Proceedings of the Eleventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2023)

Accounting for Transfer of Learning Using Human Behavior Models

Tyler Malloy[1], **Yinuo Du**[1,2], **Fei Fang**[2], **Cleotilde Gonzalez**[1]

[1] Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh PA
[2] Software and Societal Systems, Carnegie Mellon University, Pittsburgh PA
tylerjmalloy@cmu.edu, yinuodu@cmu.edu, feifang@cmu.edu, coty@cmu.edu

Abstract

An important characteristic of human learning and decision-making is the flexibility with which we rapidly adapt to novel tasks. To this day, models of human behavior have been unable to emulate the ease and success with which humans transfer knowledge in one context to another. Humans rely on a lifetime of experience and a variety of cognitive mechanisms that are difficult to represent computationally. To address this problem, we propose a novel human behavior model that accounts for human transfer of learning using three mechanisms: compositional reasoning, causal inference, and optimal forgetting. To evaluate this proposed model, we introduce an experiment task designed to elicit human transfer of learning under different conditions. Our proposed model demonstrates a more human-like transfer of learning compared to models that optimize transfer or human behavior models that do not directly account for transfer of learning. The results of the ablation testing of the proposed model and a systematic comparison to human data demonstrate the importance of each component of the cognitive model underlying the transfer of learning.

Introduction

Human behavior models seek to predict and understand how humans generate behavior in a variety of circumstances. These models can be useful for real-world applications of machine learning techniques that interact with people, such as human-in-the-loop control (Cranor 2008; Wu et al. 2022). Often, these models are trained using data from human participants in online experiments designed to account for specific behavior (Zanzotto 2019). For these methods to interact effectively with humans and achieve the goals of the system, models need to be aware of the methods that humans use to improve their learning in these tasks. One example of this is transfer of learning (ToL), which describes the ways that humans learn to learn through the application of previous experience onto new tasks to improve learning performance.

One way that humans often apply ToL is when they are taught a series of tasks to improve their performance on a more difficult task. This is a common approach to training participants in tasks used to develop human behavior models (Little et al. 2009). This method of task decomposition can lead to better performance from human participants, resulting in an increase in the accuracy of human behavior models (Dow et al. 2011). However, providing participants with a series of tasks can alter behavior as a result of the mechanisms that humans use to transfer knowledge between tasks (Bernstein et al. 2010). Failing to account for ToL could result in human behavior models that do not accurately reflect reality.

These training processes may alter behavior in ways that is not accounted for by existing computational models. For example, data from human utility judgments could be used in training recommender systems, pricing models, or financial planners (Bouneffouf, Rish, and Aggarwal 2020). However, existing methods for modeling human behavior in these tasks have not directly accounted for how human ToL may impact behavior.

To address these challenges, we introduce an experiment paradigm for testing human ToL under different conditions. These conditions vary in features that may improve performance on subsequent, more difficult tasks. Alongside this experiment paradigm, we propose a novel model of human behavior that incorporates the cognitive mechanisms of human ToL. This model is used to make predictions about how humans use their previous experience to inform their behavior in new tasks. The specific experiment paradigm is a utility-based learning task with features that inform decision-making. This is an example of a widely used class of learning tasks called contextual multi-armed bandits, in which agents make choices to maximize utility, with contextual cues informing their decisions (Bouneffouf, Rish, and Aggarwal 2020).

Humans rely on a variety of cognitive mechanisms when applying the experience of previous related tasks to new tasks (Gonzalez, Lerch, and Lebiere 2003). Two key processes enabling humans' learning to learn is their ability to compose novel concepts through combinations of previously learned concepts, and discern generalizable causal relationships that describe more than mere correlation (Lake, Salakhutdinov, and Tenenbaum 2013). Another process that is key to ToL is the decision of what content to keep in memory, and what to forget without negatively impacting future behavior (Riemer et al. 2019). The model proposed in this work is inspired by human compositional generalization, causal reasoning, and optimal forgetting. Alternative

models of human behavior have drawn from one or more of these features of human cognition, such as the impact of compositional reasoning and optimal forgetting (Niv et al. 2015). To the best of our knowledge, no existing model has incorporated all of these key mechanisms.

The model presented in this work is inspired by all these cognitive mechanisms that humans use to transfer their experience from previous tasks to new tasks. Ablation testing of the proposed model without each of these features will demonstrate their importance in predicting human-like transfer of learning. For these reasons, the proposed model is better able than existing methods to account for human ToL, resulting in better predictions of their behavior.

We will demonstrate with results of experimentation with human participants that our proposed model accounts for human ToL. Comparisons with alternative models for human learning show the benefits of the proposed model. Additionally, an ablation comparison of different model features supports the importance of each of the components of our proposed model. Finally, an analysis of model parameters when predicting human behavior will give insight into the mechanisms of the model.

Background

Contextual Multi-Armed Bandits

Many applications of machine learning that involve interaction with humans are sequential decision-making problems that can be described by contextual multi-armed bandits, such as recommendation systems, advertising, education, among others (Bouneffouf, Rish, and Aggarwal 2020). General multi-armed bandits are an extension of a sequential decision-making problem with two options introduced in (Robbins 1952), where the available options have different utility probabilities, and the goal is to maximize the observed utility.

This type of choice task with a small number of options available has been an important paradigm for research in decision making. Examples of these previous research areas include comparisons of decisions from experience and description (Barron and Erev 2003; Hertwig et al. 2004), and models of human learning (Gonzalez and Dutt 2011; Niv et al. 2015).

Early applications of multi-armed bandits include clinical trials (Gittins, Glazebrook, and Weber 1989), with algorithm design focusing on optimizing performance by minimizing the *regret* or loss of utility experienced (Kuleshov and Precup 2014). This focus on optimality is reasonable in the highly controlled setting of clinical trials, but for applications that involve interaction with human decision makers, such as prevention of poaching or illegal fishing, it is often more useful to capture the biases and constraints of human decision-making and learning (Fang, Stone, and Tambe 2015).

One method for designing models that better interact with humans is to treat their decisions as being *boundedly-rational* (Fang et al. 2016). Bounded rationality has a long history in cognitive science and human behavior modeling, focusing on an understanding of optimal behavior that takes into account the realities of access to information and the cognitive capacities of humans (Simon 1955). The model presented in this work shares a motivation of boundedly rational analysis by modeling human-like ToL in contextual multi-armed bandit tasks.

Transfer Learning in Machine Learning

In the context of data classification in machine learning (ML), the application of knowledge learned from a previous related task onto a new task is referred to as *transfer learning*. Transfer learning and related adaptation learning, are mathematically formalized in relation to input data $\mathcal{X}$ and output labels $\mathcal{Y}$ (Pratt et al. 1991).

