On the instrumental value of hypothetical and counterfactual thought

Thomas Icard (icard@stanford.edu)
Stanford University

Fiery Cushman (cushman@fas.harvard.edu)
Harvard University

Joshua Knobe (joshua.knobe@yale.edu)
Yale University

Abstract

People often engage in "offline simulation", considering what would happen if they performed certain actions in the future, or had performed different actions in the past. Prior research shows that these simulations are biased towards actions a person considers to be good—i.e., likely to pay off. We ask whether, and why, this bias might be adaptive. Through computational experiments we compare five agents who differ only in the way they engage in offline simulation, across a variety of different environment types. Broadly speaking, our experiments reveal that simulating actions one already regards as good does in fact confer an advantage in downstream decision making, although this general pattern interacts with features of the environment in important ways. We contrast this bias with alternatives such as simulating actions whose outcomes are instead uncertain.

Introduction

People spend a remarkable amount of time asking "what if?"—considering things they could have done but didn't ("counterfactuals"), or might do but haven't yet ("hypotheticals"). Specifically, they tend to simulate options that they regard as *good* (Kahneman and Miller, 1986; McCloy and Byrne, 2000; Phillips and Cushman, 2017; Icard et al., 2017). Suppose we observe a person saying something insulting and thereby getting into a terrible argument. We might immediately find ourselves thinking: "What if he had instead said something more tactful? What would have happened then?" But we would not normally show the opposite tendency. If we observe an individual saying something completely reasonable and tactful, we would not spontaneously begin thinking: "What if he had instead said something insulting?"

Here, we ask whether this tendency is an adaptive one. To answer this question, we conduct a series of computational experiments. These compare the performance of agents who show a tendency to consider options they regard as good with agents of several alternative designs. Is the "good action bias" advantageous, and when?

At the core of our approach is the idea that hypothetical simulation might prove helpful in subsequent decision-making. In our computational simulations, each agent faces a decision among a range of options. At the time of the decision, the agent is not able to reflect on all of these options, and it therefore has to make the decision in a way that involves only limited reflection. However, in the time prior to the actual decision, the agent can engage in off-line simulation of some options. The more the agent simulates a specific option, the more accurate its representation of this option becomes. The key difference between the various agents is that each one uses a different method to determine which options to simulate. We can then ask which way of selecting options at the time of off-line simulation leads to the best actions at the time of actual decision-making.

How to Improve Action by Thinking Ahead If the function of simulation is to improve future action then, broadly speaking, it must work by correcting errors in people's current assumption about the values of various actions. Such errors could take two forms: The value could be set too high (in which case the person would choose the action too often) or the value could be set too low (in which case the person might overlook the action too often). Moreover, if an error of either type exists and can be corrected merely through simulation, then the error existed because the individual had not yet devoted enough attention to simulating the action. This analysis suggests three broad heuristic strategies to allocating limited cognitive resources to simulation:

1. **Focus on actions that you haven't considered.** If you have already considered an action a large number of times, you are unlikely to learn a lot more by considering that same action again. Focus instead on considering those actions that you have thus far considered the least.

2. **Focus on actions you currently think are low value.** If you now regard certain actions as having low value, focus on simulating them and thereby learning more about them.

3. **Focus on actions you currently think are high value.** If you now regard certain actions as having high value, focus on simulating them and thereby learning more about them.

Of these three strategies, the first does the best job of maximizing the overall accuracy of the agent's representations of the expected value of each action. Thus, if the agent follows one of the other two strategies, the overall accuracy of its representations will be lower than it would have been if it had simply followed the first strategy. But, of course, the aim is not necessarily to have maximally accurate representations, but rather to have representations that are optimal for guiding action. It is therefore possible that one of the other two strategies will be more adaptive in the relevant sense.

The second strategy will tend to generate especially accurate representations of the lowest-value actions. Thus, to return to our original example, suppose that there are a number of different ways of making insulting comments (insulting the person's weight, insulting the person's family, etc.). This strategy allows the agent to develop a highly accurate representation of which of these actions would be the absolute

worst, which would be somewhat less bad, and so forth. By contrast, the third strategy will tend to leave the agent with relatively inaccurate representations of the worst actions but highly accurate representations of the best ones. An agent who followed this strategy would tend to be inaccurate about which specific insulting comment was the absolute worst. However, it would tend to be highly accurate about which specific tactful comment was the absolute best, which was only second best, and so on.

