Journal of Experimental Psychology: Learning, Memory, and Cognition

# The Impact of Disablers on Predictive Inference

Denise Dellarosa Cummins

Online First Publication, June 9, 2014. http://dx.doi.org/10.1037/xlm0000024

CITATION

Cummins, D. D. (2014, June 9). The Impact of Disablers on Predictive Inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000024

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000024

# The Impact of Disablers on Predictive Inference

Denise Dellarosa Cummins
University of Illinois at Urbana-Champaign

People consider alternative causes when deciding whether a cause is responsible for an effect (diagnostic inference) but appear to neglect them when deciding whether an effect will occur (predictive inference). Five experiments were conducted to test a 2-part explanation of this phenomenon: namely, (a) that people interpret standard predictive queries as requests to estimate the probability of the effect in the presence of the cause alone, which renders alternative causes irrelevant, and (b) that the impact of disablers (inhibitory causes) on predictive judgments is underestimated, and this underestimation is wrongly interpreted as cause neglect. Experiment 1 showed that standard predictive queries are frequently interpreted as requests to estimate the likelihood of E given C alone. In Experiment 2, a causal Bayes network overestimated predictive inference when it was queried in the standard way, but this overestimation diminished when predictive inference was queried using an alternative wording. In Experiment 3, participants judged alternative causes to be relevant to diagnostic inference and both disablers and alternative causes to be relevant to predictive inference. In Experiment 4, disablers greatly overshadowed alternative causes in predictive judgments, but their impact on diagnostic judgments was negligible. In Experiment 5, the order of disabler retrieval influenced causal judgments. Taken together, these results indicate that human causal inference cannot be adequately modeled unless the manner in which knowledge is retrieved and applied is taken into consideration.

*Keywords:* causal inference, predictive inference, diagnostic inference, disablers, causal Bayes network

Causal inference is a fundamental and ubiquitous component of cognition. We use it continually to make sense of events in our lives, such as figuring out why our car did not start, or deciding whether we will lose weight by adopting a new exercise regimen. It constitutes the foundation of cognition and perception, binding together our conceptual categories, imposing structures on perceived events, and guiding our decision making.

Reasoning from cause to effect (as in deciding whether exercise leads to weight loss) is called *predictive inference.* Reasoning from effect to cause (as in deciding whether one lost weight because one exercised) is called *diagnostic inference.* Normatively, alternative causes play a crucial role in both types of judgment. The greater the number, strength, and probability of alternatives causes for an effect, the more likely it is to occur, and the more likely it is that another cause may have been responsible for the effect rather than the one under consideration. Weight loss can be caused by many things other than exercising; hence, they should be considered when determining the likelihood that a person will lose weight even if the only thing we know for sure is that the person is exercising.

Recently, Fernbach and colleagues (Fernbach, Darlow, & Sloman, 2010, 2011a, 2011b) have shown that diagnostic inference is well captured by a normative probabilistic model that includes explicit parameters for the strength of a named cause and alternative causes, but predictive inference is best captured by a model that includes only the strength of the named cause. On the basis of this, they have argued that people do indeed consider alternative causes when making diagnostic judgments but routinely fail to consider such causes when making predictive judgments. They consider alternative cause neglect to be a bias in human predictive inference.

The goal in the work presented here is to test two alternative explanations of this apparent alternative cause neglect in predictive inference. The first is a pragmatic explanation: People frequently interpret standard predictive queries as requests to estimate the probability that the cause alone can bring about the effect rather than the probability that the cause and/or others may bring about the effect. Because this common interpretation renders alternative causes irrelevant, people do not seek or consider them. The implication of this explanation is that when predictive inference is properly queried, consideration of alternative causes will be found to impact it.

The second explanation is that the impact of disablers on predictive judgments is underestimated by probabilistic models of causal inference, and this underestimation is wrongly interpreted as alternative cause neglect. Disablers are factors that can prevent an effect from occurring in the presence of a true cause, such as icy streets disabling a brake's power to bring a moving car to a halt or improper diet negating the impact of strong exercise on weight losss (Cummins, 1995, 1997, 2010; Cummins, Lubart, Alksnis, & Rist, 1991). Cheng (1997) referred to these factors as *preventive causes.*

Denise Dellarosa Cummins is retired from the Department of Psychology, University of Illinois at Urbana-Champaign.

I thank Mike Oaksford for insightful comments on a previous version of this article.

Correspondence concerning this article should be addressed to Denise Dellarosa Cummins. E-mail: denise.cummins87@gmail.com

## The Confusability of Causal Power and Predictive Judgment Queries

Causal power is distinct from predictive inference. Causal power refers to the ability of a particular cause (when it is present) to elicit an effect, relative to other causes (Cheng, 1997). Predictive inference refers to the likelihood of an effect occurring, given what is known about the causes that can bring it about. With reference to the exercise example, a causal power judgment is an estimate of the likelihood that a new exercise regimen alone can bring about weight loss. In contrast, a predictive judgment is an estimate that weight loss will occur, given that we know for certain the person has adopted the new exercise regimen. When predicting whether the person will lose weight, decision makers should consider not just the stated cause (e.g., the new exercise regimen) but other causes that can bring the effect about as well (e.g., dieting, illness).

Despite this crucial distinction, these two inferences are typically queried in the following way:

> *Causal power query:* C occurred. How likely is it that C occurring causes E to occur?
>
> *Predictive inference query:* C occurred. How likely is that E occurs?

