine BW (2008) Differential involvement of the basolateral amygdala and mediodorsal thalamus in instrumental action selection. J Neurosci 28:4398–4405.

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941–20946.

Pohl I (1971) Bi-directional Search. In: Machine Intelligence, 6th ed. (Meltzer B, Michie D, eds), pp 127–140. Edinburgh University Press.

Pritchatt D, Holding DH (1966) Guiding Deutsch's model in reverse. Br J Psychol 57:17–23.

Reardon F, Mitrofanis J (2000) Organisation of the amygdalo-thalamic pathways in rats. Anat Embryol (Berl) 201:75–84.

Redish AD (2016) Vicarious trial and error. Nat Rev Neurosci 17:147–159 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26891625 [Accessed October 24, 2017].

Rescorla RA (1994) Transfer of instrumental control mediated by a devalued outcome. Anim Learn Behav 22:27–33.

Rogers TT, McClelland JL (2004) Semantic cognition : a parallel distributed processing approach. MIT Press.

Russell SJ (Stuart J, Norvig P (2009) Artificial intelligence : a modern approach, 3rd ed. Prentice Hall.

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge: MIT Press.

Tolman E (1939) Prediction of vicarious trial and error by means of the schematic sowbug. Psychol Rev 46:318–336.

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55:189–208.

Trapold MA (1970) Are expectancies based upon different positive reinforcing events discriminably different? Learn Motiv 1:129–140.

Urcuioli PJ (2005) Behavioral and associative effects of differential outcomes in discrimination learning. Learn Behav 33:1–21.

Wikenheiser AM, Redish AD (2015) Hippocampal theta sequences reflect current goals. Nat Neurosci 18:289–294.

Figures:

Fig. 1. Schematics of the forward-backward planning algorithm (see the subsection "Formal model and simulation details" for the general formal algorithm). Expanding a backward tree with depth *D*, started from an arbitrary goal state, s_G, the algorithm finds all states like s_B and s_C that can potentially lead to that goal state by taking a sequence of *D* actions. Similarly, expanding a forward tree with depth *D*, started from the current state s_S, the algorithm finds all states like s_A and s_C that are reachable by performing a sequence of *D* actions from the current state. If the forward and backward trees overlap, as in s_C, the forward-backward planning algorithm has actually found a sequence of (at most) *2D* actions for reaching the goal-state, s_G, from the current state s_S.

Fig. 2. An example of bidirectional vs. unidirectional planning. A 20x20 map was used with the black states as blocks. The large and small red states contain a large (50 units) and a small (25 units) reward, respectively. The agent is assumed to have perfect knowledge of the map (i.e., reward and transition function). Each episode starts from a random starting state (e.g., the blue state in this figure). In each state, the agent can choose among four actions: up, down, left, and right, except when block states or map boundaries restrict the agent. Each action has a small constant cost of 0.05 units, and takes the agent to the targeted next state deterministically. Reaching any of the two reward states results in receiving reward and resetting the episode (i.e., returning to start). The agent is assumed to have a fixed tree-depth of 10. Examples of the expanded decision tree when the agent is in the start state are shown in red for backward (A) and in blue for forward planning (C). When bidirectional planning is used (B), forward and backward trees meet and the value of the big reward back-propagates to the actions available at the starting state, and therefore guides behavior toward the big reward. In the case of forward planning (C), however, only the small reward is in the mental "visual-field" of the agent and therefore, behavior is directed toward that goal. (D) As a result, the total reward accumulated over time (i.e., number of moves) is highest for a bidirectional planner, and lowest for a backward planner (in the context of this environment). Each curve is averaged over 100 simulations.