Given the data and label pair $(\mathcal{X}, \mathcal{Y})$, transfer learning supposes a difference between the source and target joint probability distributions $P(\mathcal{X}_{\text{source}}, \mathcal{Y}_{\text{source}}) \neq P(\mathcal{X}_{\text{target}}, \mathcal{Y}_{\text{target}})$ (Zhang and Gao 2022). In *Domain Adaptation*, a special case of transfer learning, there is a difference between source and target marginal distribution $P(\mathcal{X}_{\text{source}}) \neq P(\mathcal{X}_{\text{target}})$ but a similar category space between domains $P(\mathcal{Y}_{\text{source}}|\mathcal{X}_{\text{source}}) = P(\mathcal{Y}_{\text{target}}|\mathcal{X}_{\text{target}})$ (Zhang and Gao 2022).

Decision making problems can be taken to be as an instance of classification, as actions a are assigned onto every possible state S in the same way labels y are assigned onto the data-set X (Kouw and Loog 2018). Under this conceptualization, transfer learning refers to the condition where there is a different optimal policy $\pi^*(s, a)$, which is a function that maps the state of the agent s onto the optimal action a. When this optimal policy is changed between the source and target task, we have $\pi^*_{\text{source}}(s, a) \neq \pi^*_{\text{target}}(s, a)$, and domain adaptation becomes the special case where $\pi^*_{\text{source}}(s) \neq \pi^*_{\text{target}}(s)$ but $\pi^*_{\text{source}}(a|s) = \pi^*_{\text{target}}(a|s)$.

With respect to the specific decision making task of contextual multi-armed bandits, transfer learning can refer to changes in the values associated with choice options. The expected utility of an option $\mathbb{E}[x]$ depends on the features of that option $f \in x$. The transfer learning task can involve a difference in the expected value of entire options $\mathbb{E}_{\text{source}}[x] \neq \mathbb{E}_{\text{target}}[x]$. Alternatively, domain adaptation in contextual bandits involves the same expected utility for some option features $f \in x$, but not all, giving $\mathbb{E}_{\text{source}}[f_0] \neq \mathbb{E}_{\text{target}}[f_0]$ but $\mathbb{E}_{\text{source}}[f_1] = \mathbb{E}_{\text{target}}[f_1]$ for some $[f_0, f_1] \in x$.

The experiments introduced in this paper utilize a domain adaptation approach in the setting of contextual multi-armed bandits. Through the connections of these different definitions of transfer learning, it is possible to relate the behavior of human participants and cognitive models in the proposed experiment onto alternative applications of transfer learning in other domains.

Evaluating Transfer Learning

Methods for evaluating transfer learning in ML in utility learning tasks include a variety of measures of performance (Taylor and Stone 2009). Comparing only one metric of transfer learning may not fully capture the differences between specific models of human behavior and the reality of their behavior. For that reason, we will compare the proposed model using three metrics, jumpstart Performance,

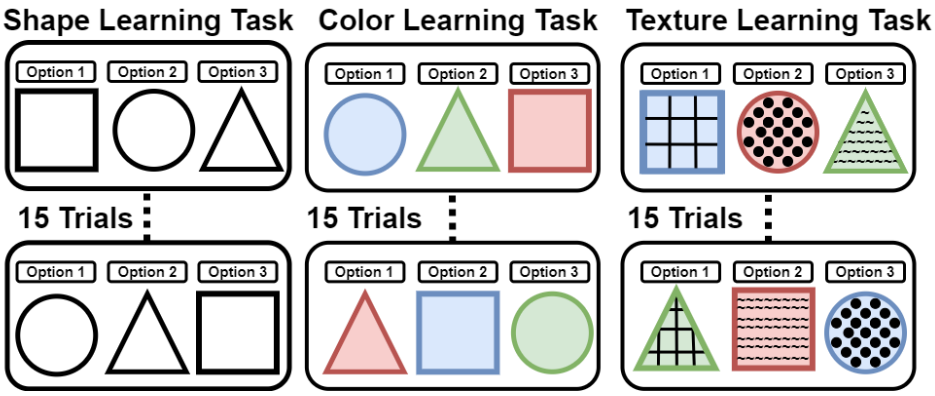

Figure 1: Contextual bandit tasks with shapes (left); shapes and colors (middle); and shapes, colors, and textures (right). An episode of the task consists of 45 total trials completing 15 shape judgements, 15 color judgements, and 15 texture judgements.

Asymptotic Performance, and Episodic Performance. These three metrics represent a sufficient range of types of measurements of transfer learning that are relevant to the utility judgment tasks described by contextual bandits.

Jumpstart Performance is defined as the initial performance of an agent in a target task (Taylor and Stone 2009). In the contextual bandit experiment used in this work, the jumpstart performance is calculated as the average of the first third observed utility in trials after the task switches.

Asymptotic Performance is defined as the final learned performance of an agent in a target task (Taylor and Stone 2009). In the contextual bandit experiment, the asymptotic performance is calculated as the average of the final third observed utility in trials before the next task switch.

Episodic Performance is defined as the average performance over an episode and is analogous to the total reward metric commonly used (Taylor and Stone 2009). This value is calculated as the average of the observed utility over the episode, which will all have the same number of choices in the contextual bandit experiment described later.

Learning Task and Experiment Methods

The contextual bandit transfer learning task, adapted from (Niv et al. 2015), is designed to elicit transfer of learning from human participants. There are three learning tasks that increase in difficulty as the task progresses. In all three types of tasks, there is one choice option that has a higher expected utility compared to the other two, based on its features. Participants are tasked with selecting the higher utility option. Participants used their keyboard to select one of the three options (A, S, and D) and had an unlimited time to make their selection, with an mean and standard deviation of reaction time of 1173 ± 2121 ms. After the participant makes their choice selection, they observed the utility feedback for 1 second before proceeding to the next choice selection.

Initially, the contextual bandit consists of three options each with one of three possible shapes (square, circle, triangle). One of the three shapes is randomly associated with a higher expected utility of 6 compared to the other two shapes which have an expected utility of 4. Utility values have a Gaussian noise added $\mathcal{N}(1, 1)$ to introduce difficulty to the learning task.

This learning task begins with 15 selections of the three options with shape features (left: Fig. 1). After each selection the order of the shapes is randomized so that the same shape does not always appear in the same position. This requires the agent to learn the values associated with the option features, and not the position of each option individually.

The experiment progresses to the second learning task including shape-color options (red, green, blue shown in middle: Fig. 1) and then finally the shape-color-texture learning task (hatched, dotted, wavy shown in right: Fig. 1). Each learning task consisted of 15 trials for a total of 45 trials in each experiment block.