As is apparent, these queries are readily confusable. Both queries can be readily interpreted to mean "estimate the probability that the effect occurs in the presence of this cause alone." This means that both queries may be readily interpreted as requests to estimate the power of the stated cause to bring about the effect. If this is the case, people are not being irrational when they give the same estimate to what appear to be two identical queries.

Fernbach et al. (2010, 2011b) referred to this matter as the "pragmatic explanation of alternative cause neglect." They investigated this explanation using a variety of manipulations intended to cue participants into considering alternative causes. None of these manipulations had appreciable effects. The authors interpreted these null effects to mean that alternative neglect is not due to pragmatic considerations but instead constitutes a genuine bias in the reasoning process.

The difficulty with this interpretation is that all of their modified problem formats still employed easily confusable predictive and causal power queries like those shown above. One could argue, therefore, that these manipulations failed to produce a difference because predictive inference was still queried in the standard, confusable way. For example, Fernbach et al. (2010; Experiment 1) told participants that a patient had been diagnosed with depression and asked them to judge the likelihood that she would present with lethargy. This is readily interpreted as a request to estimate the causal power of depression alone to elicit lethargy.

It is important to recognize that if people interpret causal power and predictive inference queries this way, they are not committing errors when they return the same values for each. Instead, they are being consistent in their responses, because they believe they have been asked to give the same judgment twice. We cannot fault people for ignoring alternatives if we have posed the problem in a way that invites them to do so. Another way to put this is that what appears to be the wrong answer to a problem may turn out to be the right answer to the problem as construed by the problem solver. For example, when young children are given problems such as

> Mary and John have 5 marbles altogether.
>
> Mary has 2 marbles.
>
> How many does John have?

they frequently return the first number (5) as the answer because they frequently misinterpret "altogether" to mean "each" (Cummins, 1991; Cummins, Kintsch, Reusser, & Weimer, 1988). From their perspective, this problem expresses a contradiction between the first and second lines, but the answer to the third line is surely "5." Similarly, they frequently interpret statements such as "Mary has 5 more marbles than John" as "Mary has 5 marbles. The quantity of John's marbles is unknown." Problem-solving performance can be improved simply by clarifying the meaning of the problem text, rather than remediating mathematical or logical thinking skills (Cummins, 1991). Moreover, common arithmetic word problem errors can be better simulated using deficient text comprehension strategies than deficient logico-mathematical strategies (Dellarosa, 1986). Similar arguments concerning the necessity of investigating how participants interpret decision-making queries in experiments have been put forth by Oaksford and Chater (1994) and McKenzie and Mikkelsen (2007).

Testing this pragmatic explanation of alternative neglect requires examining (a) how people typically interpret standard predictive queries, (b) which types of information they consider relevant to this type of judgment, and (c) whether changing the problem format in ways that avoid using standard predictive queries improves performance.

## Underestimating the Impact of Disablers on Predictive Inference

Causal Bayes networks are the dominant normative models of causal inference, and human predictive reasoning has been termed biased because it departs from modeled likelihood estimates. For this reason, it is important to appreciate how causal Bayes nets model causal decision making.

The underlying assumption of causal Bayes nets is that our beliefs are represented probabilistically (rather than as simply true or false) and that inferences based on them can be captured by Bayesian reasoning (Glymour, 1998, 2001; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005, 2009; Lagnado & Sloman, 2004; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Pearl, 1988, 2000; Rehder, 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann & Martignon, 1998).

A causal Bayes net consists of (a) a set of nodes representing variables, (b) arrows indicating the direction of causal connections among them, and (c) functions that define the probability distribution for each variable. Causes can be generative or inhibitory. Generative causes increase the probability of their effects, while inhibitory causes decrease their probability. Nodes that have output nodes but no input nodes (parent nodes) in the graph are called exogenous nodes. Nodes that have both parent nodes and output nodes are called endogenous nodes. Functions for endogenous variables are conditionalized on their parent nodes. The conditional probability distribution associated with a node can be any probabilistically sound function of its parents. Assigning a probability function to a node is called parameterization. When modeling causal learning, these parameters are estimated from observed contingency data that show the frequency with which causes and effects occur (or fail to occur) together. When modeling causal inference, the node parameters are given, and the conditional likelihood functions are inferred.

Figure 1 shows a simple graph for generative causation. The graph shows that the effect occurs probabilistically in its causal background (alternative causes), and the addition of the cause of interest increases the likelihood of its occurrence. $W_c$ is a parameter that is associated with the strength of the cause to elicit the effect, and $W_a$ is a parameter associated with the strength of the causal background to elicit the effect. If these variables independently contribute to generating the effect, then the probability of the effect can be captured by the following noisy-OR function (Pearl, 1988; Tenenbaum & Griffiths, 2001):

$$\text{CBN-P} = P(e|c, a; W_c, W_a) = W_c + W_a - W_cW_a. \tag{1}$$

If the weights are equal to 1, the function is simply the function for inclusive-OR. (Inclusive-OR simply means, e.g., “soup or salad, or both.”) When the weights range between 0 and 1 (as in the case of probabilities), the function is termed noisy-OR. It constitutes a measure of the strength of a causal relationship, assuming that a causal relationship exists (Griffiths & Tenenbaum, 2005). Put simply, this equation states that the likelihood of the effect is equal to the causal power of the cause plus the strength of alternative causes, minus their intersection.

