Are Biases When Making Causal Interventions Related to Biases in Belief Updating?

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

Abstract

People often make decisions with the goal of gaining information which can help reduce their uncertainty. However, recent work has suggested that people sometimes do not select the most diagnostic information queries available to them. A critical aspect of information search decisions is evaluating how obtaining a piece of information will alter a learner's beliefs (e.g., a piece of information that is redundant with what is already known is useless). This suggests a close relationship between information seeking decisions on one hand, and belief updating on the other. This paper explores the deeper relationship between these two constructs in a causal intervention learning task. We find that patterns in belief updating biases are predictive of decision making patterns in tasks where people must make interventions learn about the structure of a causal system.

Keywords: information search; causal interventions; causal learning; hypothesis testing

Introduction

A growing body of work has explored how people use *interventions* to learn about causal structures in their environment (Bramley et al., in press; Coenen et al., 2014; Steyvers et al., 2003). An example of an intervention would be to take a vitamin in the morning to see how it makes you feel, or tapping on a button in a video game to see what happens. Critically, interventions are frequently made to gain information about the underlying causal structure of world.

The calculus of causal Bayesian Networks and information theory can yield precise normative predictions for the most diagnostic interventions to help a learner figure out a causal structure (Pearl, 2000; Murphy, 2001). However, in previous work (Coenen et al., 2014), we found that people do not always choose the most diagnostic interventions predicted by such theories. Instead, they frequently intervened on variables that would potentially lead to a lot of predicted effects in at least *one* hypothesized structure, but without necessarily distinguishing it from others. We called this tendency a "positive-testing strategy" (PTS), because its desire to "make expected effects happen" mimics the classic finding in the rule learning literature that people often ask questions that are likely to yield "yes" answers under a given hypothesis (Wason, 1960; Klayman & Ha, 1989).

What motivates participants to make positive tests instead of maximizing information during causal structure learning (or hypothesis testing in general)? This article explores one possible explanation of this behavior: It is possible that learners *do* try to maximize the information of their interventions but errors in the way that people update their beliefs draw them astray from optimal behavior. This hypothesis exposes the deeper relationship between intervention decisions/hypothesis testing and belief updating which has remained somewhat ambiguous in past research.

In the next section we will illustrate how biased belief updating might give rise to positive hypothesis testing. Then, we will describe two experiments that test whether or not people are biased in how they update their beliefs about causal structures, and if so, whether that bias influences their intervention strategy.

Information search & belief updating

Note that historically, confirmatory behavior has been identified in both information *search* and information *evaluation* (or belief-updating). For example, people seem to actively seek positive evidence for individual hypotheses (e.g., when they selectively research websites supporting their viewpoint rather than seeking opposing ideas). In addition, people also interpret evidence in a way that gives more weight to positive rather than negative outcomes (Nickerson, 1998; Klayman, 1995). For example, upon passively hearing of a politician's new budget proposal, people might focus on aspects that support their hypothesis about the general motivations of the politician. However, the link between these two different notions of "confirmation bias" is often unclear.

To illustrate how these two behaviors might be connected with each other and how belief-updating in particular can influence a learner's intervention strategy, consider the equation of Expected Information Gain (EIG). EIG is one of the most prominent models for human decision making during information search, including causal intervention learning (Steyvers et al., 2003; Nelson, 2005; Markant & Gureckis, 2012; Oaksford & Chater, 1994).

In a causal setting, the model calculates the relative informativeness of different interventions that a person could make on a causal system. Thus, the model is primarily a decision making model, where the goal is to obtain useful information. Formally, the model assumes that the learner's hypothesis space, G, consists of a set of possible underlying causal graphs. For a learner trying to minimize their uncertainty about G (i.e., discern which graph or set of graphs is most likely) the model calculates a score, $EIG(a)$ for each possible action or intervention a, taking into account all possible intervention outcomes, Y:

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

where $H(G)$ refers to the Shannon entropy over the posterior of the hypothesis space and $P(g|a,y)$ is obtained using Bayes' rule: $P(y|g,a)P(g)/P(y|a)$. The dependency of this equation on $P(g|a,y)$ highlights how the EIG equation always involves an explicit belief updating step. Intuitively, EIG evaluates an action by imagining how different outcomes as a result of that action would change the learner's belief. Thus, according to the model, belief updating is *fundamental* to judging the information value of an action or decision. The goal of this paper is to explore if this intrinsic relationship holds for human reasoners.

