Decisions to intervene on causal systems are adaptively selected

Anna Coenen, Bob Rehder, and Todd Gureckis
Department of Psychology, NYU, 6 Washington Place, New York, NY 10003 USA
{anna.coenen, bob.rehder, todd.gureckis}@nyu.edu

Abstract

How do people choose interventions to learn about a causal system? Here, we tested two possibilities: an optimal information sampling strategy which aims to *discriminate* between multiple hypotheses, and a second strategy that aims to *confirm* individual hypotheses. We show in Experiment 1 that individual behavior is best fit using a mixture of these two options. In a second experiment, we find that people are able to adaptively alter the strategies they use in response to their expected payoff in a particular task environment.

Keywords: causal learning; information sampling; interventions.

Introduction

Interventions are an important instrument for learning about causal structures. By manipulating causal variables we can better discover the relationships between them. This ability is crucial in many areas of human inquiry including empirical science, medical reasoning, or simply when learning how a new mechanism, like a smartphone, works. In this paper we ask how people decide *which* variables to manipulate when they want to test specific hypotheses about a causal system.

Previous work has most often sought a *single* strategy or model that describes how people search for information when learning. In particular, two competing perspectives have emerged. One set of models assumes that people select information to optimally *discriminate* between different hypotheses. Such rational sampling norms have been used to model information search in many different domains (for an overview see Nelson, 2005), including learning of causal structures. For example, Steyvers and colleagues (2003) argue that participants use an *Information Gain* strategy (IG) when choosing causal interventions. This strategy aims to minimize a learner's uncertainty about which out of a number of graph descriptions (hypotheses) underly a particular causal system.

On the other hand, research on hypothesis testing in other domains, particularly in rule-learning tasks, has often argued that people use *confirmatory* strategies to search for information (Nickerson, 1998). For example, they might use the *positive testing strategy* (PTS) which makes search queries that they expect to be true under one hypothesis, irrespective of whether it helps to discriminate between different hypotheses (Klayman & Ha, 1989; Wason, 1960). Although PTS can be optimal under certain circumstances (Navarro & Perfors, 2011; Oaksford & Chater, 1994), it often runs counter to optimal sampling norms such as IG.

This paper challenges the view that people use a single strategy to test causal hypotheses via interventions, and instead finds that people simultaneously use both discriminatory and confirmatory reasoning when making interventions. Furthermore, we show that this strategy mixture is not fixed but that people can change their strategies in response to the payoff structure in a given environment. On the basis of these findings we argue that people have a flexible and adaptive repertoire for causal structure learning, rather than relying on a single strategy. The structure of this paper is as follows. First, we will define two computational models of intervention selection. We then report an experiment aimed at distinguishing the models. Based on the results, we present a second study in which we manipulate the expected payoff from each strategy to investigate the impact on people's intervention decisions.

Two models of intervention-based causal learning

Information Gain The IG model predicts that learners should choose interventions that they expect to maximally reduce their current uncertainty, $H(G)$, about a set of causal hypotheses or graphs, G (Murphy, 2001; Tong & Koller, 2001). The expected Information Gain of an intervention a can be calculated as:

$$EIG(a) = H(G) - \sum_{y \in Y} P(y|a) H(G|a,y) \tag{1}$$

where $P(y|a)$ is the probability of outcome $y \in Y$, given action a. Calculating EIG requires knowing the new uncertainty after making intervention a and observing outcome y:

$$H(G|a,y) = \sum_{g \in G} P(g|a,y) log \frac{1}{P(g|a,y)} \tag{2}$$

where $P(g|a,y)$ is the probability of graph g given intervention a and resulting outcome y. To calculate $P(g|a,y)$, Bayes' rule can be applied, yielding $P(g|a,y) = P(y|g,a)P(g)/P(y|a)$. Finally, $P(y|a)$ can be computed by marginalizing over all possible graphs and their likelihood of producing outcome y given intervention a, $P(y|g,a)$.

Positive testing There is no existing definition of positive testing as a causal intervention strategy in the literature. PTS has mainly been articulated in rule learning tasks, where it constitutes a preference for search queries that lead to positive outcomes (i.e. "yes" rather than "no") under a given hypothesis (Klayman & Ha, 1989; Wason, 1960).