I refer to the model above as *causal Bayes network–predictive*, or CBN-P, because Fernbach et al. (2011a) proposed that predictive inference could be modeled this way under the following assumptions: First, events are binary (either they happen or they do not); this allows the probability distribution for exogenous nodes to be parameterized as their prior probabilities. For the cause of interest, this is $P_c$. In the case of predictive inference (where the cause is explicitly stated to have occurred), $P_c = 1$. Second, the prior probabilities and strength of alternative causes are collapsed into a single parameter, P(Effect|~Cause). In other words, alternatives are always present (their priors are equal to 1), but they elicit the effect probabilistically. Third, the cause of interest and the causal background are assumed to contribute independently to the effect. Fourth, causal power, $W_c$, is parameterized as P(Effect|Cause, ~Effective Alternatives), which corresponds to the maximum likelihood estimate of $W_c$ (Cheng, 1997; Glymour, 1998; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001). Finally, enablers and disablers are presumed to be implicit in $W_c$ and $W_a$. Enablers are factors that must be present for a cause to be effective, as in the case of oxygen for combustion. Disablers are factors that can prevent an effect from occurring in the presence of a generative cause (Cummins, 1995, 1997, 2010; Cummins et al., 1991). This means that causal power is parameterized as $W_c$ = P(Effect|Cause, *Disablers*, ~Effective Alternatives). Notice that a causal power value $< 1$ implies that disablers occur probabilistically to inhibit the effect.

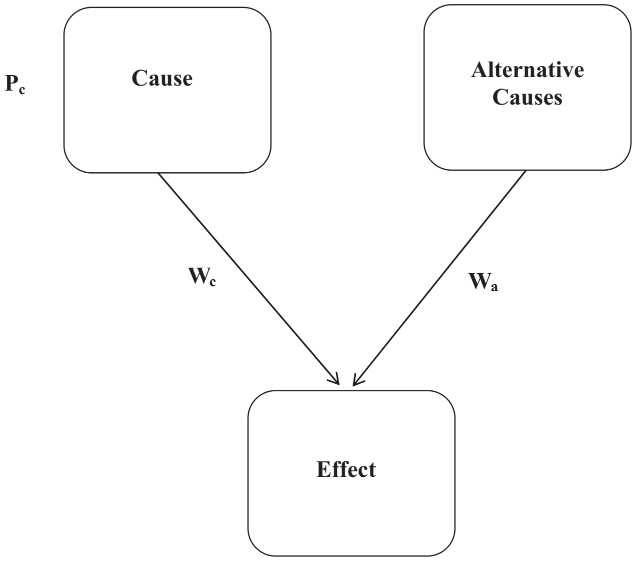

*Figure 1.* A Bayes net model of generative causation. $P_c$ represents the prior probability of the cause, $W_c$ represents the power of the cause to elicit the effect, and $W_a$ is the strength of alternatives. $W_a$ is an aggregate of causal power and prior probabilities of all alternative causes. The effect is generated by a noisy-OR function of the cause and alternatives. Adapted from “Neglect of Alternative Causes in Predictive but Not Diagnostic Reasoning,” by P. M. Fernbach, A. Darlow, and S. A. Sloman, 2010, *Psychological Science, 21,* p. 5. Copyright 2010 by the Association for Psychological Science. Adapted with permission.

Using the weight loss example, let us assume that the decision maker is told that the person is exercising ($P_c = 1$), the causal power of exercise to yield weight loss is estimated to be $W_c = .8$, and the combined causal power of alternatives to yield weight loss is $W_a = .5$. Then the probability that a person will lose weight is $.8 + .5 - .8 * .5 = .9$.

Diagnostic inference (CBN-D) is modeled by working backward through the net from the effect to its causes. Using Bayes theorem, the noisy function for diagnostic inference is as follows (Fernbach et al., 2011a or Waldmann, Cheng, Hagmayer, & Blaisdell, 2008):

$$\begin{aligned}\text{CBN-D} &= P(c|e, a; W_c, W_a) = 1 - P(\sim\text{Cause}|\text{Effect}) \\ &= 1 - \frac{(1 - P_c)W_a}{P_cW_c + W_a - P_cW_cW_a}.\end{aligned} \tag{2}$$

The denominator of this expression is P(Effect), which is captured by a noisy-OR function of the prior probability of the cause, the causal power of the cause, and the strength of alternatives. The numerator is the conjunction of the probability that the cause did not occur and the probability of the effect without the cause. The ratio of these two constitutes the probability P(~Cause|Effect).

Put simply, this expression says that the effect occurred, and the probability that the cause occurred is equal to 1 minus the probability of the effect occurring in the absence of the cause. As Fernbach et al. (2011a) pointed out, the presence of the effect cannot decrease the probability of the cause (given the nature of causation), so the value of CBN-D will always be higher than $P_c$ and will increase as $P_c$ increases. The value of CBN-D also decreases with the probability that alternative causes elicited the effect and hence constitutes the strength of evidence that the candidate cause was indeed responsible for the effect. Using again the example of weight loss, if we further assume that the decision maker believes the prior probability of weight loss is .7, then the probability that it was exercise—rather than other factors—that caused a person’s weight loss is $1 - [((1 - .7) * .5)/((.7 * .5) + .5 - (.7 * .8 * .5))] = .74$.