The experiment consists of 6 total experiment blocks of the same ToL condition (explained in the next section). At the end of each learning task of 15 trials, participants were asked which of the features was associated with a higher utility outcome. After the end of all experiment blocks, participants were asked three questions to determine their awareness of the type of ToL condition of the task.

ToL Conditions

There are three experimental ToL conditions that vary the value of the features across the learning tasks. A table of the exact utility calculations for each condition is provided in the supplementary materials. The difference between these conditions is related to our main hypotheses and research question regarding the difficulty of differences in transfer of learning, as well as the ability of our proposed model to capture these differences in difficulty.

Positive Transfer: In this condition, the high utility shape continues to be associated with a higher utility in the shape-color and shape-color-texture trials. This means that if a square is associated with a higher expected utility initially, then red squares will have a **higher** expected utility than red triangles. The same is true for the higher utility color once texture is introduced, meaning that red squares of any texture will have a higher expected utility than other shape-color pairs of the same texture.

Negative Transfer: In this condition, the high utility shape becomes associated with a decrease in utility in the shape-color and shape-color-texture trials. This means that is a square is associated with a higher expected utility initially, then red squares will have a **lower** expected utility than red triangles. The same is true for the higher utility color once texture is introduced, meaning that red squares of any texture will have a lower expected utility than other shape-color pairs of the same texture.

Null Transfer: In this condition, the shape that was previously associated with a higher utility is irrelevant for the expected utility of an option. This means that is a square is associated with a higher expected utility initially, then red squares will have **the same** expected utility as red triangles. The same is true for the higher utility color once texture is introduced, meaning that red squares of any texture will have the same expected utility of other shape-color pairs of the same texture.

This learning task allows for a comparison of how agents and human participants transfer their previous experience onto learning the values of options. Some biases and assumptions that are made when transferring previous experience may be helpful for learning the transfer conditions, and some may make learning more difficult. This experiment structure allows for comparison of human behavior models and the behavior of human participants based on the metrics of transfer learning previously mentioned.

When learning the contextual bandit transfer task, agents and human participants will have 6 total blocks, each block consisting of the three learning tasks: shape, color, and texture features. For a single participant, on all experiment blocks the type of transfer (positive, negative, and null) will be the same, allowing the agent to learn the transfer type over the course of the 6 experiment blocks.

Regarding ToL conditions, we hypothesized that the positive condition conforms most to participants expectations of the continued relevance of previously learned features. Then the null condition somewhat breaks this assumption by having previously learned features be irrelevant. And finally the negative transfer condition should go against participant assumptions to the highest degree. We additionally hypothesized that in all three conditions, participants would be able to improve their performance across the experiment blocks, as they learn the type of ToL they are engaging in.

In the contextual bandit transfer learning task, there is a slightly different maximum observable utility based on the direction of the transfer (positive, negative, null). For that reason, we report adjusted reward measures for all metrics of performance and ToL ability for the remainder of the paper. These adjusted rewards are calculated by scaling the two lower maximum reward conditions to the same scale as the highest.

Baseline Behavior Models

The following baseline behavior models are used as a comparison against the proposed model. After describing each of these approaches, their behavior in the transfer learning contextual bandit will be presented. In the following section, these models will be further compared in their ability to predict individual participant's behavior by fitting model parameters to a subset of human behavior, and using those models to predict unseen human participant decisions. All model implementations uses an action soft-max temperature parameter β to control the likelihood that the model assigns onto options based on their value estimate.

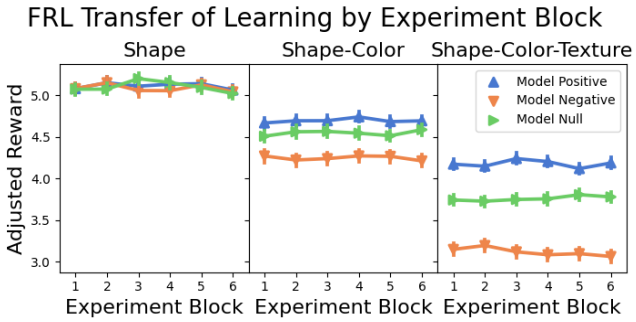

Figure 2: Mean adjusted reward observed by the FRL model by transfer condition. Error bars are standard deviation.

Feature Reinforcement Learning (FRL)

Feature Reinforcement Learning ((Niv et al. 2015)) is an extension of the framework described by Reinforcement Learning (RL) which seeks to solve a problem description defined by a Markov Decision Process (MDP) consisting of an agent performing actions in a changing environment and observing a reward (Sutton and Barto 2018). The goal of the RL agent is to maximize the reward they observe while acting within the MDP by choosing actions that result in high rewards, as well as favorable subsequent states. An MDP is formally described by a state in a set of states $s \in S$ and an action in a set of actions $a \in A$.

Feature Reinforcement learning is a human behavior model for contextual bandit games that learns the values associated with options features according to the equation (Niv et al. 2015):

$$W(f) \leftarrow W(f) + \alpha[o - V(X_{\text{chosen}})] \; \forall f \in X_{\text{chosen}} \quad (1)$$

where α is the learning rate, and o is the outcome from selecting option X_{chosen}. The FRL model functions by predicting the value of an option as the sum of the values of each of the features of that option. $V(X) = \sum_{f \in X} W(f_i)$ Instead of modeling the way in which humans use their previous experience to learn the causal structure of contextual bandit tasks, this model assumes that participants are given that information (Niv et al. 2015).

The model behaviour results in Figure 2 compare the adjusted reward of the experiment blocks in trials of shape judgements (left), then shape-color (middle), then shape-color-texture (right). This format will be used to compare model and human participant ToL ability across the three types of ToL (positive, negative, and null).

From the behavior of the FRL model, we can observe an effect in the type of ToL on the performance in the more challenging shape-color and shape-color-texture tasks. However, there is no improvement over the course of the full

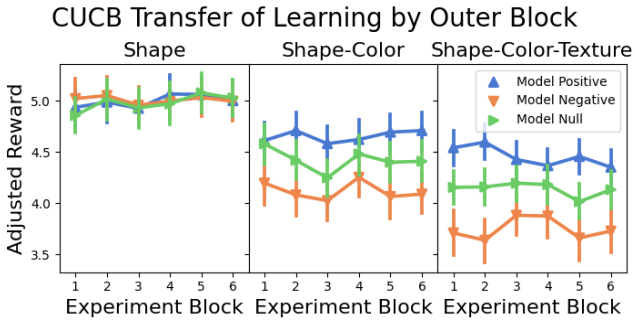

Figure 3: Mean adjusted reward observed by the CUCB model by transfer learning condition. Error bars represent standard deviation.