We propose that such a preference for positive outcomes might translate into a preference for creating positive *effects* in the causal learning scenario. Consequently, PTS could manifest as a preference for nodes that have high *causal centrality* in a hypothesis that is currently under evaluation (Sloman, Love, & Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, 2000), where centrality is measured by the number of di-

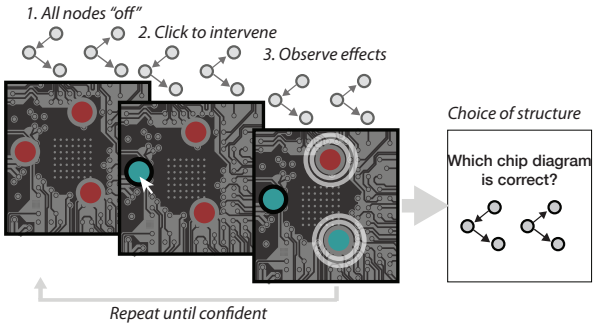

Figure 1 An example trial. A chip was presented with all components off. Two possible wiring diagrams were displayed above the chip. Participants selected a chip component to "activate" and observed the resulting effects on the system. Later participants were asked to choose which of the two diagrams best described the operation of the chip.

rect and indirect descendant causal links. If learners make interventions on high-centrality nodes, they can gather positive evidence for a hypothesis by producing the expected effects that are entailed by these descendant links (e.g. if a system has an on/off structure, then turning on a high centrality node should cause its direct and indirect children to turn on also). Because graphs can differ in the number of links that can be tested in principle, we will consider causal centrality relative to the total number links in a given graph. Thus, the PTS value of intervening on a node n is determined by that node's maximum relative causal centrality over all graphs that are currently under consideration:

$$PTS(a) = \max_{g} \left[\frac{DescendantLinks_{n,g}}{TotalLinks_{g}} \right] \quad (3)$$

where descendant links are all the links that lead to direct or indirect children of the node that is intervened on. To illustrate the concept of PTS, a node will have a value of 1.0 if, by intervening on it, all possible links of at least one hypothesized graph can be activated. If it can activate at most one out of two links, it receives a score of .5, and a score of zero if it cannot lead to any outcomes at all. According to this strategy, nodes become attractive if they have a high PTS score in *at least one* hypothesis that is currently evaluated, irrespective of the differences between hypotheses.

Choice model For both models, the probability of choosing one intervention a_i out of a range of possibilities, A, given a measure of the usefulness of the action, $V(a_i)$, is:

$$P(a_i) = \frac{exp(V(a_i)/\tau)}{\sum_j exp(V(a_j)/\tau)} \quad (4)$$

where $V(a)$ is determined by either Eqn 1 or 3. Parameter τ determines the degree to which behavior resembles guessing rather than choosing the action with the highest $V(a)$ score.

Experiment 1

Method

Participants One hundred and five participants were recruited via Amazon Mechanical Turk. Participants received $2 for participation with the option of earning up to $1 bonus based on performance in the task.

Stimuli and Materials On each trial of the experiment, participants were shown a simple causal system ("computer chip") and were asked to learn how it worked (see Figure 1). Each chip had three components (nodes), which could either be on or off as indicated by their color. Each chip could behave according to one of two possible causal structure hypotheses (wiring diagrams) that were visible at all times. One of the two wiring diagrams was randomly selected to be the true underlying structure. On each trial, participants interacted with the system to determine which diagram best described its operation.

All three-node causal Markov structures with one or two links were used in the experiment, yielding four basic structure types (Chain, Common Cause, Common Effect, and One-Link) that were exhaustively paired with each other to form 27 unique hypothesis pairs (see Figure 2, top). All links had causal strengths of 0.8 and there were no background causes that could activate a node spontaneously (i.e., without an intervention or active parent node). Participants were told and quizzed about these details before starting the task.