Fig. 3. Experimental results on factual and counterfactual PIT in rats (reprinted from (Laurent and Balleine, 2015)). (A) Experimental design. In an initial Pavlovian training phase, animals learned that CS1 and CS2 predict O1 and O2, respectively, whereas combined presentation of CS1 and CS3, as well as CS2 and CS4 predict absence of any outcome. In a separate instrumental phase, animals learned that actions A1 and A2 lead to O1 and O2, respectively. In a test phase performed in extinction, the animals' rate of choosing A1 vs. A2 was measured under four different cue-presentation conditions. Results showed that animals' response rate, as compared to baseline (i.e., in the absence of Pavlovian cues), was enhanced when CS1 or CS2 were presented, but only for the response that led to the outcome that was predicted by the cue (B). Furthermore, presentation of CS1-CS4 (or CS2-CS3) in conjunction increased seeking the outcome that was associated with CS1 (or CS2) (C, left), whereas presentation of CS1-CS3 (or CS2-CS4) together increased seeking the outcome that was associated with CS2 (or CS1) (C, right). Reprinted with permission from Elsevier.

Fig. 4. Backward planning explains factual and counterfactual PIT. (A) The experimental design was captured by a Markov-Decision Process (MDP) where reaching a goal state like G_1 requires a sequence of actions including approaching lever number one (AL1), pressing the lever (PL1), then approaching the magazine (AM) and consuming the outcome (CO1). Red and blue boxes show the domain of forward and backward trees when the agent is in the starting state, s_0, and the depth of planning is two. Simulating the backward-only (C) and forward-backward (D), but not the forward-only (B), models in this environment captures the essential behavioral patterns observed in rats (Laurent and Balleine, 2015).

Fig. 5. Backward planning explains differential outcomes effect. (A) In the differential group, the presented CS is predictive of the outcome (i.e., O1 vs. O2. Receiving no reward is indicated by X) to be received upon choosing the correct action. (B) In the nondifferential group, however, the CS predicts, given the correction action, receiving either of the two outcomes with equal probability. (C) Simulating the forward-only, backward-only, and forward-backward planning algorithms show that backward planning is a critical component for explaining experimental evidence.

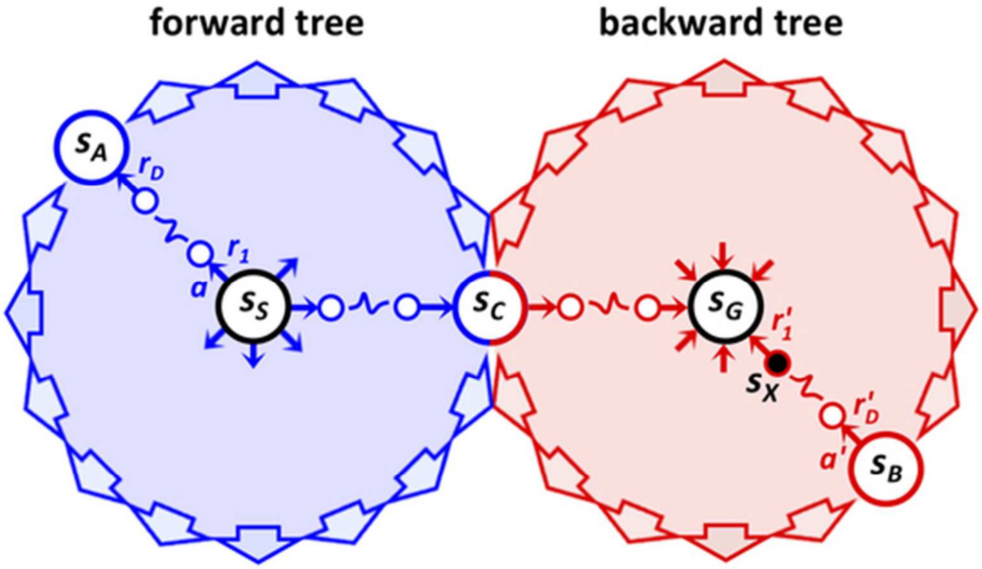


Fig. 1. Schematics of the forward-backward planning algorithm (see the subsection "Formal model and simulation details" for the general formal algorithm). Expanding a backward tree with depth D, started from an arbitrary goal state, s_G, the algorithm finds all states like s_B and s_C that can potentially lead to that goal state by taking a sequence of D actions. Similarly, expanding a forward tree with depth D, started from the current state s_S, the algorithm finds all states like s_A and s_C that are reachable by performing a sequence of D actions from the current state. If the forward and backward trees overlap, as in s_C, the forward-backward planning algorithm has actually found a sequence of (at most) 2D actions for reaching the goal-state, s_G, from the current state s_S.