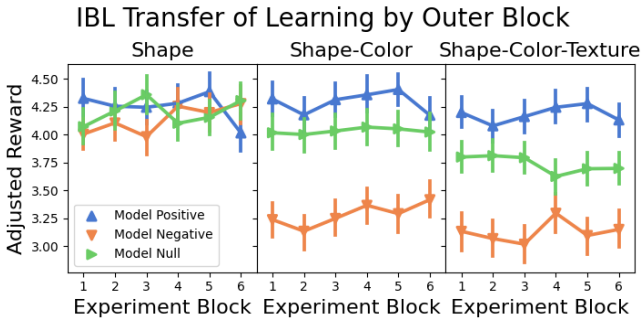

Figure 4: Mean adjusted reward observed by the IBL model across experiment blocks, by transfer learning condition. Error bars represent standard deviation.

experiment in any of the transfer learning conditions. This is due to the assumption made by the FRL model of the function that maps feature values onto option values.

Causal Upper Confidence Bound (CUCB)

The Upper Confidence Bound (UCB) algorithm (Sutton and Barto 2018), selects options X_t by balancing the expected utility measure $Q_t(x)$ with the number of times that option has been selected $N_t(x)$ according to the equation:

$$X_t = \text{argmax}_x \left[Q_t(x) + c\sqrt{\frac{\ln t}{N_t(x)}} \right] \tag{2}$$

where $N_t(a)$ is the number of times the action a has been taken at time t, $Q_t(a)$ is the mean value of the action a, and c is a parameter that balances action selection.

Upper confidence bound models can be structured to learn the type of causal relationship that is required to transfer learning in the contextual bandit setting of this experiment (Zhang and Bareinboim 2017). This Causal UCB (CUCB) learning model attempts to optimally transfer learned causal relationships in contextual bandit tasks.

The behaviour of the CUCB model in Figure 3 does display a difference in performance based on the transfer condition. However, the optimal transfer learning of the CUCB means that it learns the transfer condition quickly, within the first experiment blocks, and does not improve performance throughout the experiment blocks. Additionally, there is a large standard deviation of model behavior, likely due to the Stochastic nature of the contextual bandit utility function.

Instance-Based Learning (IBL)

The Instance Based Learning model used for comparison of behavior in the contextual bandit ToL experiment is based on Instance Based Learning Theory (IBLT) (Gonzalez, Lerch, and Lebiere 2003). This theory is described in full in the supplementary material. The implementation of an IBL model in this work predicts the value of options in the contextual bandit task according to the value function:

$$V_{k,t} = \sum_{i=1}^{n_{k,t}} p_{i,k,t} x_{i,k,t} \tag{3}$$

where x are the outcomes, and the probability of retrieval is $p_{i,k,t}$. This value estimation requires a complete memory of the history of all option selections and utility outcomes.

In the IBL model used in this work, we take the model instance in the contextual bandit task as consisting of three features (shape, color, and texture). The IBL model uses a similarity function that matches experience based on the number of features that match, weighted to account for instances in memory that do not have all three features.

The IBL model uses parameters to model memory decay d, by decaying older memory instances from contributing to the activation value, and behavioral noise σ that adds a random value to the activation.

The behavior of the IBL model shown in Figure 4 shows stable performance on each ToL condition, without improving over the experiment blocks. This is likely because the IBL model does not explicitly rely on compositional reasoning. It is possible that an alternative method of measuring similarity could capture some compositional-like reasoning, though that was not observed in the method used in this implementation of IBLT.

Causal-Compositional RL (CC-RL) Model

The proposed model is based on Reinforcement Learning, as was the case with the similar FRL model. Where the proposed model differs from FRL is that it does not assume prior knowledge of a function that maps contextual bandit option features onto their utility. Instead, the model uses compositional reasoning and causal inference to learn this relationship in a way that can be applied onto ToL tasks.

Compositional Reasoning

Compositional Reasoning would allow an RL model to systematically apply transfer of learning by creating novel concepts through composition of previously learned concepts (e.g. “red” and “square” can be combined to form the concept of “red square”). Humans use compositional reasoning in a variety of domains related to language such as letter writing and sentence generation (Lake, Salakhutdinov, and Tenenbaum 2013; Piantadosi, Tenenbaum, and Goodman 2016). Recent methods in reinforcement learning have sought to apply these concepts to improve generalization

in machine learning methods (Malloy et al. 2022; Ito et al. 2022).

We incorporate compositional reasoning by utilizing a special case of MDPs, the problem setting of RL, called the Factored Markov Decision Process (FMDP). The FMDP setting is a special case of MDP formed by relating it to a *dynamic Bayesian network* defined by a *directed acyclic graph* G_T with nodes $\{X_1, X_2, ..., X_n\}$ and scopes $S_1, ..., S_n$ (Kearns and Koller 1999). A scope S_i of this network describes the dependencies of future state features or rewards based on previous features and actions, with $x[S_i]$ signifying the features of state x corresponding to the scope S_i. This allows for a definition of the factored reward function $R(x)$ as follows (Sallans and Hinton 2004):

$$R(x) = \frac{1}{n}\sum_{i=1}^{n} R_i(x[S_i]) \tag{4}$$

In the proposed model, the strengths of these causal relationships are updated in a manner similar to traditional RL (Equation 1):

$$Q(x[S_i]) \leftarrow Q(x[S_i]) + \alpha(Q(x[S_i]) - \Phi_i o_{m,n}) \tag{5}$$

where α is a parameter controlling the speed of the factored q-function update, and Φ_i is the weight of the option feature S_i, which describes the causal relationship between features and outcomes. These factored representations can be leveraged to significantly improve sample efficiency when the causal structure is provided (Chen et al. 2020). However, it can be difficult to learn these factored representations from scratch (Malloy et al. 2022). The proposed CC-RL model addresses this issue using causal inference as a method of update the feature weights Φ.

Causal Inference

Causal inference involves the understanding of causal relationships of events as happening as a result of the factors that caused that event (Lake et al. 2017). In contextual bandits, causal inference involves learning how the actions performed by the agent are causally linked to outcomes they observe (Lattimore, Lattimore, and Reid 2016).

It is possible to learn a strategy in bandit games that does not attempt to represent the causal relationship of the task, by learning the correlation of contexts and outcomes (Sutton and Barto 2018). However, this correlation learning is not as generalizable as a model of the causal relationships that can be applied onto a learned compositional representations to allow for systematic generalization of learned concepts (Malloy and Gonzalez 2023).