Procedure A trial began with all components of the chip switched off (red). The participant could then intervene on one component by clicking on it and thereby turning it on (green). After a short delay (500ms) an animated white ring appeared around all other components to indicate that they were updated as a consequence of the intervention. Components that were activated by the intervention changed their color to green while all other components remained red. All components had to be reset to their original state (off) using a button press before another intervention could be made. Participants made as many interventions as they desired. Afterward, participants indicated which wiring diagram they felt was correct by clicking on one of the two options. They then rated their confidence.

To ensure that participants chose their interventions carefully, they were offered a bonus of up to $1 from one randomly chosen structure comparison at the end of the experiment. The bonus was only paid if they chose the correct structure at the end of the selected trial, and it was further reduced by $0.10 for every intervention made in that trial.

Results

Overall, participants were highly accurate in identifying the correct wiring diagram after interacting with a chip. The percentage of correct choices averaged across individuals was 87% (SD = 0.14, MD = 92%). Participants' confidence ratings mirrored their choices, with higher confidence ratings on correct trials (M = 80.22), versus incorrect trials (M = 72.62),

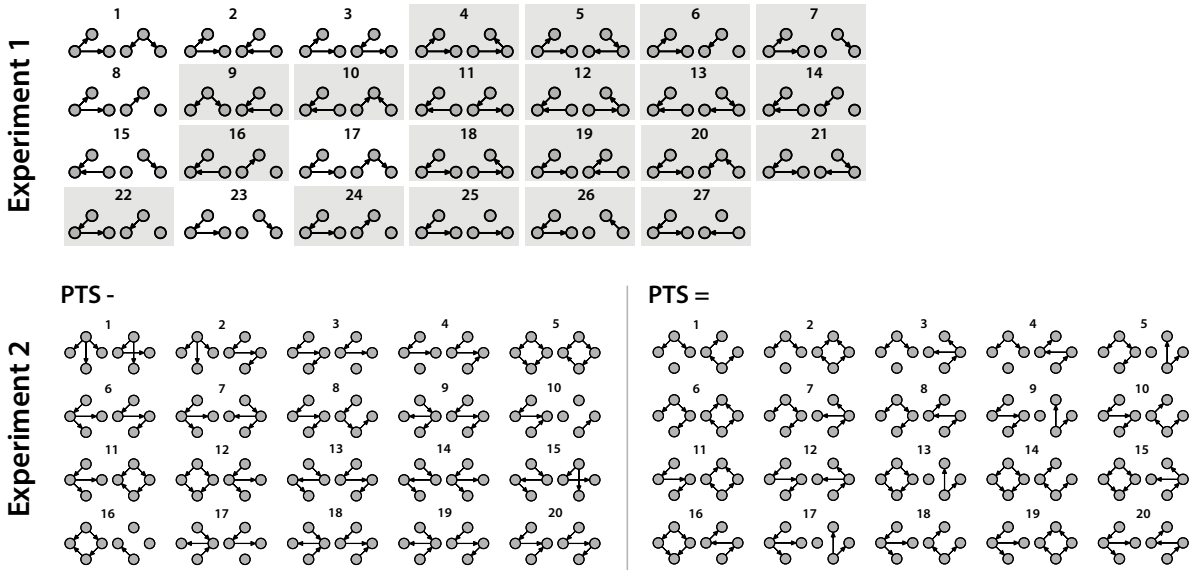

Figure 2 *Top*: All 27 structure comparisons used in Experiment 1. Each numbered comparison represents a trial in which participants were asked to intervene to decide between the depicted two causal hypotheses. Comparisons highlighted with a grey box were also used in the second phase of Experiment 2. *Bottom*: the problem types used in Experiment 2. The PTS- set specifically discourages the use of the PTS strategy (nodes highly valued under this strategy result in confounded evidence). The PTS= set is effectively neutral with respect to IG and PTS (using PTS is not harmful in this case).

$t(89) = 3.66$, $p<.001$.