46x27mm (300 x 300 DPI)

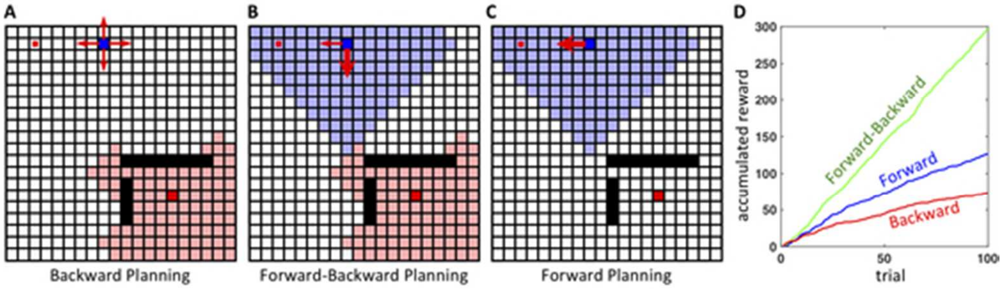
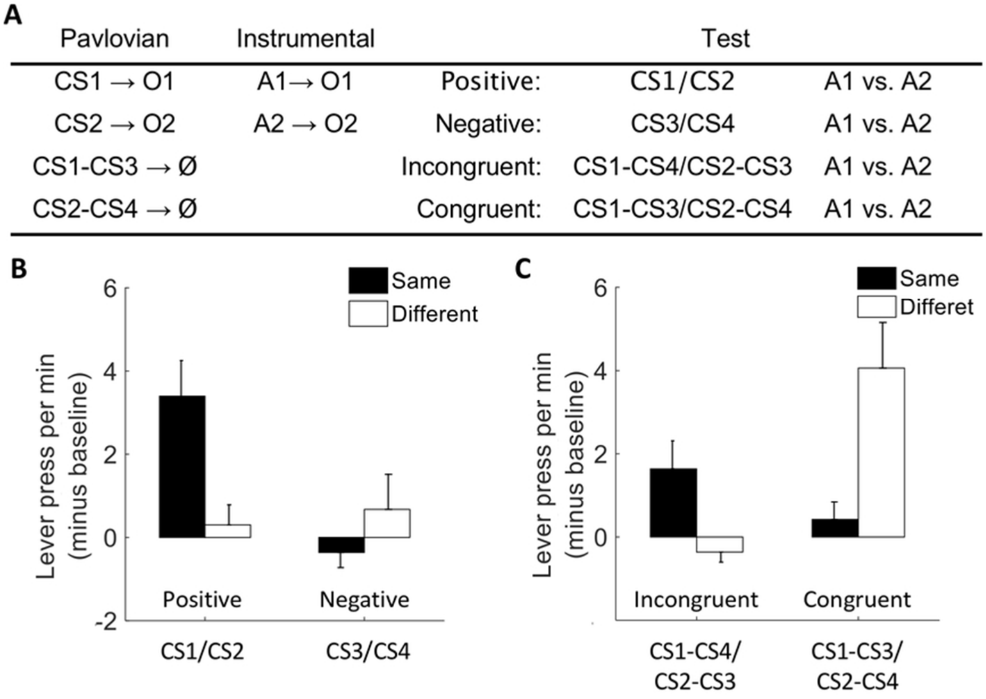


Fig. 2. An example of bidirectional vs. unidirectional planning. A 20x20 map was used with the black states as blocks. The large and small red states contain a large (50 units) and a small (25 units) reward, respectively. The agent is assumed to have perfect knowledge of the map (i.e., reward and transition function). Each episode starts from a random starting state (e.g., the blue state in this figure). In each state, the agent can choose among four actions: up, down, left, and right, except when block states or map boundaries restrict the agent. Each action has a small constant cost of 0.05 units, and takes the agent to the targeted next state deterministically. Reaching any of the two reward states results in receiving reward and resetting the episode (i.e., returning to start). The agent is assumed to have a fixed tree-depth of 10. Examples of the expanded decision tree when the agent is in the start state are shown in red for backward (A) and in blue for forward planning (C). When bidirectional planning is used (B), forward and backward trees meet and the value of the big reward back-propagates to the actions available at the starting state, and therefore guides behavior toward the big reward. In the case of forward planning (C), however, only the small reward is in the mental "visual-field" of the agent and therefore, behavior is directed toward that goal. (D) As a result, the total reward accumulated over time (i.e., number of moves) is highest for a bidirectional planner, and lowest for a backward planner (in the context of this environment). Each curve is averaged over 100 simulations.