The three relationships shown in Figure 5 demonstrate three possible instances of a causal relationship. These relate to contextual bandits by taking the Xs to be the features of choices, and the Ys to be outcomes. The structure of the causal relationship is defined by the connections between option features and the outcome. The weights Φ_i are updated based on the relative difference of the historic mean of the outcomes from selecting that feature $\bar{\Phi}_i$ and the most recent outcome $o_{m,n}$ through the equation:

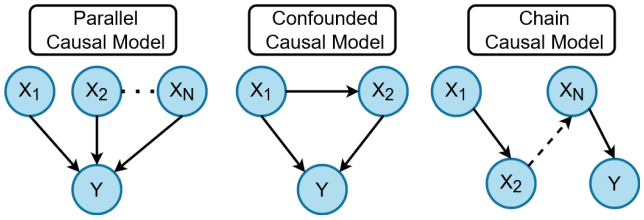

Figure 5: Three causal relationships, as graphs. Adapted from a figure in (Lattimore, Lattimore, and Reid 2016).

$$\Phi_i \leftarrow \bar{\Phi}_i + \omega(\bar{\Phi}_i - o_{m,n}) \tag{6}$$

Where ω is a parameter that controls the speed of the weight updating. This method of updating weights based on the historic mean of outcomes is related to human behavior models based on reinforcement learning that add episodic memory to improve the speed of learning new related tasks (Gershman and Daw 2017).

With this updating method, the causal structure is initially assumed to be fully connected, such that all features of a option are equally causally linked to the outcome. Causal links are updated based on experience by removing the causal links $x[S_i]$ according to the following process:

$$x[S_i] \quad \forall_i \Phi_i < \tau \tag{7}$$

Where τ is a threshold for relevancy of a causal link, set to $1 * 10^{-6}$ and Φ_i is the vector of weights that is used in the factored reward function defined in Equation 7. This vector is updated based on the predictions of the factored reward and utility outcomes according to Equation 5.

When these weights drop below the threshold, the causal relationship is updated to reflect the lack of a causal connection between option features and the outcome. We expect that this process would allow the CC-RL model to learn the causal relationship between context features and outcomes, and reflect human-like ToL between contextual bandit tasks.

Optimal Forgetting

The concept of optimal forgetting has a long history within cognitive science as a tool for learners to optimize their decision making by choosing some information to forget (Underwood 1957). In the proposed model, this optimal forgetting is implemented by a decay parameter δ that controls a decrease in the predicted value of the option features that are not chosen by the model. This allows for a bias in the estimation of features that are not experienced as often as other features, done using the equation:

$$\forall_{S_j \notin x_{m,n}} Q(x[S_j]) \leftarrow \delta Q(x[S_j]) \tag{8}$$

Where $x_{m,n}$ is the option selected in the step n of episode m, and δ is a parameter that can be fit to optimally decay options, reflecting human biases in remembering more recent events. The result is that the decay is only applied onto the features $Q(x[S_j])$ that are not within the selected option.

Algorithm 1: Causal Compositional RL (CC-RL)

Require: Options X, Features F
  Initialize weights $\Phi \leftarrow [1, ..., 1]$
  Initialize fully-connected DAG $x[S_i] \leftarrow S \; \forall_i$
  Initialize $Q(x) = \frac{1}{n} \sum_i^n \Phi_i * Q_i(x[S_i])$
  for episode $m = 1, ..., M$ **do**
    for step $n = 1, ..., N$ **do**
      Select option $x_{m,n} = argmax_{x \in X} Q(x)$
      observe outcome $o_{m,n}$
      $\Phi_i \leftarrow \bar{\Phi}_i + \omega(\bar{\Phi}_i - o_{m,n})$
      $Q(x[S_i]) \leftarrow Q(x[S_i]) + \alpha(Q(x[S_i]) - \Phi_i o_{m,n})$
      $\forall_{S_j \notin x_{m,n}} \; Q(x[S_j]) \leftarrow \delta Q(x[S_j])$
    end for
    Prune DAG $x[S_i] \; \forall_i \Phi_i = 0$
    Reset $Q(x)$
  end for=0

CC-RL Algorithm

The CC-RL model (Algorithm 1) for predicting human ToL incorporates the three proposed equations as follows:

Model Biases

CC-RL uses weights to balance the relevance of different option features on utility outcome predictions, and these weights are initialized with a value of 1. This initialization results in a bias that reflects a human-like assumptions that all features are relevant.

Initially, there is a fully connected DAG used to represent the causal relationship between option features and utility outcomes (see Figure 5). This fully connected initialization represents the presumption that human learners are initially biased to assume that all features of their choices are relevant to outcomes, and only update this causal inference assumption based on experience.

The initial structure of the weights and DAG of this model results in two assumptions regarding the behavior of human participants, as previously mentioned. Additionally, the method of updating these weights, and pruning the nodes of this DAG, makes specific predictions of human behavior. It is possible that human learning is more flexible than this, and

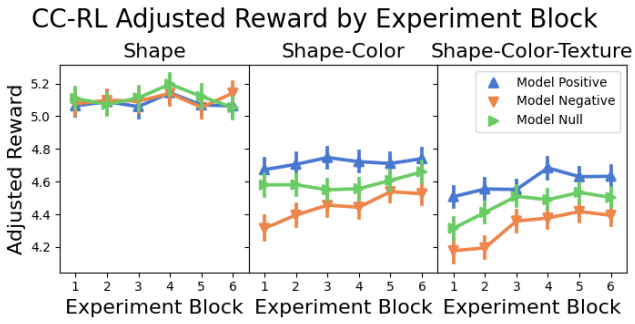

Figure 6: CC-RL model adjusted reward by transfer learning condition compared. Error bars are standard deviation.

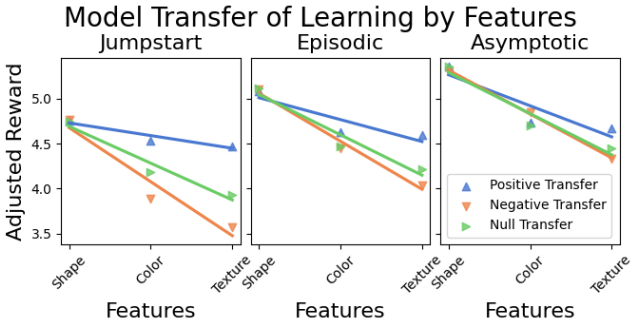

Figure 7: CC-RL performance on ToL metrics by the type of transfer, across shape, color, and texture decisions.

that alternative representations of weight updates or DAG pruning are more fitting. However, for the contextual bandit experiment in this paper, it is reasonable that participants may assume a fully connected causal structure initially.