There is reason to expect that this third strategy might be especially adaptive. After all, the task at the time of decision-making is to pick out the best of the available actions. Given this task, it is far more important to be able to accurately distinguish the best from the second-best than it is to accurately distinguish the worst from the second-worst.

Within existing research in computer science, reinforcement learning, and planning, there is a sizable literature on questions related to ours. Some of this research goes under the heading of "pure exploration" (e.g., Bubeck et al. 2011), where an agent may test the effects of various actions without facing any concrete consequences. In this context, and in other contexts related to planning, one popular approach blends the first and third possibilities sketched above. That is, simulation is devoted to actions (or sequences of actions, e.g., chess plays) that are assumed to be good, but also relatively unexplored (see Browne et al. 2012 for an overview).

Below, we explore the performance of agents showing (1) a bias towards good actions, or (2) a bias towards bad actions, (3) a bias towards unexplored actions, (4) a blend of good and unexplored actions, following the current state of the art in computer science, as well as (5) a baseline of random action selection. Options 1, 3, 4 and 5 give us something approximating a factorial design, crossing the factors of whether the agent shows a tendency to focus on actions regarded as good and whether the agent shows a tendency to focus on actions that have not been considered. This design thereby gives us a number of different opportunities to explore the impact of a tendency to focus on actions regarded as good. We can ask whether it is better to focus on actions regarded as good than simply to select actions at random (comparing 1 to 5). We can ask whether it is better to focus on actions regarded as good than to focus on actions that have not yet been considered (comparing 1 to 3). Finally, we can ask whether it is better to adopt a blend of focusing on actions that have not yet been considered and focusing on actions regarded as good than it is to focus only on actions that have not yet been considered (comparing 4 to 3). We now elaborate on the details of our setting and each of the five algorithms.

Computational Experiments

Consider a simple scenario in which some agent will be facing some future decision problem, but before this has the opportunity to perform a number of hypothetical simulations, punctuated by occasional concrete actions and observation of their consequences. These concrete actions we assume are chosen by the agent itself in such a way as to give the highest chance of gaining a reward, choosing the action with highest estimated reward, $v(A)$, based on observations made so far. But the hypothetical simulations can be determined in any number of ways. We study five agent types representing different simulation strategies. Three of these use $v(A)$ itself.

1. **Softmax**: Stochastically chooses an action with probability proportional to estimated success probability—that is, choose A with probability $\propto \exp(v(A)/\tau)$. Throughout we set the "temperature" parameter τ to 0.1.

2. **Softmin**: Stochastically chooses an action with probability proportional to estimated *failure* probability—just like softmax, but $1 - v(A)$ is used in place of $v(A)$.

3. **Infomax**: Deterministically chooses the action that will maximize expected information gain. If action A has been observed to succeed s times and fail f times, the expected information gain for A is $\left(\frac{s}{s+f}\left(H(\frac{s}{s+f}) - H(\frac{s+1}{s+1+f})\right)\right) + \left(\frac{f}{s+f}\left(H(\frac{s}{s+f}) - H(\frac{s}{s+f+1})\right)\right)$, where $H(p)$ is entropy of p.

4. **Upper Confidence Bound (UCB)**: Deterministically chooses an action that balances estimated goodness with expected information gain by maximizing $v(A) + c\sqrt{\frac{2\ln L}{L_A}}$, where L is the total number of observations, and L_A is the number of times A has been observed (Browne et al., 2012).

5. **Random**: Chooses an action uniformly at random.

When making a concrete decision, all agents softmax-select an action using the currently estimated success probabilities.

There are three key parameters in our scenario that are worth highlighting. The first is the number N of possible actions that the agent might consider. The second is the distribution of rewards among these actions. For instance, do we expect there to be many good actions or only relatively few? The third parameter of interest is the number R of actions that the agent is able to retrieve for deliberation at decision time. We assume that online deliberation reveals the true expected value of an action to an agent, and thus the practical constraint on optimal decision-making is that the agent cannot deliberate about all actions—i.e., $R < N$. Thus, as R approaches N, the importance of prior simulations decreases.

