Reinforcement Learning Models Capture Human Decision-Making Processes

Post by Shireen Parimoo

What's the science?

How do people flexibly plan their actions in service of novel goals? According to reinforcement learning (RL) models of human behaviour, actions are chosen to maximize reward in the long run. **In standard RL algorithms, actions are guided by knowledge of the environment, where the outcomes achieved are either known (model-based) or learned when they occur without much prior knowledge (model-free).** These algorithms can require a lot of resources and also run the risk of under- or over-generalizing to new tasks.

Two new algorithms have been proposed to model human cross-task generalization. **One approach extracts similarities across tasks to inform future actions using universal value function approximators (UVFAs).** Let's consider this example: you are both tired and hungry, and want to make a decision between going to a Burger Shop (known for food), a coffee shop (known for coffee) or a diner (known for both). You know the Burger Shop is good when hungry and the coffee shop is good when tired, so you **select an action using UVFA**: you look for a place similar to both these places when you are both hungry and tired (the diner). In other words, you use previous values to extrapolate and predict a new value. **Another approach is to keep track of actions associated with commonly encountered tasks using a generalized policy improvement algorithm (GPI): this predicts the outcome of an action from learned experience. Here's an example of selecting an action using GPI**: You are hungry and looking for a place to eat. You have been frequenting the diner lately and you have a 'policy' of going there when hungry. Now, you're tired and would like to get coffee. You can also envision the outcome of going to the diner - there is coffee available, and people tend to drink coffee there. Therefore, you might choose to apply this same policy of going to the diner, in your decision to get coffee. **In other**

words, you are generalizing the policy to a new task. This is the distinction between UVFA and GPI: UVFA uses previously learned values to approximate a solution, while GPI evaluates a previously learned policy (the relationship between an action and an end state or solution), and applies this policy to a new situation.

This week in Nature Human Behavior, Tomov and colleagues tested the generalizability of standard and new RL algorithms across tasks and compared their performance to human behavior.

How did they do it?

In a set of four online experiments, **over 1100 participants played a resource-trading game set in a castle.** Before each trial began, participants saw the "daily market price" for the resources – the amount of money they could expect to either receive or pay for each resource (wood, stone, and iron). For example, they might see "Wood: $1, Stone, $2, and Iron, $0", indicating that they would receive $1 for each wood and pay $2 for each stone they had at the end of the trial. Each trial consisted **of a two-step decision-making process**: Participants were instructed to choose a) between three doors to enter one of three rooms, and then b) choose between another three doors per room to enter a final room containing resources. **Importantly, each final door always led to the same amount of resources in the corresponding final room (e.g., door 2 in room 3 always contained 100 wood, 40 stone, and 0 iron across all trials).** The amount of money received or paid, however, would change from trial to trial because the 'daily market value' changed each trial. In our example, participants might receive $1 x 100 wood ($100), -$2 x 40 stone (-$80), and $0 x 0 iron = $20 total.

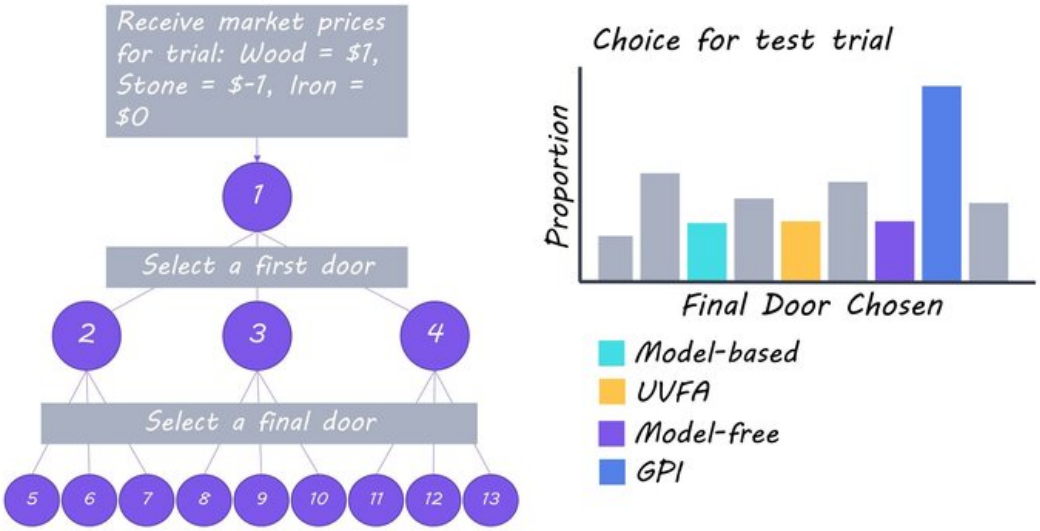
The researchers carefully selected a few sets of daily market prices for the resources for each experiment in order to manipulate the computational demands required and to vary the degree of difficulty in successfully mapping actions to outcomes across the four experiments. In the training phase of each of the four experiments, participants completed 100 trials, each of which was randomly assigned one of the pre-selected sets of daily market prices. For example, the profit from finding wood might double earnings on one trial but cost participants money on the next trial. Thus, **participants had to learn which final door would ensure maximum profit.**

The goal of the experiments was to determine which RL model the participants likely had used (model-free, model-based, UVFA, or GPI). The experiments were designed in such a way that the door participants chose on a test trial,

completed after the 100 training trials, would be likely to reflect the underlying RL algorithm that best modelled how they chose actions to maximize reward. For example, the third door in the second room in one experiment would in fact result in the highest profit in the final test trial, but it would only be selected by a model-based learner who had successfully learned the entire structure of the environment.

What did they find?

In the first experiment, participants were most likely to select actions leading to the final doors predicted by the model-based and the GPI algorithms. **In more difficult experiments, however, participants were by far more likely to choose the final door predicted by GPI.** Interestingly, the final door chosen by the model-based and UVFA algorithms would have been the most rewarding, yet participants did not choose those actions more frequently than would be expected by chance. In comparing the different algorithms, the authors found that **participants learned to select the final door predicted by GPI faster than that predicted by the model-based algorithm, which is consistent with the fact that model-based algorithms tend to require more resources.** Finally, as GPI makes predictions based on learned experience, the authors compared participants' choice history during the training phase to their actions on the test trial. Here, **learned experience did indeed drive choice at test time; participants' tendency to choose the same door during training predicted the probability that they would select that door in the test trial.** This indicates that participants kept track of the different situations they encountered during the training phase, along with the associated action-state mapping, which informed their behavior during the test.



What's the impact?

People use their knowledge of frequently encountered experiences to make predictions about future outcomes and inform their decisions. One of the interesting outcomes of this study is **that people do not necessarily make the most rewarding decisions, but rather they tend to map previously used policies onto new scenarios**. This finding provides exciting new insight into how reinforcement learning captures human decision-making processes in complex and changing environments.

Tomov et al. Multi-task reinforcement learning in humans. Nature Human Behavior (2021). Access the original scientific publication here.

Share

Newer Older

For any questions or interest in advertising contact founders@brainpost.co

Copyright © 2017, BrainPost. All Rights Reserved.
Reproduction of materials found on this site, in any form, without explicit permission is prohibited.