Updating Beliefs With Causal Models: Violations of Screening Off

Clare R Walsh
University of Plymouth

CHAPTER
21

Updating Beliefs With Causal Models: Violations of Screening Off

Clare Walsh
University of Plymouth, England

Steven Sloman
Brown University

Upon encountering a line of research in cognitive psychology, it is rare to find that it does not trace back in one way or another to Gordon Bower. Gordon is like the taproot of cognitive psychology, stouter than other cognitive psychologists and tending to develop along straight lines.

Gordon is not only an experimentalist's experimentalist and the granddaddy of computational models of cognition, he is also a patriarch of mathematical psychology. Gordon was developing Markov models of memory processes before the field was old enough to say "rehearsal buffer." Indeed, as a new student of Gordon's, the second author remembers vividly sitting in Gordon's office discussing a methodological issue. Gordon said, "Oh yeah, we worked that out about 25 years ago," took out a pad of paper and quickly proceeded through three pages of mathematical analysis to prove to the young upstart that aggregate and individual performance could not be distinguished (or some topic like that. Gordon remembers detailed proofs; the second author only vaguely remembers discourse topics.)

At one point, Gordon suggested to the second author that he do something worthwhile and rigorous for his dissertation, like develop a probabilistic model of his idea. At the time, Gordon's idea was rejected as old-fashioned. Fifteen years later, the second author's research has turned to probability models, a topic he wishes he had studied more thoroughly as a younger man.

In the early 1960s, Markov models were all the rage at Stanford, models that are explicit about how much of an individual's history is required to determine the mental state of a cognizer. One basic idea of such models is that we can model the probability of some mental state by considering only a finite set of prior mental states, often only one. Call that set of relevant states R. Then mental states prior to the critical states R are rendered independent of the current mental state by knowledge of R. This is a form of what is these days called "screening off": Once you know R, events prior to R cannot tell you anything about events subsequent to R. States in R are able to screen off prior states from the current state. Like all ideas that interest Gordon, this is not only a nice general property that, if true, would reveal a lot about the nature of cognition, it is also highly testable as it asserts a clear prediction about human behavior. In Gordon's field of memory, the prediction in the simplest case is that what people remember at time *t* should be a function of what they know about events at time $t-1$. Once a person's knowledge at time $t-1$ is specified, events before that should provide no further information about what people will say at time *t*. Events at $t-1$ screen off events at time *t* from events prior to $t-1$. Screening off is a property that holds of a large class of graphical probability models, models that attempt to represent human knowledge and cognitive processes using nodes and links in a way that relates them intimately to probability distributions. For instance, screening off holds not only of Markov models, but of Bayes's nets (Glymour, 2001; Sloman, 2005).

BELIEF UPDATING

In this chapter, we examine the validity of screening off as a descriptor of how people revise their beliefs about simple events that they have causal knowledge of. How do people update their beliefs in the face of new information? Imagine for example that four students move into the apartment above you. You imagine that their apartment may be quite untidy. You may also worry that they may play music late in the evening and so on. Imagine they then have a party during their first weekend there. When people encounter information that is consistent with their expectations, then that information may reinforce their existing views.

However, sometimes people encounter information that is inconsistent with what they expected. Imagine that during the first weekend after the students move in you don't hear a sound. In some cases, people may choose to discredit this information (Lord, Ross, & Lepper, 1979). For example, they may still expect the students to be noisy. However, if they accept that their new neighbors are genuinely quiet, then they will need to change some of their existing beliefs. But what beliefs will they change? For example, will they also revise their belief that the students' apartment is untidy?

The question of how people update their beliefs when they encounter unexpected information has been addressed by psychologists (e.g., McKenzie, 2004), philosophers (e.g., Harman, 1986), and artificial-intelligence researchers (e.g., Gardenfors, 1988). The predominant view is that people accommodate new information by making the minimal possible change to their existing beliefs. When they discover that a cause occurs but its usual effect does not, one option is to adjust belief in the strength of the relation between these two events. For example, if we know there was heavy rain but the tennis match wasn't canceled, we may revise our belief that rain causes these matches to be canceled and we may lower our expectation that the match will be canceled next time it rains. However, when people encounter an unexpected outcome, they tend to construct explanations for it. These explanations often describe disabling conditions (i.e., factors that prevent a cause from producing its usual effect; Byrne & Walsh, 2002; Walsh & Johnson-Laird, 2007). For example, they may suggest that this tennis match was played under cover. We propose that these explanations may determine how people update their beliefs.