47x13mm (300 x 300 DPI)



A

Pavlovian	Instrumental	Test		
CS1 → O1	A1→ O1	Positive:	CS1/CS2	A1 vs. A2
CS2 → O2	A2 → O2	Negative:	CS3/CS4	A1 vs. A2
CS1-CS3 → Ø		Incongruent:	CS1-CS4/CS2-CS3	A1 vs. A2
CS2-CS4 → Ø		Congruent:	CS1-CS3/CS2-CS4	A1 vs. A2

Fig. 3. Experimental results on factual and counterfactual PIT in rats (reprinted from (Laurent and Balleine, 2015)). (A) Experimental design. In an initial Pavlovian training phase, animals learned that CS1 and CS2 predict O1 and O2, respectively, whereas combined presentation of CS1 and CS3, as well as CS2 and CS4 predict absence of any outcome. In a separate instrumental phase, animals learned that actions A1 and A2 lead to O1 and O2, respectively. In a test phase performed in extinction, the animals' rate of choosing A1 vs. A2 was measured under four different cue-presentation conditions. Results showed that animals' response rate, as compared to baseline (i.e., in the absence of Pavlovian cues), was enhanced when CS1 or CS2 were presented, but only for the response that led to the outcome that was predicted by the cue (B). Furthermore, presentation of CS1-CS4 (or CS2-CS3) in conjunction increased seeking the outcome that was associated with CS1 (or CS2) (C, left), whereas presentation of CS1-CS3 (or CS2-CS4) together increased seeking the outcome that was associated with CS2 (or CS1) (C, right). Reprinted with permission from Elsevier.

74x52mm (300 x 300 DPI)