To summarize, according to Fernbach and colleagues (Fernbach et al., 2010, 2011a, 2011b), normative predictive inference can be calculated using the power of the cause to elicit the effect ($W_c$) and the strength of alternatives ($W_a$). The prior probability of the cause is equal to 1 because it is known to have occurred, and the prior probabilities of the alternative causes are set to 1 on the assumption that they are always part of the causal background. The value of $W_c$ and $W_a$ are also a function of disablers. Normative diagnostic inference can be calculated using these same factors and to the prior probability of the cause ($P_c$).

When fitted to observed judgment data, CBN-D has been found to be an excellent model of diagnostic judgments, while CBN-P provides a poor fit to predictive judgment data (e.g., Fernbach et al., 2010, 2011a, 2011b). Instead, the best fitting model for predictive judgments frequently turns out to be $W_c$, the estimate of causal power. Fernbach and colleagues interpreted this pattern of results to mean that people consider alternative causes when making diagnostic judgments but neglect to consider them when making predictive judgments.

The second alternative explanation proposed here is that the model underestimates the impact of disablers on predictive inference. Ample research based on causal conditional reasoning shows that different types of knowledge are activated when reasoning from cause to effect than when reasoning from effect to cause. Specifically, when reasoning from cause to effect, disablers are spontaneously activated; when reasoning from effect to cause, alternative causes are spontaneously activated.

For example, acceptance rates for "cause therefore effect" conditional arguments such as "If Susan exercises hard, then she will lose weight/Susan has been exercising hard/Therefore, Susan will lose weight" are inversely proportional to the number of disablers activated in memory (factors that could prevent weight loss even though a person exercises hard). In contrast, acceptance rates for "effect therefore cause" conditional arguments such as "If Susan exercises hard, then she will lose weight/Susan lost weight/Therefore, Susan has been exercising hard" are inversely proportional to the number of alternative causes for weight loss that are activated in memory. This pattern of results has been observed in adults (e.g., Chan & Chua, 1994; Cummins, 1995, 1997; Cummins et al., 1991; De Neys, Schaeken, & d'Ydewalle, 2002, 2003; Elio, 1998; Fernbach & Erb, 2013; Markovits & Potvin, 2001; Quinn & Markovits, 1998; Verschueren, Schacken, De Neys, & Y'dewalle, 2004) as well as children (Janveau-Brennan & Markovits, 1999; Markovits, 2000; Markovits, Fleury, Quinn, & Venet, 1998). Moreover, Fernbach and Erb (2013) reported that acceptance rates for "cause therefore effect" arguments were best predicted by causal power alone, which in turn was a function of disabler strengths and disabler base rates. In contrast, acceptance rates for "effect therefore cause" arguments were best captured by the CBN-D model shown above.

This pattern of results is directly relevant to descriptive models of predictive and diagnostic probabilistic decision making. Predictive queries require reasoning from cause to effect (e.g., the likelihood that exercising will yield weight loss). The causal argument literature suggests that such a scenario should spontaneously activate disablers that could prevent the effect from occurring. This means that predictive inferences are made within a cognitive context heavily populated by spontaneously retrieved disablers. Alternative causes, even if they are activated, are likely to be overshadowed by disablers during the inference process. Because disablers outnumber and/or outweigh alternative causes during predictive inference, the resulting likelihood estimated will be underestimated relative to normative estimates.

Diagnostic queries, on the other hand, ask the decision maker to estimate the likelihood that a particular cause was responsible for an effect (e.g., the likelihood that exercise was responsible for a person's weight loss). This scenario should spontaneously activate alternative causes for the effect, meaning that probabilistic diagnostic judgments are made within a cognitive context that is heavily populated by spontaneously retrieved alternative causes. Disablers, even if they are activated, are likely to be overshadowed by alternative causes. It should come as no surprise, therefore, that alternative cause strength impacts diagnostic likelihood judgments.

The greater implication of this analysis is that the parameters for alternative causes, disablers, and perhaps even causal power are not static; the weight assigned to each may depend on the type of inference requested. Diagnostic queries may activate alternative causes, thereby increasing the weight attributed to alternative strength. Predictive queries may activate disablers, thereby diminishing the contribution of alternative strength to the resulting judgment.

## Overview of Research

Five experiments were conducted to test this two-part explanation of alternative neglect bias. Experiment 1 provided decision makers the opportunity to explicitly and directly indicate how they interpreted standard predictive, diagnostic, and causal power queries by selecting which alternative rewording most closely matched their understanding of what they were being asked to do. If it is the case that causal power and predictive inference queries are interpreted in the same way, the same choice should be the preferred selection for both. Experiment 2 allowed direct testing of CBN-P using standard and modified predictive queries. If alternative neglect is due to ambiguity in standard predictive queries, removal of the ambiguity should bring predictive inference estimates more in line with normative estimates. Experiment 3 provided participants the opportunity to explicitly choose which type of information they believed was relevant to predictive and diagnostic judgments: alternative causes, disablers, both, or neither. If people do not believe alternative causes are relevant to predictive inference, this should be reflected in their choices. Experiment 4 allowed direct assessment of the degree of impact exacted by disablers and alternatives on predictive and diagnostic judgments. Finally, Experiment 5 afforded participants the opportunity to explicitly list the disablers considered when making predictive inferences and to estimate their prior probabilities and strengths. This allowed comparison of competing models that purport to capture causal power, an important component of predictive and diagnostic inference.