In the current experiments, we constructed pairs of causal statements that shared a cause. Bayes's nets provide an excellent medium for representing this kind of causal knowledge. According to Bayes's net theory, when two events have a common cause, they tend to be correlated. The reason for this is that they both tend to be present when their cause is present and absent when their cause is absent. However, in cases in which we know whether or not the cause has occurred, this correlation disappears. Hence, we can no longer infer anything about the likelihood of one effect from the knowledge that the other event did or did not happen. In this case, the cause screens off judgments about one event from information about the other (see Fig. 21–1). Based on this principle, if a cause occurs (e.g., it rains) and one effect does not (e.g., the tennis match is not canceled), people's strength of belief that the second event will occur (e.g., umbrellas are sold) should be left unchanged. However if people construct explanations for why the effect didn't occur, they may use these explanations to revise their expectations. Along with examining whether people's judgments are consistent with the screening-off principle, our studies were designed to examine whether explanations play a role in belief updating.

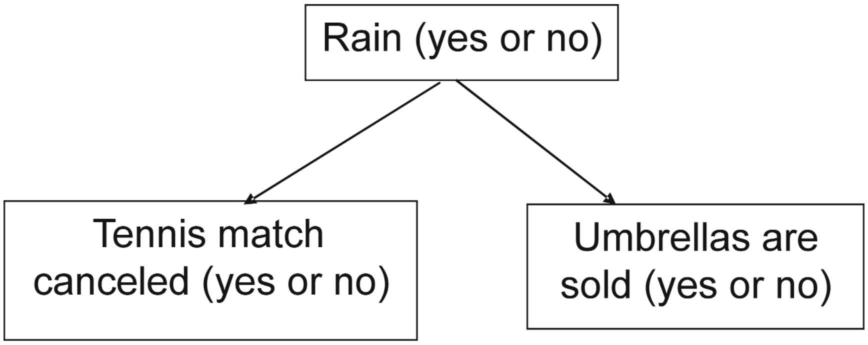
Rain (yes or no)
Tennis match canceled (yes or no)
Umbrellas are sold (yes or no)

Figure 21–1. Screening off in a common cause model. On a particular day, if we know whether it has rained or not, there's nothing more we can learn about whether umbrellas are sold from knowledge about whether the tennis match was canceled. For instance: P(Umbrellas are sold | It rained) = P(Umbrellas are sold | It rained, Tennis match not canceled); Rain (yes or no); Tennis match canceled (yes or no); Umbrellas are sold (yes or no).

UPDATING BELIEFS

In our first experiment, we presented 23 Brown University undergraduates with pairs of causal statements (methodological details are reported in Walsh & Sloman, 2004). In each case, the two statements shared the same cause but had different effects. Hence, they described a common cause model. An example is:

Following this diet causes you to have a good supply of iron.

Following this diet causes you to lose weight.

After presenting participants with these statements, we then asked them two questions. First, we measured their initial belief in one of the statements by asking them to judge the probability of the effect given that the cause occurred. For example:

John followed this diet. What is the probability that he lost weight?

We then presented participants with new information. Specifically, we told them that the second effect (i.e., the one that we had not previously asked about) did not occur. Our goal was to see how people revise their judgments in the light of this new information. Participants judged again the probability of the same effect, this time given that the cause occurred and the second effect did not. For example:

> *Tom followed this diet but he did not have a good supply of iron. What is the probability that he lost weight?*

If probability judgments are compatible with the screening-off principle, then judgments before and after the new information should be the same. Once it is known that the cause occurred (Tom followed the diet), any correlation between the two effects (supply of iron and weight loss) will disappear. Hence, knowing that one was absent should have no effect on the likelihood of the second. However, if people do not abide by the screening-off principle, then their judgments may change.

The results showed that participants lowered their judgments in the second question on 56% of trials (means are presented in Table 21–1). Twenty-one out of 23 participants lowered their judgments in at least one problem and the remaining 2 showed no change in any problem. There was no difference in the change in probability across the different problem contents. We also gave participants two control problems that contained no unexpected information. For example:

> *Tom followed this diet and he did have a good supply of iron. What is the probability that he lost weight?*

In this case, there was no significant change in judgments after reading the new information (see Table 21–1). When events were consistent with causal expectations, people did screen off.