Potential updating biases and their impact on decision making. It is interesting to consider the ways that incorrect belief updating would alter causal intervention decisions. Let us momentarily take it as given that people use the Equation 1 to evaluate the informativeness of an intervention. How could it be that their choices appear suboptimal?

One possibility is that people may be biased in how they assess $P(y|g,a)$, the likelihood for an outcome to occur after an intervention on a specific graph. Specifically, they might assign higher likelihoods to outcomes that activate larger portions of a graph. Such a tendency would be in line with previous work showing that people are more strongly affected by positive rather than negative evidence when evaluating hypotheses (for a summary, see Klayman, 1995). If outcome likelihoods are higher for positive observations (e.g. full activation of a graph), EIG will also be higher for interventions that lead to those observations.

As another possibility, learners may also deviate from a normative account in how they incorporate the prior probability of an intervention outcome, $P(y|a)$. Previous work has shown that people often exhibit a tendency to disproportionately evaluate hypotheses based on their likelihood of producing an event or outcome alone, irrespective of its probability of occurring under alternative hypotheses (Kahneman & Tversky, 1973, 1972; Doherty et al., 1979). In this case, if learners decrease or ignore $P(y|a)$, they will be more likely to choose interventions that are not actually diagnostic, because the same outcome is predicted by multiple graphs. In that case, they might take an outcome to be supportive of *one* of the graphs, while ignoring that it's equally well supported by an alternative one.

In summary, different biases in belief updating could, according to the theory, alter decision making strategies even if people otherwise perfectly followed the basic (normative) tenets of Equation 1.

Explaining intervention data. To make this discussion more concrete, we will describe two examples of biased intervention strategies from our past findings (Coenen et al., 2014). In this previous study participants were presented with a virtual "computer chip" made up of various components. In addition, they were provided with two possible wiring diagrams that could detail how the components were connected to one another. Interventions activated a component which in turn (might) activate other components. Participants' task was to intervene on different chip components in order to learn the true wiring diagram.

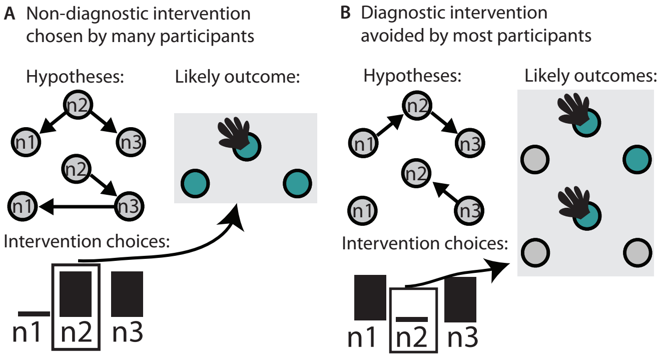

Figure 1: Example intervention problems with choice data from Experiment 1 in Coenen et al. (2014). Participants were presented with two possible wiring diagrams (shown under "Hypotheses") and were asked to make an intervention on the system to determine the correct causal structure. The height of the black bars under "Intervention choices" show the frequency by which different elements of the causal system were intervened on by participants. For example, in panel A, no participants selected node $n1$, but people were split between $n2$ and $n3$.

In our previous study, we found that participants chose interventions that have low EIG if these queries offer the possibility of activating a full causal graph. For instance, in Figure 1A many participants chose to intervene on the root node, n2, of both graphs, which most frequently led to confounding evidence of *all* nodes turning on. This behavior could have been caused by the two types of updating biases described in the previous section. That is, if participants ignore the fact that the same outcome be produced by both hypothesized graphs (neglect of $P(y|a)$) and they give more weight to positive outcomes (increase of $P(y|g,a)$ if a graph is completely activated), intervening on n2 becomes an attractive option because it can produce all predicted effects in either graph. Learners might think that the confounding "all-on" outcome actually provides strong evidence for *one* of the graphs (which may be chosen randomly, or based on a learner's prior preference).

We also found that by choosing interventions with expected positive outcomes, participants often forgo queries that have very high EIG, such as portrayed in Fig 1B. Here, intervening on n2 has the potential of discriminating between the two graphs, but does not offer a chance to see either of them be completely turned "on", because the outcome would either be a half-activated chain graph or nothing happening whatsoever. If people weigh positive evidence higher than negative (or less positive) evidence then perhaps both of these outcomes would be deemed less informative than they actually are, which could explain why most participants chose one of the other two interventions, n1 or n3.