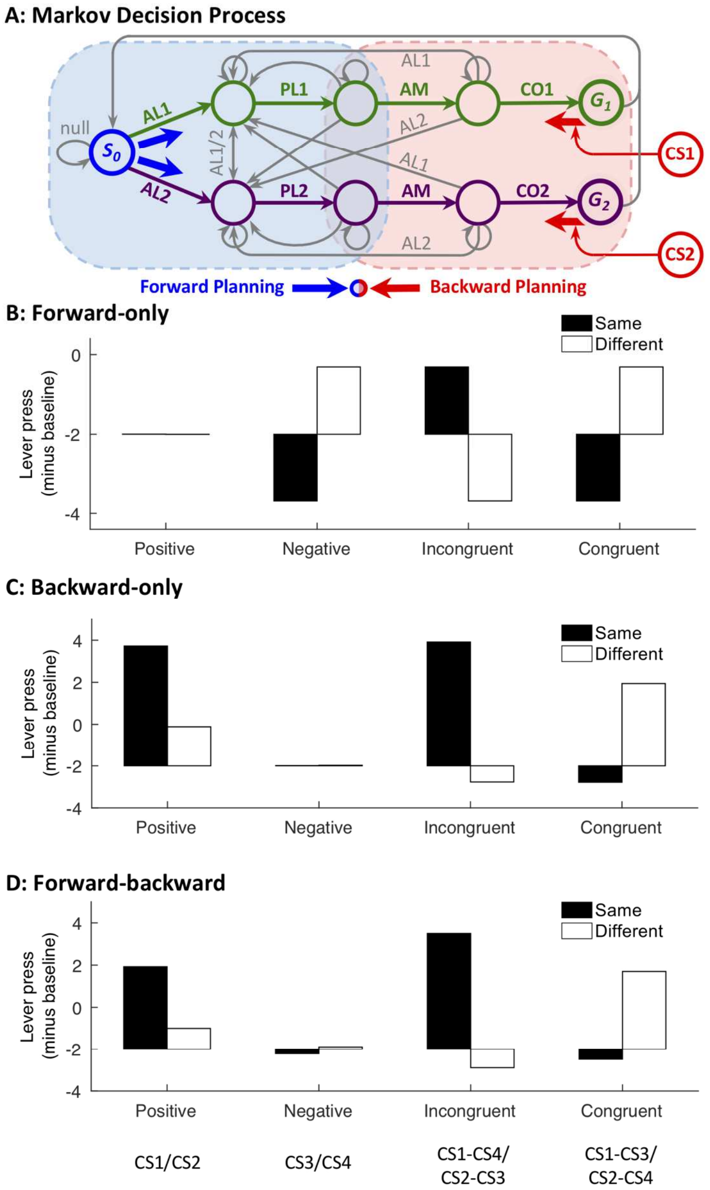
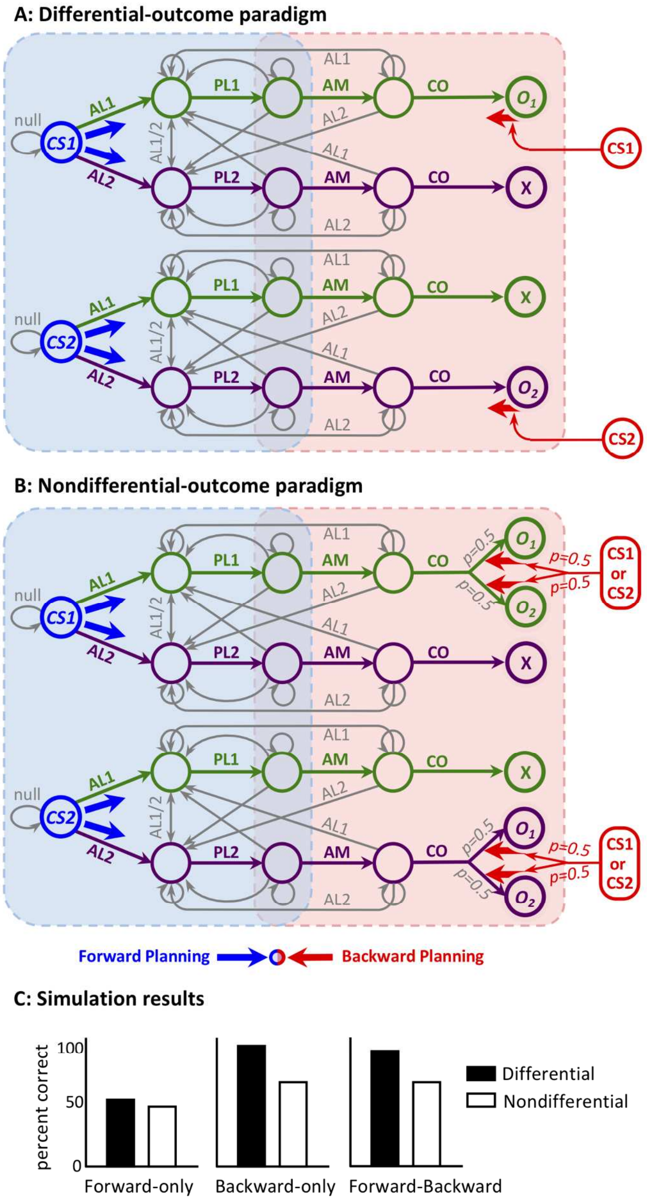


Fig. 4. Backward planning explains factual and counterfactual PIT. (A) The experimental design was captured by a Markov-Decision Process (MDP) where reaching a goal state like G1 requires a sequence of actions including approaching lever number one (AL1), pressing the lever (PL1), then approaching the magazine (AM) and consuming the outcome (CO1). Red and blue boxes show the domain of forward and backward trees when the agent is in the starting state, s0, and the depth of planning is two. Simulating the backward-only (C) and forward-backward (D), but not the forward-only (B), models in this environment captures the essential behavioral patterns observed in rats (Laurent and Balleine, 2015).