Table 21–1
Mean Judged Probability of Event B Under Various Conditionalizations for Experimental and Control Problems in Experiment 1

Abstract problem format	
If A then B	If A then C
Experimental Problems	
Given A	77
Given A and not C	59
Control Problems	
Given A	81
Given A and C	82

The results show that when events were inconsistent with causal constraints, judgments were not consistent with the screening-off principle. One explanation for this result is that people do not understand this principle. They may not realize that once a cause is known to have occurred, then any effects of that cause are no longer correlated; people continue to believe that knowing whether one effect occurred provides information about other possible effects. However, an alternative possibility is that when people encounter unexpected outcomes they construct explanations for why they occur. These explanations often refer to disabling conditions; they may conclude that this diet affects only some people. As a result, explanations may lead people to add new factors to their causal models and to update their beliefs on the basis of this new model. When two effects share a common cause, then disabling conditions for one outcome may often disable the second outcome also and as a result people may adjust their judgments regarding the likelihood of a second outcome, that everyone will lose weight on this diet. The aim of our second experiment was to test this hypothesis. We compare it to an alternative account of our result, that contradictory information simply reduces confidence in all related judgments.

UPDATING BELIEFS AND EXPLANATIONS

In this experiment, our aim was to test the role of explanations in belief change. We did so by explicitly asking participants to generate explanations when we gave them the unexpected information that one of the effects did not occur. The study addressed two main questions. First, we wanted to know whether causal judgments depend on the explanation that was generated for the contradiction and on that explanation alone. If their probability judgments depend on their explanations, then their judgments should be predictable from their explanation regardless of the contradicted fact. In contrast, if a contradiction just reduces confidence, then their probability judgments should vary with contradiction, and not with the explanation. We tested this by explicitly asking participants to generate an explanation for the contradiction before making a causal judgment, for example:

Jogging regularly causes a person to increase their fitness level.

Jogging regularly causes a person to lose weight.

(Q2) Anne jogged regularly but she didn't lose weight.

Why?

What is the probability that her fitness level increased?

We then asked participants to use this explanation to make another causal judgment. For example, if a participant gave the explanation that Anne's appetite increased, then we asked them the following question:

(Q5) John jogged regularly and his appetite increased.

What is the probability that his fitness level increased?

The full sequence of questions is illustrated in Table 21–2. If people use their stated explanation (and not the contradicted fact) to make the causal judgment in Question 2 and they don't consider any other hypotheses, then we expect the probability judgments in Questions 2 and 5 to be equal. Previous research has shown that people frequently neglect alternative hypotheses (e.g., Klayman & Ha, 1987). However, if reasoners do consider other explanations or if their causal judgments are reduced merely because they have less confidence in what they have been told, then we expect these judgments to differ.

The second question that we addressed in this study was whether people draw on information that already exists in their causal model to generate an explanation for a contradiction or whether resolving a contradiction leads people to revise the causal model itself. We did this by asking participants two further questions. Before reading the contradiction we asked them the following:

(Q1) Tom jogged regularly.

What is the probability that his fitness level increased?

And after reading the contradiction and generating the explanation, we asked them the following:

(Q3) Mary jogged regularly and you don't know if her appetite increased.

What is the probability that her fitness level increased?

If a reasoner's causal model already contains information about the relation between appetite and fitness level and they use this information in answering Question 1, then we expect their responses to Questions 1 and 3 to be equal. But if they change their causal model when resolving the contradiction, we expect their answer to these two questions to be different.

The results of this study replicated the results of the previous study (see Table 21–2). The probability of the second effect was rated as significantly higher in Question 1 before reading the new information ($M = 85\%$) than in Question 2 after reading the new information ($M = 63\%$).

Our second finding was that responses to Question 2 and Question 5 ($M = 62\%$) did not differ significantly and this pattern occurred for all six types of problem content. For problems in which the contradiction reduced the judged probability of B (response to Question 2 was lower than to Question 1), participants gave the same answer to Questions 2 and 5 for 53% of problems. We would expect greater variety if participants were considering multiple hypotheses. Hence the results are consistent with the view that in many cases people consider just the one hypothesis given in their explanation and they fail to consider other possibilities. They allow this hypothesis to mediate their later causal judgments without considering the possibility that they are wrong (see also Shaklee & Fischhoff, 1982).

Table 21–2
The Format of the Problems Used in Experiment 2.

	Mean
Worrying causes difficulty in concentrating.	
Worrying causes insomnia.	
1. Mark was worried. What is the probability that he had difficulty concentrating?	85%
2. Kevin was worried but he didn't have insomnia. Why? What is the probability that he had difficulty concentrating?	63%
3. Frank was worried but *you don't know if the explanation holds.* What is the probability that he had difficulty concentrating?	71%
4. Helen was worried and *you know that the explanation does not hold.* What is the probability that she had difficulty concentrating?	85%
5. Evelyn was worried and *you know that the explanation does hold.* What is the probability that she had difficulty concentrating?	62%

Note. In question 2 participants provide an explanation and this is used in questions 3, 4 and 5. Mean judgments for questions of each type are reported.