Intervention decisions - Individual models Our critical question concerns how people decide which node(s) to intervene on given a specific pair of hypotheses (wiring diagrams)[1]. Since we are modeling only a single choice per participant per problem type, there are unavoidable amounts of measurement noise. For example, even a participant perfectly following the IG strategy might not choose the IG-maximizing choice on a single trial (e.g., assuming Eqn 4). To ensure that our model fit measures correctly accounted for this type of noise and uncertainty, we adopted a fully probabilistic model-assessment approach using a hierarchal Bayesian model (for an informal discussion see Coenen & Gureckis, 2013). The generative model assumed a population level distribution over the τ parameter (Eqn 4) such that for participant i, $\tau_i \sim Gamma(\alpha, \beta)$. This sampled value of τ_i was then used along with the $V(a)$ scores for the IG and PTS models (Eqns 1 and 3) to derive choice probabilities for each of the three nodes on the circuit board. Finally, a single choice was sampled from these probabilities via a categorical distribution.

We performed Bayesian inference, conditioned on the behavioral data, to obtain posterior estimates of these hyperparameters α and β, as well as τ_i for each participant. To then assess the quality of the model fit, we used the method of posterior predictive model assessment (Gelman, Meng, & Stern, 1996) and compared samples of each model to the data. This analysis (not shown graphically, in the interest of space, but summarized in Coenen & Gureckis, 2013 for the IG model) revealed that both models fit the data well on some of the 27 problems, but also missed the key behavioral profile for other problems. As a result neither the IG nor PTS model provided a credible account of the empirical data (i.e., for both models there was more than one problem for which behavior was well outside the 95% confidence contour even after accounting for measurement noise).

Intervention decisions - Combined model Given that neither the PTS or IG model provided a sufficiently credible fit to our behavioral data, we considered a range of alternative models. In the interest of space, we describe here the best alternative model from our exploration, shown in Bayesian Hierarchical form in Figure 3 (middle panel). This model represents a linear combination of IG and PTS with a mixture weight θ which determines the degree to which participants match IG compared to PTS. Eqn 3 in the same Figure shows how choice probabilities arise from this linear combination. The strategy reduces to a pure version of IG or PTS when $\theta = 1$ or $\theta = 0$, respectively. Both θ and τ were fit individually for each subject. At the hyperparameter level θ was fit using a beta distribution that was reparametarized by its mean μ and standard deviation κ. The distribution of τ was parameterized as described above.

Using the posterior-prediction method described above, we then also evaluated this combined model. Unlike for the in-

[1]On average, participants made 1.56 (SD = 0.59) interventions during a single chip test, and the majority of structure choices were made after only one intervention. Given this, we focused our analysis on the *first* intervention that participants made in any game.

dividual models, none of the empirical data from individual problems appear implausible in relation to this model (i.e., the behavioral data lie within the range of plausible data patterns generated from the model).

Individual differences in strategy use The inferred θ parameters in the combined model provide an estimate of participants' individual tendency to behave according to IG compared to PTS. The top plot of the left panel in Figure 3 shows a histogram of the best-fitting values of θ for each participant based on maximum-likelihood estimation[2]. Interestingly, rather than dividing into two groups, many participants fall on a continuum between the two strategies. Thus, behavior does not only resemble a strategy in the aggregate; it does so at the individual level, as well.

In support of the parameter estimates, we found that participants' intervention strategies, measured by θ, were related sensibly to other behavioral variables. For example, higher weightings of IG were negatively correlated with response time, $r(103) = .23$, $p < 0.05$, but had a positive impact on accuracy, $r(102) = .44$, $p < 0.001$ (after controlling for τ, which indicates participants' tendency to guess rather than decide in line with any of the two models). This matches the intuition that IG is computationally more expensive than PTS, but also more effective for learning the correct structure.

We also found a relationship between θ and measures of successful belief-updating in line with an optimal learner using Bayes' rule. For instance, θ was positively correlated with the proportion of times a participant chose the hypothesis with the higher posterior probability given their intervention outcomes, $r(102) = .4$, $p < 0.001$ (again, controlling for τ). This shows that differences in people's intervention strategies might also be connected with differences in their ability to learn from these interventions, and that using PTS corresponds to a higher tendency to deviate from optimal behavior.