## Experiment 1

The purpose in Experiment 1 was to test the first alternative explanation for apparent alternative cause neglect in predictive inference, namely, that the way predictive judgments are queried invites alternative neglect. To test this, people were provided the opportunity to select the statement that best reflected their understanding of standard predictive and diagnostic queries. Predictive inference was queried in the standard way: *C occurred. How likely is it that E will occur?* Participants were asked to indicate what they believed the query was asking them to do by selecting from among three alternative interpretations: (a) the likelihood that C can cause E, (b) the likelihood of E given C alone, and (c) the likelihood of E given C and/or other causes. The normative interpretation is (c); this is what noisy-OR means. Note that (a) is a causal power query, but it differs from the way such queries are typically worded (i.e., *How likely is it that C occurring causes E to occur?*). The term *can cause* was used, as it more strongly implies that the cause has the power to bring about the effect. Likelihood estimates were then obtained for each choice option.

In the diagnostic condition, people were given the opportunity to indicate how they interpreted the standard diagnostic query: *E occurred. How likely is it that C occurred?* Their choices were (a) the likelihood that C can cause E, (b) the likelihood that the effect is due to C as opposed to other causes, and (c) the likelihood that the effect is due to C and/or other causes. The normative interpretation is (b).

## Method

**Participants.** One hundred Amazon Mechanical Turk workers served as participants in the experiment. Participation was restricted to fluent speakers of English, residing in the United States, who had previously completed over 50 Amazon tasks and whose task approval rate was greater than 80%. Six were eliminated for failing to complete the task. Of the remaining 94, 33.0% ($n = 31$) were female. Age categories (and sample sizes) were as follows: 18–25 ($n = 46$), 26–35 ($n = 36$), 36–45 ($n = 8$), 46–55 ($n = 2$), 56 and older ($n = 2$). They were paid $.25 for their participation.

**Design.** The experiment employed two between-subject variables and one repeated measure. The between-subject variables were judgment type (predictive or diagnostic), and causal stories chosen from Fernbach et al. (2010; using Rosetta Stone to learn French, using a nicotine patch to quit smoking, or exercising hard to lose weight). The repeated measures variable was query (interpretation, casual strength, cause alone/opposed to others, and cause and/or others).

**Materials and procedure.** Participants were asked four questions about a single causal scenario. In the predictive judgment condition, they were asked an interpretation question, a causal power question and two modified predictive judgment questions. Using the nicotine patch scenario as an example, the questions were as follows:

> *Imagine you wear a nicotine patch to quit smoking. How likely is it that you will quit smoking?*
>
> *Interpretation question: What is this question asking you to do?*
>
> *a. Judge how likely it is that wearing a nicotine patch can cause you to quit smoking.*
>
> *b. Judge how likely it is that you will quit smoking if you wear a nicotine patch and do nothing else.*
>
> *c. Judge how likely it is that you will quit smoking if you wear a nicotine patch and/or do other things that can cause you to quit smoking.*
>
> *Causal power estimate: How likely is it that wearing a nicotine patch* can *cause you to quit smoking?*
>
> *Cause alone estimate: How likely is it that you will quit smoking if you wear a nicotine patch and do nothing else?*
>
> *Cause and/or alternatives estimate: How likely is it that you will quit smoking if you wear a nicotine patch and/or do other things that can cause you to quit smoking?*

As in Fernbach et al. (2010), the last three questions were accompanied by a 0–100 likelihood scale, marked in 10-point increments (0, 10, 20 . . .). The word Impossible appeared above 0, and the word Definite appeared above 100.

In the diagnostic judgment condition, the questions were as follows:

> *Imagine you quit smoking. How likely is it that you wore a nicotine patch?*
>
> *Interpretation question: What is this question asking you to do?*
>
> *a. Judge how likely it is that wearing a nicotine patch* can *cause you to quit smoking.*
>
> *b. Judge how likely it is that your quitting smoking is due to wearing a nicotine patch as opposed to other possible causes.*
>
> *c. Judge how likely it is that your quitting smoking is due to wearing a nicotine patch and/or to other possible causes?*
>
> *Causal power estimate: How likely is it that wearing a nicotine patch* can *cause you to quit smoking?*
>
> *Cause opposed to alternatives estimate: How likely is it that your quitting smoking is due to wearing a nicotine patch as opposed to other possible causes?*
>
> *Cause and/or alternatives estimate: How likely is it that your quitting smoking is due to wearing a nicotine patch and/or doing other things that can cause you to quit smoking?*

The same likelihood scale appeared beneath each of the last three questions. The questions were presented without the labels shown above. Participants accessed the study materials through a Survey Monkey link posted on Amazon Mechanical Turk. The first page was the consent form. The next page presented the interpretation question. Each participant answered either a predictive or a diagnostic question about a single causal story. Story was randomized across participants, as was the order of interpretation choices. The remaining three questions appeared on the following page in random order. The next page requested sex and age information. The last page provided debriefing information. Participants could not go back to previous pages.

## Results and Discussion

All $t$ tests are two-tailed. All $\chi^2_{(1)}$ tests include correction for continuity.

**Interpretation question.** Because story did not contribute significant variability, it was collapsed in subsequent analyses (see the Appendix). The frequency with which each option was selected is presented in Table 1. The first row corresponds to the causal power ("can cause") question. The second row corresponds to the question of evaluating the impact of the cause alone ("the cause alone" for predictive inference and "the cause as opposed to other causes" for diagnostic inference). The third row corresponds to the impact of the cause in consideration with other causes ("and/or").