CC-RL Model Results

The CC-RL model behavior in the contextual bandit ToL task (Fig. 6), demonstrate the increased difficulty of the shape-color and shape-color-texture learning conditions, as indicated by the lower adjusted reward. Additionally, the degree of this decrease in performance depends on the type of transfer learning.

The highest performance on the more difficult tasks is observed in the positive transfer condition, then null transfer, then negative transfer. Additionally, the proposed model is able to improve performance on all three transfer conditions throughout the experiment, indicating that it is displaying a ToL ability.

Comparing the adjusted reward of the three conditions based on the three metrics of ToL shows the largest difference between ToL conditions in the jumpstart performance metric (Figure 7). As the metric of comparison takes into account decisions made when feature utility is learned by the model, the episodic and then asymptotic reward metric shows a decrease in the difference between transfer conditions.

Finally, comparing the asymptotic performance shows the least difference based on the ToL condition. These modeling results indicate that the proposed model demonstrates a larger initial effect (Jumpstart) on performance based on the type of ToL, but that this effect is diminished as the metric of comparison attends more to decisions made later in the trial block.

Human Experimentation

This section describes the experiment involving human participants learning the proposed contextual bandit task involving different types of ToL. This experiment paradigm was designed to compare our proposed CC-RL method against alternative models of human behavior and machine learning methods for ToL.

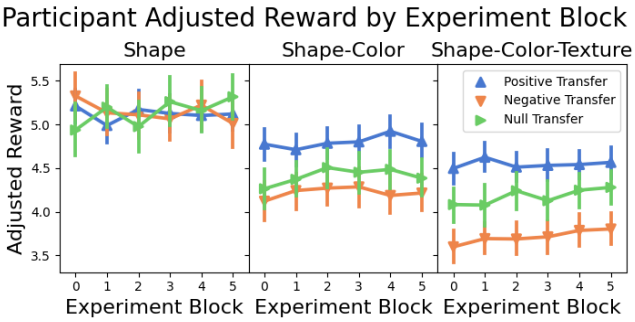

Figure 8: Participant transfer of learning by experiment block, divided between learning shape, then color, then texture and the type of ToL condition. Error bars represent standard deviation.

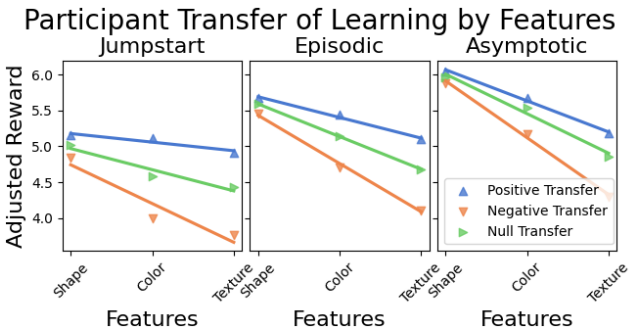

Figure 9: Participant performance on three ToL metrics by the type of transfer, on the shape, color, and texture tasks.

Methods

180 participants were recruited online through Amazon Mechanical Turk, 3 participants failed to complete the study and their responses were not collected. All participants resided in the United States of America and were above the age of 18. The mean participant age was 40.35 and the standard deviation of ages was 11.33 years. There were 86 female, 69 male, and 2 non-binary respondents. Participants were paid a base payment of $4, with a possible additional bonus of up to $3 based on performance in the learning task.

Participants completed a demographics form before reading through instructions on the game design. After this, there was a short quiz on the contents of the instructions that participants needed to get all correct before continuing onto the main experiment. This experiment was approved by the Carnegie Mellon University Internal Review Board. The experiment protocol was preregistered on OSF, and all participant data and a full experiment protocol OSF[1]

The first learning task was the shape learning in which participants selected one of the three shapes available and observed a reward. No time restrictions were imposed. After selecting one of the three choices, participants observed utility outcome feedback for 2 seconds. Then participants progressed to the color task and then the texture task. After the experiment, participants completed a short questionnaire to test their awareness of the ToL condition.

Human Experiment Results

Results from human participant behavior (Figure 8) in this experiment demonstrate key features of ToL. The order of performance for the positive, null, and negative transfer conditions conforms to our hypothesis regarding the difficulty of each condition. The positive transfer condition conforms the most to expectations of the relevance of previous experience and is the easiest to learn. Similarly, the null transfer condition is slightly more difficult to learn, and the negative transfer condition which reverses humans' assumptions of continued relevance, is the most difficult.

[1]https://osf.io/mt4ws/

The difference between participant performance in these three conditions is related to the second main effect of the transfer of learning, which takes place within a single experiment block. When participants learn the value of the shape first and then apply it to the decisions they make when presented with shape-color options and shape-color-texture options, they demonstrate their transfer of learning within an experiment block. This is a key component of human generalization, as it describes the biases and assumptions that humans make about how their previous experience relates to future situations.

Figure 9 compares participant ToL by the learning of shape, color, then texture. These results demonstrate the impact of the transfer condition on the three main metrics of ToL, jumpstart, episodic, and asymptotic performance. At the beginning of trails, there is a large difference in the performance of participants based on the ToL condition. However, this difference narrows as the metric of comparison includes more of the later trials in the episodic and then asymptotic performance metrics. This indicates that human participants are able to improve their episodic and asymptotic performance on the shape-color and shape-color-texture trials relative to their jumpstart performance.

Model Predictive Accuracy

To compare the baseline models and the ablated versions of the CC-RL model against baseline human behavior, we compared models based on their accuracy in predicting participant behaviour. To determine a model's predictive accuracy, model parameters are fit for each individual participant by minimizing the Negative Log Likelihood (NLL) of the probability $p(x)$ that the model assigns onto the choice that was selected by the participant, using the Python Scipy library (Virtanen et al. 2020). Then, those model parameters are used to predict a different set of participant decisions. This is a common evaluation method in human behavior models of contextual bandit tasks (Niv et al. 2015; Niv 2019).

Table 1 lists all model parameters that are fit for each model. A longer description of each fit model parameter can be found in the sections describing each model, and in the supplementary material.

In Figure 10, there are three different splits of data for fitting model parameters and data for predicting human partic-

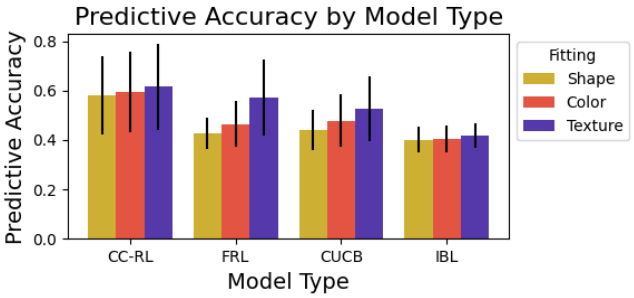

Figure 10: Fit accuracy: model's predictive accuracy when fitting model parameters to human decisions made only in one of the three types of decisions, and using those parameters to predict behavior in the other conditions.