We did not find a significant relationship between θ and τ, $r(103) = -0.07$, $p > 0.05$ (calculated using $log(\tau)$, since the estimates were strongly positively skewed).

Discussion

The first experiment offered two insights. First neither the IG nor PTS model alone seem to provide a plausible account of participant's intervention decisions. Instead, behavior was best explained by a mixture of these two strategies that varied somewhat between participants. Second, we developed a new model assessment approach based on hierarchical Bayesian modeling and posterior predictive simulation to assess model fits. This approach enabled us to evaluate both population and individual level difference in strategy use from relatively sparse data (a single choice per problem).

Somewhat counter to the conclusion of past work on modeling intervention choices (Steyvers et al., 2003), we did not find strong support for the IG model. This was surprising because many aspects of our experiment were selected to make it particularly easy for participants to use IG (e.g., small number of hypotheses explicitly visible at all times, economic incentive to be efficient). One crucial question raised by this finding is what factors determine what strategy people use. In particular, we wondered whether the tendency to use a confirmatory strategy like PTS is a stable "bias" in how people approach such problems or whether it can be altered through more, or different, experience with the task. This issue is explored in Experiment 2.

[2]Note, we used the maximum-likelihood estimates of θ instead of the fits from the Bayesian analysis in Figure 3 because the latter were influenced by the hyperparamer distributions and thus did not reflect the closest match to a participant's actual choice data. MLE values of θ and Bayesian fits were very highly correlated, however.

Experiment 2

One important property of Experiment 1 was that even if participants used PTS, the cost they incurred in accuracy compared to using IG was relatively small. The present experiment was designed to test if people can learn to switch from using PTS to a more discriminatory strategy when PTS more obviously impairs performance. The experiment closely followed the design of Experiment 1. However, participants first completed a set of novel causal intervention problems that were either designed to make PTS a lot less effective than IG (*PTS-* condition) or to make the strategies almost equally useful (*PTS=* condition). In both conditions, participants were then tested on a critical subset of problems from Experiment 1. If strategy use is flexible and adaptive to experience, we expected that interventions would be more in line with the IG model in the PTS- condition compared to *PTS=*. However, if strategy use is a stable trait or bias, we expect to find no difference between the conditions.

Method

Participants We recruited 122 participants via Amazon Mechanical Turk. Compensation and incentive structure were the same as in Experiment 1.

Stimuli, Materials, and Procedure In total, participants completed 40 intervention problems. In the first half of the experiment they were given 20 new problems consisting of pairs of four-node causal networks (see Figure 2, bottom). In the *PTS-* condition, each problem was designed such that choosing interventions using PTS would often lead to outcomes that do not differentiate between the hypotheses. A simulation of an optimal learner choosing interventions on these problems resulted in only 62% accuracy after one intervention using PTS compared to 91% using IG (assuming the learner always chooses the option with the highest IG/PTS score on each trial). In the PTS= condition, simulated accuracy of PTS after one intervention was 93%, compared to 95% with IG. To compare, in Experiment 1, PTS would have led to 85% and IG to 92% accurate choices, so the payoff structure more closely resembled the PTS= condition.

In addition to these new problems, participants were then tested on a selection of 20 of problem types from Experiment 1 (specifically the subset for which IG and PTS made differ-