160x271mm (300 x 300 DPI)



A: Differential-outcome paradigm

B: Nondifferential-outcome paradigm

C: Simulation results

Fig. 5. Backward planning explains differential outcomes effect. (A) In the differential group, the presented CS is predictive of the outcome (i.e., O1 vs. O2. Receiving no reward is indicated by X) to be received upon choosing the correct action. (B) In the nondifferential group, however, the CS predicts, given the correction action, receiving either of the two outcomes with equal probability. (C) Simulating the forward-only, backward-only, and forward-backward planning algorithms show that backward planning is a critical component for explaining experimental evidence.

177x321mm (300 x 300 DPI)

Supplementary information for:

Behavioral Signatures of Backward Planning in Animals

Authors: Arsham Afsardeir[1], Mehdi Keramati[2,3]*

[1] Control and Intelligence Processing Center of Excellence, School of ECE, College of Engineering, University of Tehran, P.O. Box 14395-515, Tehran, Iran.

[2] Gatsby Computational Neuroscience Unit, Sainsbury Wellcome Centre, University College London, 25 Howland Street, London W1T 4JG, UK.

[3] Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square London WC1B 5EH, UK.

* Correspondence to: mehdi@gatsby.ucl.ac.uk

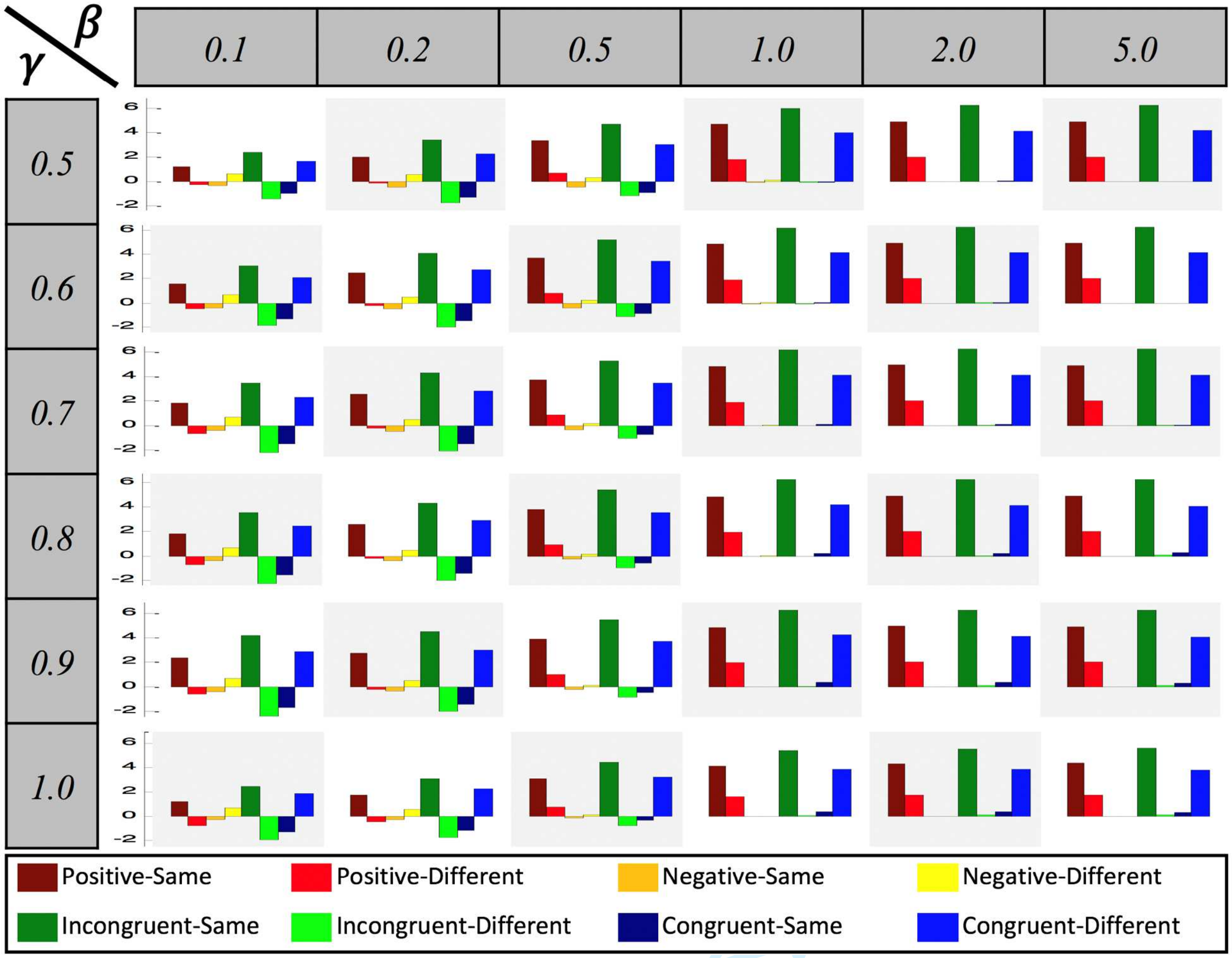
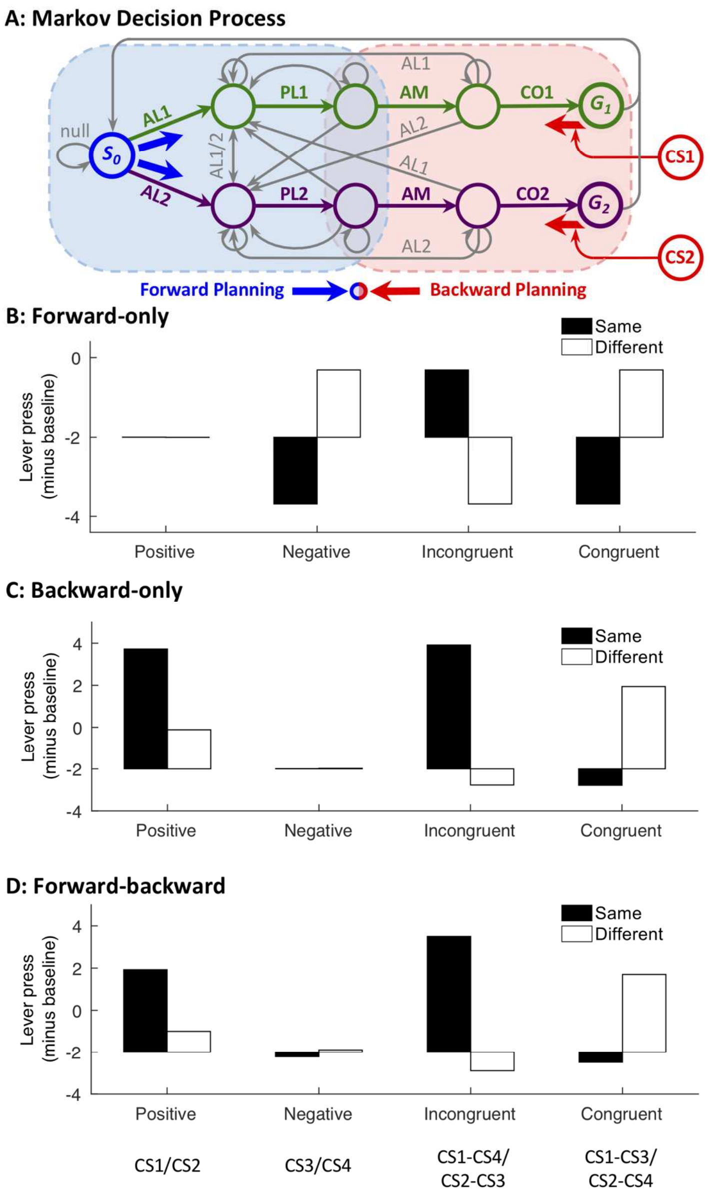


Fig. S2. Parametric exploration of the model for explaining specific PIT. The essential behavioral patterns observed in rats (Laurent and Balleine, 2015) can be replicated by simulating the forward-backward algorithm with a wide range of parameters γ (temporal-discounting factor) and β (inverse temperature in *softmax* action-selection).

We suggest that humans/animals use backward planning: they start planning from a goal state and think backward to find out possible ways to get there. We show that several robust behavioral patterns, namely Pavlovian-to-Instrumental transfer and differential-outcome effect, can be parsimoniously explained by this algorithm.



A: Markov Decision Process

B: Forward-only

C: Backward-only

D: Forward-backward

160x271mm (300 x 300 DPI)