Before going into the details of our experiments, here is a summary of the most important general patterns we observe:

1. When comparing two agents that differ only in whether they incorporate a bias toward the good actions (Random vs. Softmax, or Infomax vs. UCB), the agent that focuses on good actions almost always performs better.

2. Even an agent biased toward good actions who ignores uncertainty altogether (Softmax) generally outperforms an agent who minimizes uncertainty without focusing on good actions (Infomax). While this result appears to hold "on average," there are scenarios where we see the opposite.

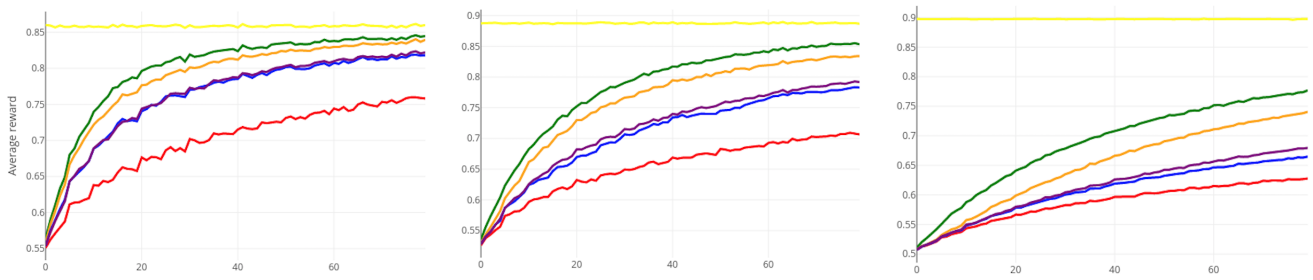

Figure 1: Performance of different agent types over 80 learning trials (x-axis), with $N = 10$, 25, and 100 actions. The color key is as follows: yellow=Omniscient, orange=Softmax, red=Softmin, blue=Infomax, green=UCB, purple=Random.

3. Whereas good-seeking and information-seeking tendencies are both generally helpful, the helpfulness of a good-seeking bias (unlike that of the information-seeking bias) seems to depend on the idea that the number (R) of actions we can retrieve at decision time is relatively small.

These results thus demonstrate the sense in which the empirical tendency we observe may indeed be adaptive. The results also raise a number of subtle empirical and theoretical questions, which we will discuss further below.

Initial Experiment: Actions Uniformly Distributed

Let us begin by assuming that at decision time each agent will merely (softmax) retrieve a single action (i.e., $R = 1$). Intuitively, these are scenarios that allow no time for deliberation, and decision making thus depends entirely on the agent's view of the various actions as established through prior simulation and observation. Suppose each of the N actions has some "success" probability, and as a very simple first assumption, suppose that these success probabilities are drawn uniformly from the unit interval $(0,1)$.

Fig. 1 shows the results. A learning trial involves observing of an action outcome (success/fail), where this action is chosen according to one of the strategies above, or (every fifth trial) by softmax selecting an action using current value estimates. The latter is intended to simulate observations of actual (not just hypothetical or counterfactual) choices. At the end of L trials (x-axis) we assess each of the five agents according to how well they perform by using their currently estimated action values. (Success probabilities are drawn independently 5,000 times for each value of L, ensuring no correlation.) We include results for an "omniscient" agent ($R = N$), revealing the average maximal success probabilities.

The most striking pattern here is that a bias toward good actions is helpful no matter whether there are 10, 25, or 100 actions. That is (as a first illustration of point 1 above), the Softmax Agent, who biases simulation toward good actions, is significantly better than the Random agent; and the UCB agent, who mixes uncertainty minimization with a bias toward good actions, drastically outperforms the pure Infomax agent. Moreover (in line with point 2 above), the Softmax agent has a significant advantage over the Infomax agent, whose performance is even slightly worse than Random.