Overview of the present study

The two experiments reported in this paper attempt to assess the relationship between belief updating and intervention choice. In particular, we assumed that there are individual differences in belief updating that might explain variation in decision strategies.

In order to quantify difference in belief updating, in Experiment 1 we created a causal learning task where people were instructed to make interventions on a causal system and to subsequently rate the posterior belief about the correct graph. Trials were constructed to be revealing about biases towards the overweighting of positive but non-diagnostic evidence compared to less positive (or negative) diagnostic evidence.

In Experiment 2 we attempted to establish a link between updating strategies and intervention choice strategies. In addition to the instructed trials designed to assess belief updating, participants also had to make a range of self-selected interventions. Using model-based analyses we quantified the tendency of those choices to conform to either information-maximizing interventions (in line with EIG) or PTS interventions and related these scores to the estimates made in the belief updating phase of the task.

Experiment 1

Participants

Sixty-one US-based participants were recruited via Amazon Mechanical Turk. They were paid $2 for participating in the study.

Stimuli

The same set of 20 causal intervention problems (10 critical trials, and 10 control trials) were given to each participants. Each problem consisted of two causal hypotheses (three-node graphs), a prescribed intervention (node), and the outcome of that intervention (which nodes turned "on" as a consequence). The *critical* trials were specifically designed to reveal any belief updating bias that we hypothesized to underlie positive hypothesis testing, that is, they were interventions in which the outcome was either diagnostic but without activating any graph completely (like in Figure 1B), or it was non-diagnostic, but showing all expected outcomes of at least one graph (like in Figure 1A). The *control* trials were either diagnostic and showing a graph activated completely or non-diagnostic and without any full graph's activity. Thus, they did not pose the same conflict between diagnosticity and positive outcomes.

The two trial types were randomly interspersed without the knowledge of the participants. When causal graphs were presented on the screen the order of nodes was always randomized and they were also randomly placed in three out of five positions on the screen.

Procedure

Participants were told that they had to test a series of computer chips in a chip factory to figure out how they worked.

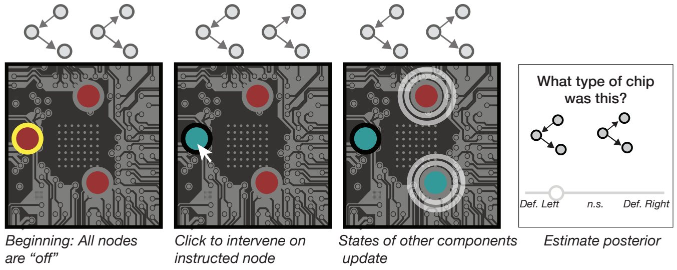
What type of chip
was this?
Def. Left
n.s.
Def. Right
Beginning: All nodes are "off"
Click to intervene on instructed node
States of other components update
Estimate posterior

Figure 2: Schematic procedure for the instructed intervention trials. Participants were asked to intervene (click on) the node with the yellow outline and subsequently observed the outcome of their intervention (other nodes turning on or remaining turned off). Using a slider, they then gave their posterior estimate about which hypothesized graph was more likely to underlie the chip they just intervened on.

For each chip, they were given two possible arrow diagrams that illustrated the two hypothesized causal graphs that explain the working of a chip (see Figure 2). They were then instructed to turn on one specific "component" (node) on the chip by clicking on it and observe the outcome (other components also turning on or remaining turned off). Next, they were asked to rate which chip diagram was an accurate description of the chip they just tested. They gave their answer using a continuous slider with three labels on the left, middle, and right ("definitely Type 1", "not sure", "definitely Type 2").

Before starting the task they were explicitly told that causal links between components only worked 80% of the time and that components could only be turned on by each other or an intervention, but not by any other background causes. Participants had to accurately pass a short quiz about the task before being allowed to proceed to the main part of the experiment.

Results & Discussion

Figure 3 shows histograms of participants' posterior estimates after each intervention/outcome trial. The red dot indicates the true posterior probability of the graphs derived using Bayes' rule and the correct outcome likelihoods of the two graphs given each intervention. The red arrows in the critical trials indicates the direction in which one would expect participants to be biased based on our predictions outlined above. We would like to particularly guide the reader's attention to three findings.

First, especially on the control trials participants' estimates track the true posterior probabilities of the graphs well, that is, the highest density of choices is typically found at or very close to the red dots in the plot. This shows that generally participants understood the task and were able to evaluate intervention outcomes correctly.