A statistical dependency was observed between judgment type and question type, $\chi^2_{(2)} = 9.95$, $p < .01$, Cramer's $V = .32$. As predicted, two interpretations of the predictive query were equally likely, (a) "estimate the likelihood of the effect given the occurrence of the cause alone" ($n = 19$) and (b) "estimate causal strength" ($n = 20$), $\chi^2_{(1)} < 1$. Rarely did participants interpret the predictive judgment question as a request to consider the cited cause and/or alternatives ($n = 7$).

The most frequent interpretation of the diagnostic query was as a request to estimate the impact of "this cause as opposed to alternatives" ($n = 27$), which was almost twice as frequent as "this cause and/or alternatives" ($n = 14$); this difference was marginally significant, $\chi^2_{(1)} = 3.52$, $p = .06$. Rarely did participants interpret diagnostic queries as requests to estimate causal power ($n = 7$).

**Likelihood estimates.** Estimates for predictive and diagnostic judgments were analyzed separately via analyses of variance (ANOVA) using story (French, nicotine, or exercise) as a between-subject variable and question type as repeated measures. For predictive judgments, the question types were causal power, cause alone, and cause and/or alternatives. For diagnostic judgments, the question types were causal power, cause opposed to alternatives, and cause and/or alternatives. For diagnostic judgments, the main effect of story was not significant, nor did it interact with question type, Fs $< 1$. For predictive judgments, the main effect of story was significant, $F(2, 43) = 7.31$, $MSE = 405.61$, $p < .01$, but it did not interact with question type ($F < 1$). Because story did not interact with question type for either type of judgment, this variable was collapsed in subsequent analyses. The means and standard errors are presented in Table 2.

For predictive judgments, the main effect of question type was significant, $F(2, 90) = 6.74$, $MSE = 263.48$, $p < .002$, $\eta^2 = .13$. As predicted, a planned comparison indicated that participants estimated the likelihood of the effect given the cause and/or alternatives ($M = 70.6$) to be greater than the likelihood of the effect given the cause alone ($M = 58.4$), $t(45) = 3.24$, $p < .002$, Cohen's $d = .96$, Cohen's $r = .43$. For diagnostic judgments, the main effect of question type also was significant, $F(2, 94) = 8.50$, $MSE = 339.97$, $p < .0001$, $\eta^2 = .15$. A planned comparison indicated that participants estimated the likelihood that the effect was due to the cause and/or alternatives ($M = 73.1$) to be greater than the likelihood that it was due to the cause as opposed to alternatives ($M = 58.5$), $t(47) = 3.71$, $p < .001$, Cohen's $d = 1.08$, Cohen's $r = .47$.

Table 1
*Frequency of Interpretation Selection for Predictive and Diagnostic Queries in Experiment 1*

| Predictive[a] | $n$ | Diagnostic[b] | $n$ |
|---|---|---|---|
| C can cause E | 20 | C can cause E | 7 |
| C alone caused E | 19 | C caused E as opposed to alternatives | 27 |
| C and/or alternatives caused E | 7 | C and/or alternatives caused E | 14 |

[a] $n = 46$. [b] $n = 48$.

Table 2
*Likelihood Estimates for Predictive and Diagnostic Judgments in Experiment 1*

| Predictive[a] | Likelihood | Diagnostic[b] | Likelihood |
|---|---|---|---|
| "Can cause" | 66.7 (2.8) | "Can cause" | 70.4 (2.9) |
| Cause alone | 58.4 (2.7) | Cause opposed to alternatives | 58.5 (2.7) |
| Cause and/or alternatives | 70.6 (2.7) | Cause and/or alternatives | 73.1 (3.2) |

*Note.* Numbers in parentheses are standard errors of the mean.
[a] $n = 46$. [b] $n = 48$.

A planned contrast was also conducted to compare the impact of alternatives on predictive and diagnostic judgments. The *and/or* queries invite consideration of alternative causes, while the *cause alone* and *cause as opposed to* queries invite ignoring alternative causes. The mean difference between *and/or* and *alone* in the predictive condition ($M = 12.2$) was not statistically different from the mean difference between *and/or* and *opposed to* in the diagnostic condition ($M = 14.6$), $t(92) = 0.43$, $p = .67$. This suggests that the impact of alternative causes on predictive judgment was equivalent to their impact on diagnostic judgment when the queries made explicit what was being requested.

Turning finally to *can cause* (causal power) judgments, participants estimated the likelihood that the cause could bring about the effect to be significantly greater than the likelihood that the cause alone would actually bring about the effect, and it did not matter whether they made a predictive or diagnostic judgment: predictive condition mean difference $= 8.2$, $t(45) = 2.72$, $p < .01$, Cohen's $d = .81$, Cohen's $r = .38$; diagnostic mean difference $= 11.8$, $t(47) = 3.71$, $p < .001$, Cohen's $d = 1.08$, Cohen's $r = .48$. Moreover, mean causal power estimates in the diagnostic judgment condition ($M = 70.4$) were not statistically different from those in the predictive condition ($M = 66.7$), $t(92) < 1$.

**Summary.** The results clearly show that people frequently misinterpret standard predictive queries as requests to estimate the likelihood that the cause alone will or can bring about the effect in question. When predictive queries are worded in a way that more accurately maps on the meaning of the noisy-OR function (C and/or alternatives), alternative causes have equivalent influence on predictive and diagnostic judgments. These results constitute a rational decision profile in that P(Effect/Cause, Alternatives) > P(Effect/Cause, No Alternatives) and P(Cause/Effect, No Alternatives) > P(Cause/Effect, Alternatives). Another way of interpreting this result is that, when queries are worded this way, participants provided an estimate that a causal mechanism was in play, as opposed to simple contingency.