Varying the Distribution on Success Probabilities

An obvious question is whether these results depend on the distribution of success probabilities being uniform. What if the distribution were instead highly skewed toward good or bad actions? Or what if most actions were intermediate with only a few very good or very bad actions?

To investigate this question it is convenient to think of success distributions as themselves drawn from a *beta prior*, $\mathsf{Beta}(\alpha,\beta)$, giving a distribution over Bernoulli success probabilities. The two parameters α and β have the following significance: $\frac{\alpha}{\alpha+\beta}$ gives the mean of the distribution, while $\alpha+\beta$ determines the shape of the distribution. For instance, if one of α or β is much higher than the other, this will result in a highly skewed distribution. The uniform distribution considered above is the simple case of $\mathsf{Beta}(1,1)$.

Two other familiar distributions on Bernoulli probabilities are the so called *Jeffreys* distribution (which is equivalent to $\mathsf{Beta}(0.5,0.5)$) and a standard bell-shaped distribution (we take $\mathsf{Beta}(2,2)$ as a representative example). These are both symmetric distributions, meaning that there are typically as many good actions as bad actions. However, they differ from the uniform distribution in that success probabilities are more concentrated either around 0.5 (bell-shaped) or closer to the extremes, 0 and 1 (Jeffreys). The results for these two environments, with $N = 25$ actions, are not appreciably different from the uniform case, as shown in Fig. 2

A somewhat different pattern is revealed in environments with highly skewed success probabilities. Again in Fig. 2 we show the results for two distributions that are skewed toward either very low or very high success probabilities (specifically these are $\mathsf{Beta}(.05,.5)$ and $\mathsf{Beta}(.5,.05)$, respectively). In these scenarios the advantage of good-seeking is either modest or not present at all. What might explain this pattern?

Consider the extremely positive skewed case (that is, $\mathsf{Beta}(.05,.5)$). In this setting, the initial expectation for every action is quite low ($< .1$), but there is a small set of actions that outperform this expectation significantly (typically at least one > 0.8). When it stumbles upon just one of these outliers in simulation, a Softmax agent will perseverate on it, to the exclusion of discovering one of the few others (at least one of which is likely to be quite superior). In contrast, the Infomax agent will gradually survey the full set of

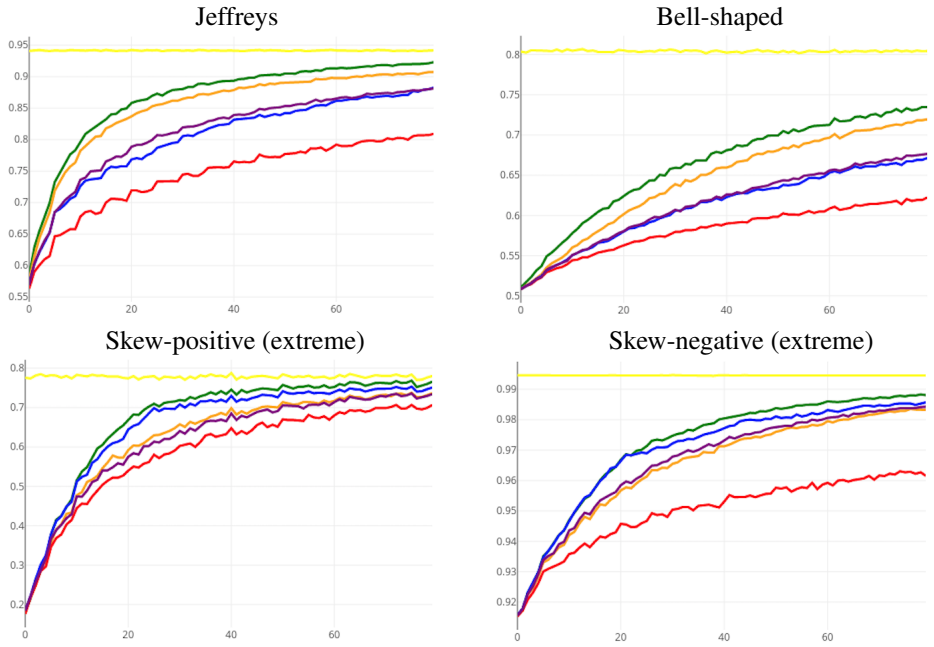

Figure 2: Performance graphs for other distributions on action success probabilities. All simulations are with $N = 25$ actions.

options and eventually discover the true optimal action. In other words, the positively skewed case presents the Softmax agent with "distractors" that capture simulation because they easily outperform the default expectation, while significantly underperforming the optimal action.