Second, the mapping between true and estimated probability is less pronounced on the critical trials (see panel A). Here, participants' estimates are often skewed in the direction of the red arrow, that is, in line with the above predictions about updating biases. Without discussing every single deviation in detail, estimates on the objectively non-diagnostic trials (i.e. those with a posterior probability of 0.5) in the first four trials from the top are particularly noteworthy. Here, the systematic deviation from the true posterior cannot be due to a tendency towards indifference (probability of 0.5). Consider for example trial no. 2, in which many participants strongly endorsed the One-Link graph on the left, even though the evidence equally supports the alternative Common Effect graph on the right. Again, the reason we suspect this deviation happens is that participants place too much emphasis on the fact that they can observe all possible activity predicted by the One-Link structure, but not of the Common Effect.

Thirdly, another interesting deviation from the true posterior is found on the very first trial type, in which all components are turned on after an intervention on the middle node, which is the root of both a Chain and a Common Cause graph. As discussed earlier, this root node intervention is one that many participants actually chose in a previous experiment (see Figure 1A) and this finding might be able to help explain why: Participants may have thought that they can actually learn something from the all-on outcome. It is puzzling, however, that participants mainly endorsed the Common Cause structure rather than the Chain, since the outcome provides equal positive evidence for both. One could speculate that the immediacy of the Common Cause links in connection to the root node may have contributed to this tendency, but ultimately this finding requires further investigation. In any case, this result demonstrates a relatively strong violation of optimal belief updating.

In sum, we find that participants do exhibit a tendency to erroneously change their probability estimates away from indifference if one or both of the structures are completely activated through an intervention. They also do not change their beliefs enough if a diagnostic outcome is not activating a complete graph. However, this bias is by no means found to be equally strong in all trials and participants' estimates were much more accurate in the control trials, showing that generally people's updating process tracks the correct posterior probability of the graphs.

Experiment 2

Having found that participants show greater deviations from optimal belief-updating when diagnosticity and positive outcomes of individual graphs were at odds with each other, the second experiment aimed to find out whether the tendency to show this bias is at all related to people's intervention strategy. Since we argue above that biased belief-updating might be the reason that people often conduct positive tests rather than maximally diagnostic ones, we predict that participants with biased belief updating should be more likely to conduct positive tests.

To test this, we added a free intervention task to the experiment which participants completed either before or after a series of instructed interventions. The goal was to quantify both a participant's intervention strategy and their belief up-