Finally, when causal power queries are worded in a way that invites the decision maker to estimate the likelihood that the cause can bring about the effect, the decision maker estimates this to be higher than the simple likelihood that the cause will (or did) bring about the effect. These results have two implications. First, they imply that causal power estimates are stable across inference types; that is, the parameter is the same for diagnostic and predictive inference. Second, they imply that when causal power queries are worded as "can cause," participants interpret them to mean "in the absence of disablers." If this were the case, their judgments again constitute rational decision profiles in that P(Effect/Cause, No Alternatives, No Disablers) > P(Effect/Cause, No Alternatives, Disablers).

## Experiment 2

In Experiment 2, participants again made diagnostic and predictive judgments, and predictive judgments were queried using standard and modified "and/or" wording. They also gave estimates for causal power, the prior probability of the cited cause, and the strength of alternative causes. These values were used as parameter estimates for $W_c$, $P_{c,}$ and $W_a$, respectively. Model predictions were calculated for predictive and diagnostic judgments using CBN-P and CBN-D, respectively. As mentioned in the introduction, alternative cause neglect in predictive inference has been inferred from the fact that observed predictive judgment estimates are typically significantly lower than model estimates. Moreover, predictive judgments are typically best captured by the single causal power parameter, $W_c$; the strength of alternatives, $W_a$, accounts for negligible variance.

In Experiment 1, we found that participants typically interpret standard predictive queries to mean "estimate the likelihood of the effect given the cause alone." This interpretation precludes consideration of alternative causes. When the query was modified to map more accurately onto the meaning intended by researchers (i.e., noisy-OR) by simply including the term *and/or,* predictive estimates rose significantly. If the difference between observed and normative predictive estimates is due to these pragmatic considerations rather than a reasoning bias, altering the query such that it maps directly onto noisy-OR interpretation should narrow or eliminate the difference between observed and model estimates.

### Method

**Participants.** One hundred twenty Amazon Mechanical Turk workers served as participants. Participation was restricted to fluent speakers of English, residing in the United States, who had previously completed over 50 Amazon tasks and whose task approval rate was greater than 80%. Six were eliminated for failing to complete the task. Of the remaining 114 participants, 35.1% ($n$ = 40) were female. Age distribution was as follows: 18–25 ($n$ = 52), 26–35 ($n$ = 45), 36–45 ($n$ = 9), 46–55 ($n$ = 4), 56 and older ($n$ = 4). Participants were paid $.25.

**Design.** There was one between-subject variable and six repeated measures. The between-subject variable was causal relationship (using Rosetta Stone to learn French, using a nicotine patch to quit smoking, or exercising to lose weight). The repeated measures were the parameter estimates and judgment queries used by Fernbach et al. (2011a) along with the additional modified and/or predictive query, which are shown in Table 3.

**Materials and procedure.** Each participant read only one story and answered the six questions shown in Table 3. Five of the six questions are identical to those used by Fernbach et al. (2011a), and the sixth constituted a modified and/or predictive. Only one question appeared on the screen at a time, and participants could not go back to previous screens. As in Fernbach et al. (2011a), they recorded their answers by selecting a number from a 0–100 likelihood scale, where 0 was labeled "Impossible" and 100 was labeled "Definite," and numbers increased in units of 10. Causal story and question order were randomized across participants.

Table 3
*Example of Questions Asked in Experiment 2*

| Judgment | Example[a] |
|---|---|
| Causal power ($W_c$) | Imagine John wears a nicotine patch. How likely is it that wearing a nicotine patch causes John to quit smoking? |
| Prior probability of cause ($P_c$) | How likely is it that John wears a nicotine patch? |
| Standard predictive | Imagine John begins wearing a nicotine patch. How likely is it that he will quit smoking? |
| And/or predictive[b] | Imagine John begins wearing a nicotine patch and/or doing other things to help him quit smoking. How likely is it that he will quit smoking? |
| Diagnostic | Imagine John quit smoking. How likely do you think it is that he began wearing a nicotine patch? |
| Strength of alternative causes ($W_a$) | Imagine John quit smoking. How likely do you think it is that he did NOT wear a nicotine patch? |

[a] In this example, each question was preceded by the following statement: "John is a smoker between the ages of 18 and 25." [b] Modified predictive query used in Experiment 2. All other questions are identical to those used by Fernbach et al. (2011a).

### Results and Discussion

Because the results of preliminary analyses indicated that the story variable did not interact with the variables of interest, subsequent analyses collapsed over the story variable. All $t$ tests are two-tailed. Mean likelihood estimates for the six questions are shown in Table 4.

**Predictive likelihood.** Mean likelihood estimates for the and/or predictive query ($M$ = 70.2) exceeded those for the standard predictive query ($M$ = 62.5), $t(113) = 4.69$, $p < .001$, Cohen's $d$ = .88, Cohen's $r$ = .40. This is consistent with the claim that people believe standard predictive queries are requests to estimate P(Effect|Cause, ~Alternatives) and that using the colloquial *and/or* expression instead makes it clear that they are to estimate P(Effect|Cause, Alternatives).

CBN-P was fit to observed predictive likelihood in the same way done by Fernbach et al. (2011a, Experiment 1). The results of these model-fitting efforts are presented in Table 5.