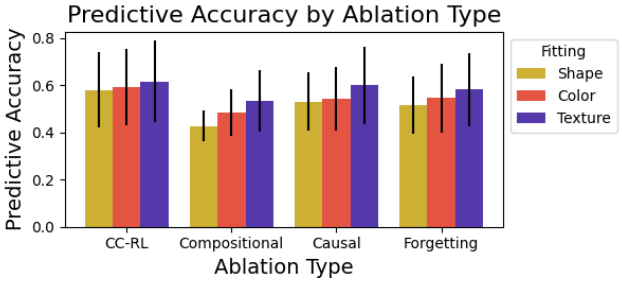

Figure 11: Accuracy of ablated versions of the CC-RL model when fitting model parameters to only human decisions made in one of the three types of decisions, and using those parameters to predict behavior in the other conditions.

ipant behavior. The splits used for fitting model parameters are the shape decisions, color decisions, and texture decisions. For each type of parameter fit, the predictive accuracy is calculated by predicting the two types of left out decisions (i.e color and texture are predicted based on a model with parameters fit using shape decisions, etc.).

Shape Decisions: The first comparison of model accuracy in predicting human behavior is done by fitting model parameters to the decisions made in the during the trials that only contain shape features. This parameter fitting is done for each individual participant, while only looking at the data for their shape utility decisions. Results from model predictions of this first fitting condition demonstrate that the proposed model has a higher predictive accuracy of all models under comparison. Additionally, this first fitting condition has the widest margin of improved predictive accuracy afforded by the proposed model.

Color Decisions: The second comparison of model accuracy is done by fitting model parameters to only the participant decisions made when making judgements of the shape and color task. Results from this fit demonstrate a similar trend as the shape behavior fitting results, with a slight increase in predictive accuracy of all models.

Texture Decisions: The final comparison of model accuracy is done by fitting model parameters to only the participant decisions made when making judgements of the shape-color-texture task. This is likely because the final task was the most difficult for participants, and fitting parameters to these decisions allowed for accurate predictions of the two easier tasks.

A mixed repeated measure analysis of variance of the effect of model type and parameter fitting condition on predictive accuracy demonstrated significant variation of model ($F = 248, p < 0.0001$) and condition ($F = 92, p < 0.0001$). A post-hoc multi-comparison Tukey test showed that the CC-RL model performance was significantly higher than each of the three baseline models (FRL $p = 0.0081$, CUCB $p = 0.0046$, IBL $p < 0.0001$).

Model	Learning	Temp	Noise	Decay	Unique
CC-RL	α	β	N/A	δ	ω
FRL	α	β	η	δ	N/A
CUCB	N/A	β	η	N/A	c
IBL	N/A	β	η	δ	σ

Table 1: Table of parameters that are fit for each model.

CC-RL Ablation Testing

To evaluate the importance of each of the three proposed cognitive mechanisms in producing human-like behavior in this ToL task, we compared three ablated versions of the proposed model and the results are shown in Figure 11.

In each case of ablation, the result is a model with one fewer parameter to fit to match human behavior. However, since we split the fitting of parameters and testing predictive accuracy based on the type of decision being made, there shouldn't be a clear effect of removing the a fit parameter beyond the impact on the model lacking the related mechanism.

No Compositional Reasoning: Removing the compositional reasoning faculty of the proposed model has the most drastic effect on the structure of the model. Without compositional reasoning, the model treats each shape as a novel object, without applying the previous experience that has been gained with options that have the same features.

No Causal Inference: The removal of causal inference requires less changes to the model structure compared to removing compositional reasoning. This can be achieved by setting the weight updating parameter ω to zero. This will result in a model that does not update the weights Φ_i, and instead leaves them at their original values.

The predictive accuracy of the resulting model shows a decrease in performance compared to the full proposed model. However, this decrease is not as significant as with the removal of compositional reasoning. This is likely due to the fact that the resulting model can somewhat compensate for the lack of a weight updating method by predicting human actions that match predict accuracy close to the mean of participant behavior as it improves.

No Optimal Forgetting: To remove optimal forgetting from the proposed model, it is simple enough to remove the feature value decay process of decaying the option features

that are not contained in the option chosen by the participant. The resulting predictive accuracy is slightly decreased from the full model across each fitting condition. This indicates that optimal forgetting may be less relevant for this specific contextual bandit ToL formulation. However, this may be due to the fact that in two out of the three conditions (positive and null), forgetting the previous learned feature values was not as relevant as in the negative condition.

A mixed repeated measure analysis of variance of the effect of model ablation type and parameter fitting condition on predictive accuracy demonstrated significant variation of model ($F = 8.171, p < 0.0001$) and condition ($F = 138, p < 0.0001$). A post-hoc multi-comparison Tukey test showed that the full CC-RL model performance was significantly higher than each of the three ablations (compositional $p < 0.0001$, causal $p = 0.0009$, forgetting $p = 0.0042$). These results demonstrate the crucial importance of each of the three features of the proposed CC-RL model in predicting how humans transfer their learning in contextual bandit tasks.

Altering Existing Models

One concrete test of the contribution of the proposed cognitive mechanisms to ToL is to alter one of the existing baseline models to incorporate that mechanism and observe the results. We introduced a compositional structure to the IBL model by training three separate models and combining these predictions compositely as is done in the proposed model in Equation 5.

The result is a Compositional Instance Based Learning (C-IBL) model that incorporates a compositional reasoning structure into the contextual bandit task. Additionally, a causal inference process can applied to this C-IBL model to additionally learn the weights of the causal relationship, resulting in the Compositional Causal Instance Based Learning (CC-IBL) model.

The results in Figure 12 demonstrate the improvement in predictive accuracy that is afforded by applying the causal compositional methodology onto the baseline instance based learning model. This highlights the usefulness of modeling causal inference and compositional reasoning.

A mixed repeated measure analysis of variance of the effect of adding compositionality and causality to the baseline Instance Based Learning model on predictive accuracy demonstrated significant variation of model ($p < 0.0001$) and condition ($p = 0.036$). A post-hoc multi-comparison Tukey test showed that the CC-IBL model performance was significantly higher than each of the other two models (causal $p < 0.0001$, baseline $p < 0.0001$) but an insignificant difference for the C-IBL and baseline IBL models ($p = 0.9$).