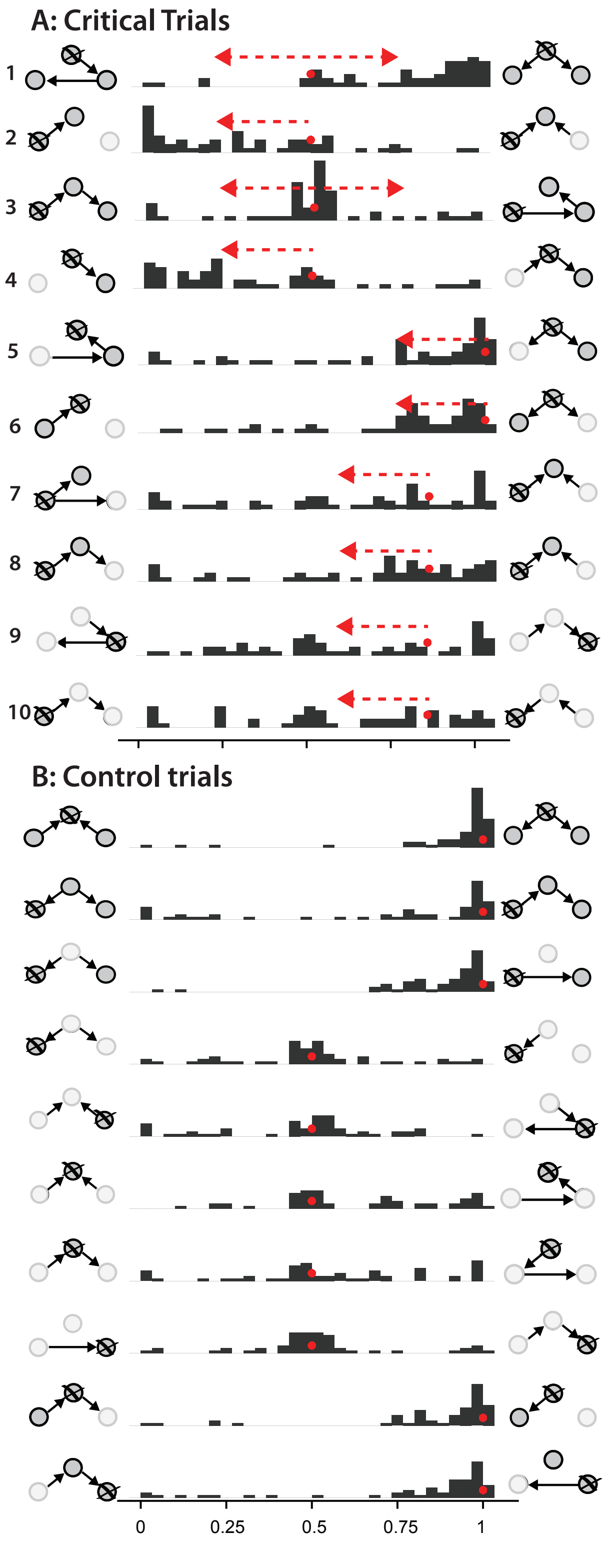

Figure 3: Histograms of participants' probability estimates. Each row represents one trial type in which two causal hypotheses were compared. Crosses indicate nodes that participants were instructed to intervene on (turn the nodes "on"). Grayed out nodes remained "off" after the intervention, the remaining nodes were also observed to be "on". Red dots in the critical trials indicate the posterior graph probability according to accurate Bayesian belief updating. Red arrows indicate the predicted direction of positive outcome bias.

dating bias to investigate if there exists a connection between the two.

Participants

121 US-based participants were recruited via Amazon Mechanical Turk. They were paid $2 for participating in the study.

Stimuli

In the instructed intervention phase, participants saw the same 20 trial types as in Experiment 1 with one small change. The two trials in which the predicted bias was bi-directional (trials 1 and 3 in Figure 3) were replaced with trials that had a unidirectional predicted bias. This was done to make it easier to quantify updating biases specifically in one direction as pertaining to our predictions, without confounding it with general noisiness in participants' use of the slider, for example.

In the free intervention phase, participants were given 20 intervention problems that were used in a previous set of experiments reported in (Coenen et al., 2014, see Figure 2), and were used to characterize people's intervention strategies on a continuum between positive testing (PTS) and information maximization (EIG).

Procedure

The procedure was identical to that in Experiment 1 except that in the free intervention trials participants were instructed to choose freely which node to intervene on. Again, there was no feedback about the correct structure at the end of a trial.

Results & Discussion

To quantify people's intervention strategies we used the method developed and reported in Coenen et al. (2014) which results in a strategy weight θ that indicates the degree to which a participant's interventions are in line with the EIG strategy of information search ($\theta = 1$) compared to a confirmatory PTS strategy ($\theta = 0$). In comparison to our own previous work on intervention strategies, the current experiment yielded a larger portion of PTS-interventions and fewer participants that were strong EIG users (as a rough indication, only 30% of participants were fit with $\theta > .5$, compared with 47% in our previous study). Due to the unequal distribution of participants' best-fitting θ values we use a median split to divide participants into two equal groups of low-θ and high-θ for the analyses reported below, bearing in mind that the high-θ group contained many participants from the middle of the distribution, however.

To quantify each participant's tendency to commit the updating error we hypothesized and found in Experiment 1, we first computed the average deviation of participants' probability estimates from the true posterior *in the hypothesized direction* on the critical trials (i.e. in the direction of the arrows in Figure 3A). Because this deviation score will be higher if a participant is generally noisy in their posterior probability estimation, we regressed this directional deviation on the average absolute deviation in the control trials (see Figure 3B) and used the residuals as a measure of positive-outcome bias.

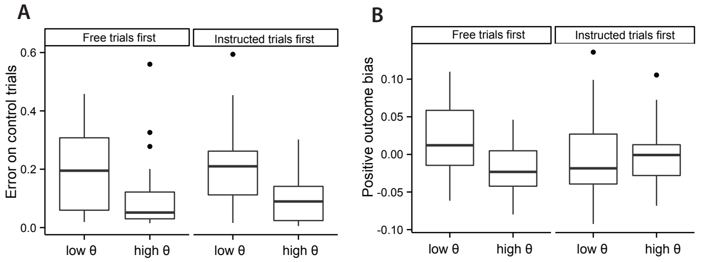

Figure 4: Updating error and positive outcome bias by strategy weight and task order. Low-θ participants were better fit by a positive testing strategy and high-θ participants by the EIG model.