These observations suggest at least two further questions. First, how "typical" are these environments? Second, does the pattern depend at all on the number (N) of actions?

Turning briefly to this second question, suppose that the distribution on actions is the same, but that there are many more possible actions, say 100. Intuitively, the relative advantage of Infomax should be reduced in this setting as the number of candidate actions increases. This is because the Infomax approach of exhaustive, balanced search becomes especially inefficient as the space of actions grows. Fig. 3 shows that this intuition is indeed borne out with 100 actions.

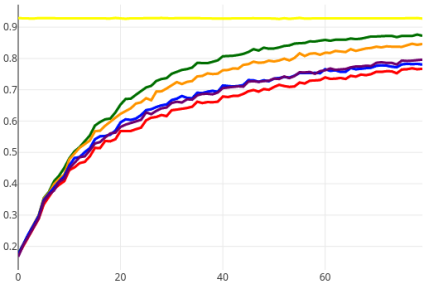

Figure 3: $N = 100$ actions, Skew-positive (extreme).

In fact, with this scenario we see no apparent advantage of Infomax even over the Random agent (similar to what we saw in Fig. 1). This suggests that the advantage of information-seeking in this scenario may not be particularly robust. This naturally leads us to the first question: what happens "on average" as we consider the entire parameter space?

As an illustration, let us return to the setting of only 25 actions. The example distributions so far have been cherry-picked based on specifically notable characteristics (flat, bell-shaped, skewed). But what happens when we average over the environment parameters, α and β? To investigate this question we present the results of a large-scale experiment in which parameters are drawn from a hyper-prior. A relatively neutral hyper-prior, used often in cognitive science (see, e.g., Griffiths et al. 2008), defines reasonably "unbiased" distributions on the mean ($\frac{\alpha}{\alpha+\beta}$) and the shape ($\alpha+\beta$), which together uniquely determine α and β. Specifically, we draw the mean from the uniform distribution on $(0,1)$, while the shape is drawn from an exponential distribution with rate parameter 1 (meaning that the density function is simply e^{-x}). The results in this broader setting are shown in Fig. 4.

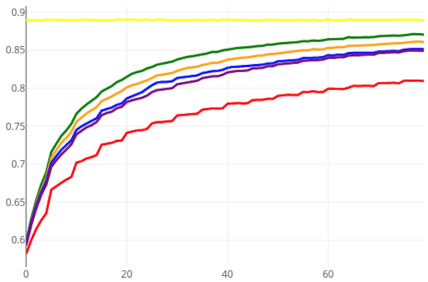

Figure 4: $N = 25$, averaged over environmental settings.

This reveals that, even in the scenario of 25 possible actions where we know the Infomax agent occasionally outperforms the Softmax agent, this makes up a relatively low-probability portion of the space. On average, Infomax performs only slightly better than a Random agent. But most importantly (and again, in line with point 1 above), both good-seeking agents (Softmax and UCB) have a clear advantage over the agents without this bias (Random and Infomax).

Adding Reflective Deliberation

Up to this point we have been assuming that each agent has the capacity to consider only one possible action at decision time. This means the quality of decision is dependent entirely on how well previous observations and simulations have biased better actions to come to mind. We now turn to scenarios in which an agent may have the capacity to deliberate over several possible actions and then choose whichever of these looks best upon reflection. In this kind of scenario, the aim of prior hypothetical simulation is somewhat broader. A strategy is effective to the extent that the set of R actions brought to mind in a decision context will likely include at least one very good action (for this particular context).