Finally, our results suggest that people resolve contradictions by making a change to their causal model. Ratings for Question 1 were significantly higher than for Question 3 when the explanation was unknown (M = 71%). People do not merely change their causal judgments about the specific case in which the contradiction occurred. They extend these changes to new situations. Responses to Question 1 did not differ significantly from responses to Question 4 (M = 85% for both). People do not generally resolve contradictions by drawing on events that they have already represented in their causal model.

We also examined the nature of explanations given for the inconsistency. The most common explanation was to introduce a disabling condition that would prevent the cause from producing its usual effect. Seventy-four percent of responses were of this type. In many cases, the conditions disabled the cause from both consequences. For example, the fact that worry did not lead to insomnia may be explained by the fact that the person did relaxation exercises. This in turn may reduce the probability that worry will lead to a difficulty in concentrating. The next most common type of response was to suggest that the level or amount of the cause was not sufficient to produce the effect; for example, the person was not very worried. Eighteen percent of responses were of this type. In both cases, the pattern of responses and significance ratings for the probability of B were the same as for the overall ratings. We return to this question in our next experiment.

TYPES OF EXPLANATIONS

In our third experiment, we examined different types of explanations and their influence on belief revision. This time we provided people with explanations for why an unexpected outcome didn't occur and we examined the effect of these explanations on belief change.

Once again in this study we presented people with pairs of causal statements such as the following:

> *Playing loud music in Paul's apartment causes the neighbors on his left to complain.*
>
> *Playing loud music in Paul's apartment causes the neighbors on his right to increase the volume on their TV.*

As in the previous studies, we compared people's judgments about the likelihood of an outcome given that the cause occurred:

> Last Friday night, Paul was having a party and he played loud music. What is the probability that the neighbors on the left complained?

to the likelihood of the same outcome given that the cause occurred and the second effect did not. However in this study, we provided people with one of three types of explanation for why the second effect did not occur and we predicted that these explanations would have different effects on belief change:

> *This Friday night, Paul was having a party and he played loud music but the neighbors on the right don't turn up the volume on their TV. [Explanation provided or generated here.] What is the probability that the neighbors on the left complained?*

The first type of explanation described an additional condition that would prevent both effects from occurring, for example:

> *You discover that Paul invited all of his neighbors to the party.*

This condition would provide a reason for why the neighbors on the right did not turn up the volume on their TV but it would also reduce the likelihood that the neighbors on the left would complain. They may have been at the party. For this reason, we expected that the probability of the first effect would be judged to be lower after learning that the second effect didn't occur.

The second type of explanation that we gave people described an additional condition that would prevent only one effect from occurring, for example:

> *You discover that they are out.*

This condition again provides a reason for why the neighbors on the right did not turn up the volume on their TV. But, the influence of this on judgments of the first effect is less clear. If judgments are based on these new facts alone, then the probability of the first effect should not change. We might expect that the probability of the neighbors on the left complaining would be the same given that Paul had a party as it would be given that Paul had a party and the neighbors on the right are out. However, an alternative possibility is that this explanation might bring possible disabling conditions to mind. In this case, people may judge the second question to ask the probability that the neighbors on the left will complain given that Paul had a party and it is not known whether the neighbors on the left were at home. If in their earlier judgment, people made the default assumption that the neighbors were at home, then they may change their judgment now as they did in Experiment 2.

Finally, the third type of explanation that we gave also prevented only one of the effects, for example:

You discover that they are not watching TV.

Again this provides a reason for why the neighbors on the right did not turn up the volume on their TV, but we would not expect these facts to influence the likelihood that the neighbors on the left would complain. This explanation differs from the previous one in that it doesn't provide a plausible disabling condition for the first effect. In other words, we might expect that the probability of the neighbors on the left complaining given that Paul was playing loud music to be the same as they would be given that Paul was playing loud music and it is not known whether or not the neighbors on the left are watching TV. If people are basing their judgments on the explanations that we provided, then we expected the probability of the first effect to be the same for both questions.

We presented 20 participants with a series of scenarios and provided them with one of these three explanations or we asked them to generate their own.

Our results appear in Table 21–3. After reading the first type of explanation, which disabled both effects, participants changed their judgment of the probability of the outcome from 89% to 34%. After reading the second type of explanation, which prevented the first effect only but provided a plausible disabling condition for the second, participants also changed their judgment of the probability of the outcome significantly from 84% to 72%. This result supports the view that people begin by making default assumptions that the necessary background conditions are present but explanations can remind them of these conditions and lead them to question their earlier assumptions. Finally, after reading the third type of explanation, participants also changed their judgments significantly from 84% to 76%, $p < .02$. The results show that even giving people explanations that are not relevant to a particular judgment can lead to changes in those judgments. One possible reason for this outcome is that explanations may have the more general effect of reminding people that things may not go as expected and that their default assumptions may be wrong. For example, people may make a new judgment about the probability that the neighbors on the left will complain given that Paul played loud music and it is not known if other background conditions are met. When participants were asked to provide their own explanation, they generated each of the three different types and accordingly, they changed their judgments significantly from 89% to 67%.