For brevity's sake, we will only report two analyses that relate the binary strategy weight with the the quality of belief-updating. Figure 4 shows how the two groups of participants (low-θ and high-θ) compare in terms of overall error (average absolute deviation from the true posterior in the control trials) and on the positive-outcome bias measure (residuals after controlling for overall error). The plots are split by the order of the instructed and free intervention tasks, which was counterbalanced between participants.

Note that there exists a significant relationship between strategy weight and error on the control trials, $t(119) = 4.84$, $p < .001$. Participants with more discriminatory intervention choices (more in line with EIG), also made more accurate probability estimates than participants better fit by a PTS strategy.

The relationship between strategy weight and positive-outcome bias, on the other hand, is much weaker. Overall, it is still significantly negative, so that bias is lower for high-θ participants, $t(119) = 2.34$, $p = .02$, but the effect seems driven by participants in the group that received the free intervention task first.

One caveat of this analysis is the way in which the positive-outcome bias is defined. In an effort to isolate it from a participant's general tendency to make errors on the posterior estimates, it is possible that some relevant variation may have been lost. For example in people who are both biased and noisy, removing the noise component may have made it look as if they showed no bias at all. Taken together with the lack of high-θ participants, the last analysis need to be treated with some caution and should be backed up by further experiments.

General Discussion

In this paper we explored how people update their beliefs after performing causal interventions and observing their outcome. Having previously found that participants often perform interventions that can cause positive outcomes in individual graphs, we predicted that this may be due to a tendency to treat these outcomes as particularly informative (whether or not they actually are).

Several findings from these experiments stand out.

First, it is important to note that participants were often very good at updating their beliefs in a normative fashion, particularly when diagnosticity and outcome positivity (i.e. the degree to which outcomes involve all effects predicted by individual graphs) were not in conflict. This finding ties in with earlier work demonstrating that people are effective causal learners who understand the basic mechanism of an intervention (see e.g., Lagnado & Sloman, 2004; Hagmayer et al., 2007; Bramley et al., in press).

However, Experiment 1 showed that there was a greater tendency to deviate from the true posterior probabilities when diagnosticity and outcome-positivity were at odds. On those trials, we found considerable deviation from optimal belief-updating in participants' posterior probability estimates. In particular, participants often endorsed graphs more strongly than they should if all of their predicted effects could be observed. This shows that people are not purely engaged in Bayesian belief-updating when observing intervention outcomes. Instead, they may sometimes be influenced by the degree to which outcomes reflect a fully activated causal structure, while ignoring the question whether or not an outcome actually discriminates between structures.

Experiment 2 further showed that people who conduct more positive tests when choosing interventions were more likely to commit belief-updating errors in general. Thus there appears to be a relationship between people's intervention strategy and their subsequent ability to learn from these interventions. Whether or not this relationship is specific to the positive-outcome updating error as we hypothesized, and which was found in Experiment 1, remains an open question. Although Experiment 2 found a significant relationship between strategy and positive-outcome bias overall, it was a relatively weak one.

As we pointed out, the data in Experiment 2 had some undesirable properties, such as a lack of variability in participants' intervention strategies, with a much larger number leaning towards positive testing, rather than discriminatory search. It also proved challenging to disentangle the general tendency to commit updating errors with the specific type of bias that we intended to isolate. Future work should therefore aim to find a better method to distinguish the two. Concretely, we suggest a follow-up experiment with a stronger manipulation, such as training participants on how to evaluate intervention outcomes and test whether that has an effect on subsequent intervention decisions. Such a manipulation would be able to show a more direct link between these two aspects of intervention learning.

In sum, we believe that these studies offer a first attempt to study how people update their beliefs about causal systems from intervention data and the experiments reported here show some noteworthy patterns of errors that affect participants in this process. Future research is needed to clarify exactly how causal learning and hypothesis-testing interact.

Acknowledgments This work was supported by grant number BCS-1255538 from the National Science Foundation, the John Templeton Foundation "Varieties of Understanding" project, and a John S. McDonnell Foundation Scholar Award.

References

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (in press). Conservative forgetful scholars - how people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Coenen, A., Rehder, B., & Gureckis, T. (2014). Decisions to intervene on causal systems are adaptively selected. In P. Bello, M. Guarini, McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society.* Austin, TX.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta psychologica*, *43*(2), 111–121.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, 86–100.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation*, *32*, 385–418.

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.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Markant, D., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th annual conference of the cognitive science society. austin, tx: Cognitive science society.*

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

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.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Cambridge Univ Press.

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

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