One straightforward way to formalize this type of scenario is to associate each action, not with a specific Bernoulli success probability, but with a distribution over such probabilities. The intended interpretation is that actions may be good or bad overall, but that the agent has the capability to figure out the actual success probability in a given circumstance, which might be quite different from what one would expect on average. Thus, in these simulations we again test how well different agents perform after learning, but decision making is assumed to work in a more sophisticated way. Instead of sampling an action in proportion to its expected goodness, each agent (softmax-)selects some number R of actions and then deterministically chooses whichever of these R actions turns out to be best in the given situation, reaping that reward. Intuitively, as R increases the learning problem becomes significantly easier, since it is easier to find oneself with good options to consider at decision time.

The setting we are now studying is similar to the scenario studied earlier (and presented in Fig 4), except that in the present case each action is associated with its own beta distribution. Specifically, the distribution for each action is drawn from a hyper-prior with $\alpha+\beta \sim \mathsf{Exp}(1)$. Suppose first that the means $\frac{\alpha}{\alpha+\beta}$ are drawn uniformly. In Fig. 5 we report on the case of $N = 25$ actions with only 25 learning trials. We show the results with $R = 1,\ldots,10$.

As is evident from the figure, the differences among algorithms is relatively pronounced at $R = 1$ and is still noticeable at $R = 4$. That is to say, even if an agent has the capability to retrieve and reflectively deliberate over 4 actions at decision time, simulating better actions may still have an appreciably positive effect on an agent's success. However, closer to $R = 10$ (just less than half the total) all differences vanish.

Once again, we can ask the question of whether this pattern depends on the specific assumption that the (average) success

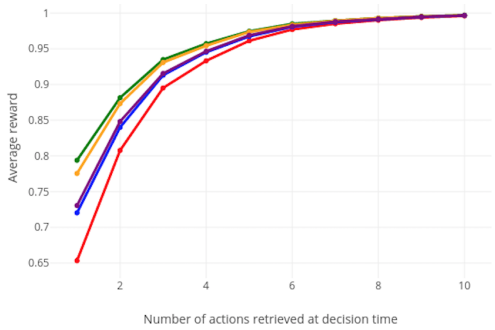

Figure 5: Uniform distribution on success probability means. R actions retrieved at decision time.

probabilities are uniformly distributed, i.e., that we expect the same number of actions for any particular range of success probabilities. As one would expect, when the distribution of means is at all favorable toward (generally) good actions, the differences among algorithms disappears even more rapidly: it is simply too easy to find at least one rewarding action.

Other symmetric distributions with mean 0.5 show the same pattern as in the uniform case. In Fig. 6 are two examples, where the means themselves are drawn from Jeffreys and bell-shaped priors (again, $\mathsf{Beta}(.5,.5)$ and $\mathsf{Beta}(2,2)$).

Also presented in Fig. 6 is the scenario where means are drawn from a highly (positively) skewed prior.[1] In such a scenario almost all of the actions are almost always very bad, but there are a few that are typically very good. As discussed above, the aim of simulation is figure out which actions should be included among the R to retrieve at decision time. If there is only one "needle in the haystack," then an agent will perform better the more likely it is to identify that uniquely good option.

Notice that the x-axis on this third graph in Fig. 6 goes all the way to $R = 20$. Remarkably, the advantage of info-seeking is still apparent even when the agent can retrieve and deliberate over 20 of the 25 possible actions. Also remarkable is the observation that, while the Softmax agent clearly outperforms the Random agent, the UCB agent's attention to goodness does not effect any significant gains over Infomax. A reasonable conclusion from this study is that, in such a scenario, efficient and exhaustive pure search is hard to beat.

This last case study uncovers an important caveat to the general finding that a bias toward good actions in hypothetical thinking is adaptive. In contrast to biases toward uncertainty-minimization, the advantage of the good-seeking bias depends on the assumption that the number (R) of actions that a person can consider in deliberation is small relative to the number (N) of possible actions that one could conceivably consider. A worthy hypothesis is that this is exactly the kind of situation people typically face.

[1] We do not show the graph for the negatively skewed prior. In the present setting such a distribution results in too many good actions, and there are no observable distinctions among agents.