Table 21–3
Mean Judged Conditional Probability of Effect B for Baseline Condition and After Different Explanations of the Non-Occurrence of Effect A in Experiment 3

Explanation Type	Before told A did not occur	After told A did not occur and explanation
Both effects disabled	89%	34%
Plausible application to B	84%	72%
Implausible application to B	84%	76%
Generate Own	89%	67%

SUMMARY AND CONCLUSIONS

In three experiments, we have provided support for the view that explanations play an important role in belief updating. In the first study, we demonstrated that when people encounter an unexpected outcome, they often fail to follow the screening-off principle and they diminish their expectation that other outcomes of the same cause will occur. One explanation for this is that people seek to explain outcomes and these explanations lead them to add new elements to their causal models. Our second study tested this hypothesis by asking people to make judgments after generating an explanation and then to make the same judgment assuming that the explanation held. The majority of responses were the same in both cases, lending support to the view that their explanations underlie their change in judgment. Finally, in our third experiment we presented people with explanations and we showed that the extent to which people change their judgments is influenced by the nature of these explanations. This suggests that explanations may change judgments through three different effects. One way is by suggesting other relevant facts that change the likelihood of an outcome. The second is by reminding people of specific conditions that could be relevant. Finally, explanations may remind people more generally that there could be conditions that are not met. In each of these ways, explanations may have an impact on the judgments that people make.

Like so many studies of human reasoning, these data undermine the old idea that human reasoning can be understood using systems of monotonic logic (see Evans, 2002). Some form of nonmonotonic inference—such as probabilistic reasoning—is necessary to describe how people think.

Can human reasoning be described using graphical representations of probability? If so, then it is important that judgments satisfy screening off, one of the most basic inferential principles of many such models. Our data indicate that they don't. The question remains whether people will obey the screening-off principle in situations in which their causal expectations are not so starkly contradicted. Some other evidence suggests that they won't. In studies examining people's causal models of category knowledge, both Chaigneau, Barsalou, and Sloman (2004) and Rehder and Burnett (2005) found small violations of screening-off. Chaigneau et al.'s violations involved causal chains. Rehder and Burnett's, like ours, concerned common cause models. Rehder and Burnett's response to these violations, again like ours, is to suggest that people are reasoning in terms of a different causal model than the experimenters initially attributed to them. All of these studies test adults who enter the laboratory with a lot of causal knowledge. That fact, combined with people's sophisticated ability to generate explanations, means that pinning down the causal model that people are actually reasoning with in any given situation is far from trivial. Subsequent tests of screening off need to study models consisting of causal beliefs that are uncontroversial and not easily revisable.

Although the flexibility of people's causal models makes testing the psychological reality of screening off harder, it doesn't make screening off immune to empirical validation. Any principle that's consistently violated in the laboratory cannot serve as a useful description of human cognition. We feel confident that Gordon Bower would agree. And that's good enough for us.

ACKNOWLEDGMENTS

This research was supported by NASA Grant NCC2–1217 to Steven Sloman. Experiments 1 and 2 were reported at the 2004 meeting of the Cognitive Science Society and the 2004 International Conference on Thinking.

REFERENCES

Byrne, R. M. J., & Walsh C. R. (2002). Contradictions and counterfactuals: Generating belief revisions in conditional inference. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 160–165). Mahwah, NJ: Lawrence Erlbaum Associates.

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General, 133,* 601–625.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128,* 978–996.

Gardenfors, P. (1988). *Knowledge in flux.* Cambridge, MA: MIT Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Harman, G. (1986). *Change in view.* Cambridge, MA: MIT Press.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109.

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 200–219). Oxford, England: Blackwell.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50,* 264–314.

Shaklee, H., & Fischhoff, B. (1982). Strategies in information search in causal analysis. *Memory & Cognition, 10,* 520–530.

Sloman, S.A. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Walsh, C. R., & Johnson-Laird, P. N. (2005). *Changing your mind.* Manuscript submitted for publication.

Walsh, C. R., & Sloman, S. A. (2004). Revising causal beliefs. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1423–1427). Mahwah, NJ: Lawrence Erlbaum Associates.