These results provide further evidence that compositionality alone is not enough to capture human-like ToL capability. Irrespective of the type of base model that is used to incorporate causality and compositionality, both are required to accurately predict human transfer of learning. Instead, both causal inference and compositional reasoning are necessary for cognitive models of human ToL.

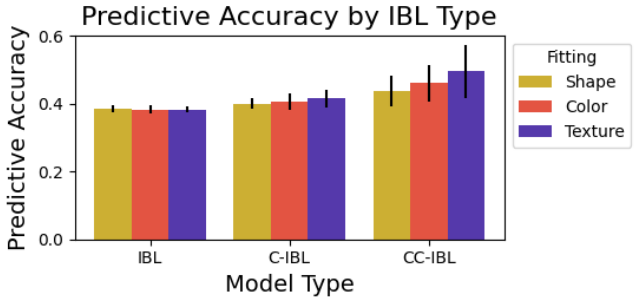

Figure 12: Predictive accuracy of IBL model, Compositional IBL model, and Causal Compositional IBL model. Error bars represent standard deviation.

Conclusions

Computational models of human behavior are applied onto a variety of domains, and all require predictions of human behavior that are as accurate as possible. However, the computational processes behind human ToL are typically not incorporated into human behavior models.

To better understand both human ToL and how it can be modelled computationally, we developed a novel experiment in a contextual bandit setting. Alongside this experiment, we proposed the CC-RL model that applies compositional reasoning, causal inference, and optimal forgetting onto predicting human behavior while accounting for transfer of learning.

Results comparing the accuracy of CC-RL predictions of human behavior in this task demonstrated a higher accuracy compared to related models of human behavior. The cognitive mechanisms that inspired the different components of our proposed model were compositional reasoning, causal inference, and optimal forgetting. These mechanisms had previously been related to human learning-to-learn and generalization in a variety of domains related to language and function learning. Our proposed model extends the application of these cognitive mechanisms onto understanding human ToL in a contextual bandit task.

Comparison with alternative accounts of human learning demonstrated that methods that do not directly account for human ToL cannot accurately reflect how humans apply it onto contextual bandit tasks. An ablation analysis confirmed the importance of each feature of our model in instantiating the proposed cognitive mechanism behind transfer of learning for utility-based tasks. Additionally, altering an IBL model to incorporate the missing cognitive mechanisms of compositional reasoning and causal inference demonstrated improved accuracy in predicting the behavior of human participants in our ToL task.

Future research in human ToL should seek to understand how humans apply the cognitive mechanisms behind transfer of learning onto tasks with different difficulty. The experiment in this work involved increasing difficulty and complexity, but previous experience can also be applied onto more simpler tasks, or tasks of similar complexity.

Acknowledgements

This research was sponsored by the Army Research Office and accomplished under Australia-US MURI Grant Number W911NF-20-S-000 and by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA)

References

Barron, G.; and Erev, I. 2003. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of behavioral decision making*, 16(3): 215–233.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322.

Bouneffouf, D.; Rish, I.; and Aggarwal, C. 2020. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. IEEE.

Chen, X.; Hu, J.; Li, L.; and Wang, L. 2020. Efficient Reinforcement Learning in Factored MDPs with Application to Constrained RL. In *International Conference on Learning Representations*.

Cranor, L. F. 2008. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, 1–15.

Dow, S.; Kulkarni, A.; Bunge, B.; Nguyen, T.; Klemmer, S.; and Hartmann, B. 2011. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 1669–1674.

Fang, F.; Nguyen, T.; Pickles, R.; Lam, W.; Clements, G.; An, B.; Singh, A.; Tambe, M.; and Lemieux, A. 2016. Deploying paws: Field optimization of the protection assistant for wildlife security. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 3966–3973.

Fang, F.; Stone, P.; and Tambe, M. 2015. When security games go green: designing defender strategies to prevent poaching and illegal fishing. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 2589–2595.

Gershman, S. J.; and Daw, N. D. 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, 68: 101–128.

Gittins, J.; Glazebrook, K.; and Weber, R. 1989. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Gonzalez, C.; and Dutt, V. 2011. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, 118(4): 523.

Gonzalez, C.; Lerch, J. F.; and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4): 591–635.

Hertwig, R.; Barron, G.; Weber, E. U.; and Erev, I. 2004. Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8): 534–539.

Ito, T.; Klinger, T.; Schultz, D.; Murray, J.; Cole, M.; and Rigotti, M. 2022. Compositional generalization through abstract representations in human and artificial neural networks. *Advances in Neural Information Processing Systems*, 35: 32225–32239.

Kearns, M.; and Koller, D. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, 740–747.

Kouw, W. M.; and Loog, M. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

Kuleshov, V.; and Precup, D. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

Lake, B. M.; Salakhutdinov, R. R.; and Tenenbaum, J. 2013. One-shot learning by inverting a compositional causal process. *Advances in neural information processing systems*, 26.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40: e253.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2009. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 29–30.

Malloy, T.; and Gonzalez, C. 2023. Learning to Defend by Attacking (and Vice-Versa): Transfer of Learning in Cyber-Security Games. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE.

Malloy, T.; Klinger, T.; Liu, M.; Riemer, M.; Tesauro, G.; and Sims, C. 2022. Learning in Factored Domains with Information-Constrained Visual Representations. In *Annual Conference on Neural Information Processing Systems Workshops*.

Niv, Y. 2019. Learning task-state representations. *Nature neuroscience*, 22(10): 1544–1553.

Niv, Y.; Daniel, R.; Geana, A.; Gershman, S. J.; Leong, Y. C.; Radulescu, A.; and Wilson, R. C. 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21): 8145–8157.

Piantadosi, S. T.; Tenenbaum, J. B.; and Goodman, N. D. 2016. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4): 392.

Pratt, L. Y.; Mostow, J.; Kamm, C. A.; Kamm, A. A.; et al. 1991. Direct Transfer of Learned Information Among Neural Networks. In *Aaai*, volume 91, 584–589.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2019. Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference. arXiv:1810.11910.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5): 527–535.

Sallans, B.; and Hinton, G. E. 2004. Reinforcement learning with factored states and actions. *The Journal of Machine Learning Research*, 5: 1063–1088.

Simon, H. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Taylor, M. E.; and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7).

Underwood, B. J. 1957. Interference and forgetting. *Psychological review*, 64(1): 49.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3): 261–272.

Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; and He, L. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Zanzotto, F. M. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64: 243–252.

Zhang, J.; and Bareinboim, E. 2017. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 1778–1780.

Zhang, L.; and Gao, X. 2022. Transfer adaptation learning: A decade survey. *IEEE Transactions on Neural Networks and Learning Systems*.