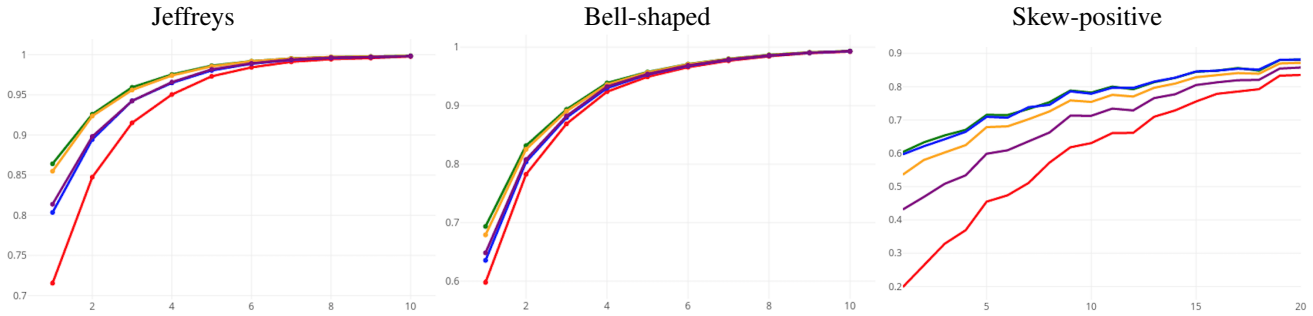

Figure 6: Distribution of means drawn from Jeffreys, bell-shaped, and positively skewed priors.

Discussion

We demonstrate three basic patterns in the way that a "good action bias" during offline simulation later improves online decision-making. First, we find that a good action bias improves performance whether or not there is an additional bias to reduce uncertainty. Second, although an uncertainty-reduction bias typically improves performance, its effect is usually smaller than the good action bias. Though we occasionally see the opposite pattern—particularly in environments with just one or two very good actions and many quite poor actions ("needle in a haystack" problems)—the trend is robust when averaging over environmental parameter settings. Third, the benefit of a good action bias depends strongly on the assumption that the agent is unable to retrieve and deliberate over a large number of actions at decision time. By contrast, the benefit that minimizing uncertainty confers (in those cases where the benefit is especially apparent) does not seem to depend on this assumption.

These results strongly confirm our hypothesis that the bias people empirically show toward thinking about actions they deem good is adaptive. But the results also raise several new theoretical and empirical questions and possibilities.

While our focus has been on the bias people show toward good actions, our results also confirm a hypothesis that hypothetical and counterfactual simulation aimed at minimizing uncertainty would be quite helpful. We saw several instances in which Infomax outperformed Softmax. But more strikingly, the UCB agent, who employs a blend of the two approaches, seeking out good as well as informative actions, outperforms every other agent in virtually every context. Indeed, this is one of the reasons UCB-type agents have been so heavily studied in computer science, including in the setting of offline planning. Moreover, it is easy to imagine that in many domains that have this "needle in a haystack" character (in areas of science, for example), thinking more about less explored possibilities could be especially rewarding.

Empirical research has clearly demonstrated the first type of bias, toward good actions. An obvious question is whether people also show a bias toward simulating more informative actions. That is, in a case where actions A and B appear equally good, but where the person simply has less information or experience revealing how A might turn out, would she then show a tendency to think about A more than about B, for example, when imagining counterfactual scenarios? We leave this as an intriguing open question.

Finally, the third result mentioned above—that the bias toward good actions is only effective when the number R of actions retrieved is relatively small—may point toward an important fact about the kinds of problems that cognition is adapted to solve. It seems evident that in any given decision making context, people can only bring to mind and deliberative over a very small number of options—certainly quite small relative to the number of conceivable actions. As our agent experiments show, these are precisely the cases where such a bias is advantageous. Perhaps this pervasive feature of our cognitive predicament explains a great deal about the nature of hypothetical and counterfactual thought.

References

Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1).

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412:1832–1852.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.

Icard, T., Kominsky, J., and Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161:80–93.

Kahneman, D. and Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psych. Rev.*, 94:136–153.

McCloy, R. and Byrne, R. (2000). Counterfactual thinking and controllable events. *Memory and Cognition*, 28:1071–1078.

Phillips, J. and Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114(18):4649–4654.