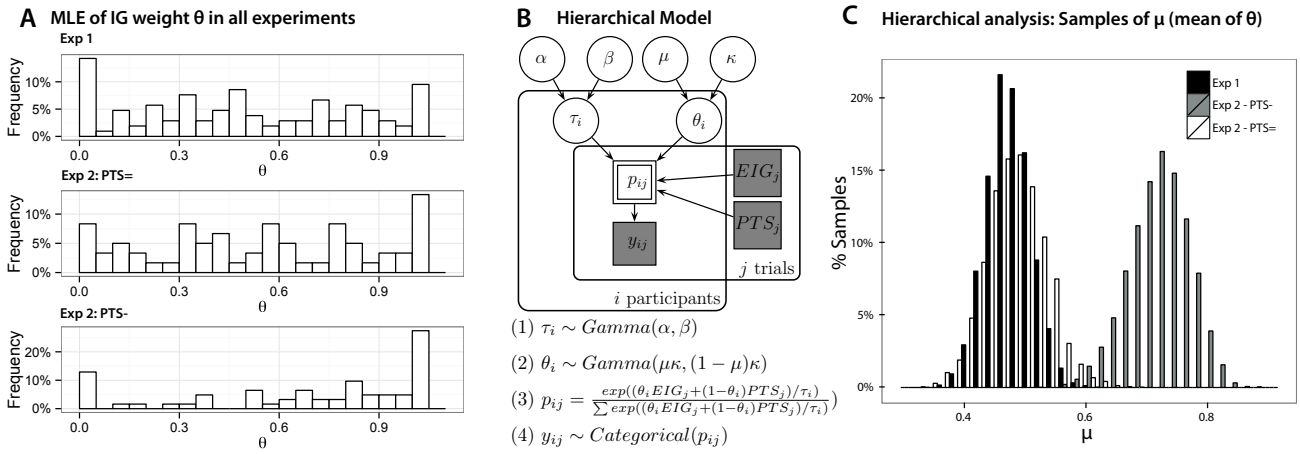

Figure 3 A: Histograms of best fitting θ parameters in all experiments. High θ indicates a better match of the data to IG compared to PTS. B: Hierarchical Bayesian model of the combination of IG and PTS. Each trial, *j*, corresponds to one problem type for which each participant chose one intervention, *y*. IG_j And PTS_j are three-vectors with model scores for the three possible intervention on problem *j*. *x* is a three-vector of choice probabilities for each intervention. C: Distribution of samples of the μ parameter (population mean of θ), fit to data in Experiment 1 and both conditions of Experiment 2.

ent predictions, assessed by their rank correlation). In Figure 2 (top) these problems are highlighted with a grey box.

The overall procedure was the same as in Experiment 1.

Results

As before, participants were very accurate at choosing the correct hypothesis at the end of a trial. In both conditions they chose the correct graph on average 88% of the time ($SD = 0.13$ and $SD = 0.12$ in PTS- and PTS=, respectively).

Our main interest in this experiment was to see which interventions participants chose on the 20 three-node problem types that were already used in Experiment 1. The distribution of best fitting θ parameters is shown in the leftmost column of Figure 3 (bottom two plots). In the PTS- condition estimates of θ are shifted considerably towards 1.0 (i.e., pure IG strategy), compared to the PTS= condition.

To assess whether, at the population level, this new distribution of strategy weights was credibly different between the two conditions, we also fit the full Bayesian model described above (see middle panel in Figure 3) to participants' choices. The right panel in Figure 3 shows the distributions of samples of the μ parameter from this Bayesian model for the two new conditions and for Experiment 1. This parameter represents the population mean of θ, that is, the overall tendency of all participants to choose interventions in line with IG, compared to PTS. As the figure shows, μ is shifted considerably towards higher IG-use in the PTS- condition of Experiment 2, compared to PTS= and Experiment 1.

Another way of testing whether this change in behavior between the two conditions is credible involves determining the 95% Highest Density Interval (HDI) of the distribution of the *difference in* μ in the PTS- and PTS= conditions. To compute this difference, we took 10, 000 samples from each model, paired the samples randomly, and computed $\mu_{PTS=} - \mu_{PTS-}$ (method is similar to Kruschke, 2013). The 95% HDI of this distribution did not include 0.0. As a result, we can be confident that there is a credible difference at the population-level in the degree to which participants used IG in the two experimental conditions, even when testing them on exactly the same problem set.

Discussion

We found that participants were more prone to behave in a discriminatory (i.e., IG) fashion, after encountering a sequence of problems for which PTS led to a lower expected payoff (PTS-) than in another condition where PTS was not as detrimental to performance (PTS=). Importantly, we observed this difference when we tested participants on the *same* set of problem types in both conditions following the same overall number of trials. This shows that experience with particular problem sets can carry over to others and induce a lasting effect on people's intervention strategies. More generally, the results suggest that a tendency towards using PTS is not a stable trait or bias and that people can adaptively select strategies based on overall features of a choice environment.

General Discussion

Previous work has argued that people adhere to optimal norms during information search in general (Nelson, 2005), as well as causal learning in particular (Steyvers et al., 2003). According to this view, people choose interventions that they expect to discriminate between a number of hypotheses and reduce their uncertainty about them. In contrast, we find that people were often better fit by a confirmatory positive testing strategy or a mixture between discriminatory and confirmatory models.

We also found large variability in the degree to which individual participants used either strategy, suggesting that it is

too simplistic to expect a *single* strategy to underlie people's interventions. Instead, our results imply that people have access to multiple ways of addressing intervention problems, and, as shown by Experiment 2, also have control over which strategy to use in a given environment. We find this latter result particularly noteworthy, because it means that people are perfectly capable of using the optimal strategy, but may only choose to do so if the payoff from the confirmatory strategy is significantly reduced.

Relation to other studies

In contrast to our results, Steyvers and colleagues (2003) found that a version of the IG model fit their participants' intervention data well. However, there are several differences between their experimental design and the present study. Most importantly, by showing people only two hypotheses with equal prior likelihood, we avoided having to make assumptions about which hypotheses participants consider at any point in time. The best fitting model of Steyvers' and colleagues relied on the assumption that participants only consider their favorite hypothesis and its subgraphs, but this was not made explicit to participants.

Similar to our findings in Experiment 2, other work has pointed out that hypothesis testing strategies can be changed from confirmatory to discriminatory behavior. However, these studies mainly manipulate different ways of framing the task altogether, for example by changing hypotheses to be normative statements (e.g. Cosmides, 1989; Cheng & Holyoak, 1985). In addition to those findings, our results show that people are sensitive to the expected payoff from a strategy, even given the same framing of the task.

Adaptive strategy selection

By definition IG is always as good or better than PTS, which raises the question why so many participants used PTS at all. Since we showed in Experiment 2 that people shifted towards IG, it does not seem to be the case that PTS use is a hard-wired "bias" or that people are universally unable to conform to the optimal model.

One possibility is that PTS might be a simpler cognitive strategy to use, since it does not involve repeatedly simulating and comparing the outcome of interventions under both hypotheses, as IG does. When the decrement in performance from using PTS is fairly small (as in Experiment 1), the extra computational costs involved in calculating something akin to IG could thus be outweighed by the benefit of simplicity.

It is also possible that some people might use PTS by default due to prior positive experience with it. Multiple authors have pointed out that PTS can be a good and even optimal strategy in situations when hypotheses are *sparse*, that is, when each hypothesis indexes only a small number of possible items in the world (Navarro & Perfors, 2011; Oaksford & Chater, 1994). Under our causal interpretation of PTS, this translates to a case where hypotheses predict very different effects (as in the PTS= condition of Experiment 2). Thus, if everyday experience prior to entering the experiment primarily consists of sparse hypothesis spaces, and assuming it has a lower mental cost to compute, this might explain the overall prevalence of PTS in Experiment 1.

Conclusion

In conclusion, we investigated how people interact with a simple causal system in order to discover how it works. We found that many participants did not behave in a purely discriminatory or confirmatory fashion, but at an individual level show behavioral signatures of both strategies. Furthermore, people are able to adapt their strategies in response to cues about the expected payoff of the strategy in the current task context. These results suggest a much more adaptive view of self-directed causal structure learning in humans than has so far been considered in past research.

Acknowledgments. This work was supported by grant number BCS-1255538 from the National Science Foundation and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023 to TMG.

References

Ahn, W.-k., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive psychology*, *17*(4), 391–416.

Coenen, A., & Gureckis, T. (2013). *When does a rational model "fit"?* Available from http://gureckislab.org/blog/?p=3710

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, *31*(3), 187–276.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Klayman, J., & Ha, Y.-w. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596.

Kruschke, J. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573-603.

Murphy, K. P. (2001). Active learning of causal bayes net structure. *Technical Report. Department of Computer Science, U.C. Berkeley.*.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review*, *118*(1), 120.

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4), 979.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453–489.

Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence* (Vol. 17, pp. 863–869).

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129–140.