When the predictive query was worded in the standard way, the CBN-P model overestimated observed predictive estimates, $t(113) = -11.44$, $p = .001$, Cohen's $d = 2.15$, Cohen's $r = .73$. In fact, standard predictive judgments were best captured by a model containing causal power only, $t(113) = -0.49$, $p = .63$. This pattern of results replicates those reported by Fernbach et al. (2010). It is consistent with the supposition that people frequently believe that causal power queries and standard predictive queries are simply two ways of requesting the same likelihood estimate.

As predicted, however, using the *and/or* query significantly raised predictive likelihood estimates over causal power estimates,

Table 4
*Mean Likelihood Estimates for Judgments in Experiment 2*

| *n* | $W_c$ | $P_c$ | $W_a$ | Standard predictive | And/or predictive | Diagnostic |
|---|---|---|---|---|---|---|
| 114 | 63.1 (1.8) | 42.8 (2.3) | 42.9 (2.1) | 62.5 (1.7) | 70.2 (1.6) | 54.3 (2.2) |

*Note.* Numbers in parentheses are standard errors of the mean. $W_c$ = causal power; $P_c$ = prior probability of alternatives; $W_a$ = alternatives strength.

$t(113) = 4.69$, $p < .001$, Cohen's $d = .88$, Cohen's $r = .40$. Despite this increase, they were still lower than the values predicted by the CBN-P model, $t(113) = -6.00$, $p = .001$, Cohen's $d = 1.13$, Cohen's $r = .49$.

The most likely explanation for this pattern of results is that CBN-P underestimates the impact of disablers on predictive judgment: Even when the query invites consideration of alternative causes, disablers overshadow their contribution to the judgment. This interpretation is supported by the fact that the difference between standard predictive judgments and model estimates (mean difference = −17.58) was significantly larger than the mean difference between and/or predictive judgments and model estimates (mean difference = −9.87), $t(114) = 5.31$, $p < .0001$, Cohen's $d = 1.00$, Cohen's $r = .45$.

Two important questions concerning predictive inference remain. The first is whether the parameters that are hypothesized to contribute to likelihood estimates do in fact explain significant variance. The second is whether the models contribute to likelihood estimates over and above the contribution of the crucially important $W_a$ and $W_c$ individual parameters. Two hierarchical stepwise regression analyses were conducted to answer these questions.

Turning first to standard predictive judgment, a hierarchical stepwise regression analysis was conducted in which CBN-P was entered in the final step after the stepwise regression had been conducted using $W_a$ and $W_c$ as variables. During the stepwise regression phase, only $W_c$ passed the criterion for selection; it accounted for 54% of the variance, $F(1, 112) = 132.78$, $MSE = 147.54$, $p < .001$. Adding the model CBN-P increased explained variance by only 1%, which was not a significant increase, $F(1, 111) = 1.77$, $p = .18$. Thus, causal power accounted for almost all of the explained variance in standard predictive judgments, a result that replicates Fernbach et al. (2011a, Experiment 1).

Turning next to and/or predictive judgment, $W_c$ again passed the criterion for selection, accounting for 36% of explained variance, $F(1, 112) = 63.83$, $MSE = 189.93$, $p < .001$. Strength of alternatives, $W_a$, accounted for an additional 2% of explained variance, which was marginally significant, $F(1, 111) = 3.45$, $p < .07$. Adding the model CBN-P produced less than 1% increase in explained variance, which was not significant, $F(1, 110) < 1$. Thus, causal power accounted for the majority of explained variance in modified predictive judgments, alternative strength accounted for a marginal amount, and the model accounted for none.

**Diagnostic likelihood.** Turning now to diagnostic likelihood estimates, these results indicate that CBN-D constituted a good fit for diagnostic judgments, as is apparent in Table 5. The same stepwise hierarchical regression was conducted as described above. Results indicated that $W_a$ passed the criterion for inclusion first, accounting for 22% of explained variance, $F(1, 112) = 31.54$, $MSE = 413.67$, $p < .001$. $W_c$ was entered next, accounting for an additional 10% of the variance, $F(1, 111) = 15.48$, $p < .001$. Finally, the model CBN-D was found to contribute 8% additional explained variance, $F(1, 110) = 14.09$, $p < .001$. Thus, diagnostic judgments are influenced by variation in causal power and strength of alternatives. Moreover, the way these parameters are combined in the CBN-D model accounts for unique variance above and beyond what can be attributed to those two parameters alone.

**Summary.** Diagnostic judgments were well captured by a noisy-OR model whose parameters consist of the prior probability of the cause, the strength of the cause, and the strength of alternatives. Predictive judgments, however, fell below values predicted by a noisy-OR model whose parameters consist of the strength of the cause and the strength of alternatives. When these judgments were queried in the standard way, almost all explainable variation could be attributed to variation in causal strength. When they were queried in a way that maps directly onto noisy-OR, observed estimates rose significantly but still remained lower than normative estimates. Although the rise in judgment values over what would be predicted solely by causal strength suggests that participants did take alternative strength into consideration, this factor contributed only marginally to explained variance. Perhaps the best interpretation of these results is that the noisy-OR model underestimates the impact of disablers on predictive inference.

## Experiment 3

The results of Experiments 1 and 2 showed that predictive likelihood estimates approach normative levels when they are queried using wording that maps more accurately onto the "and/or" meaning intended by researchers. These results constitute strong support for the first alternative explanation of apparent alternatives neglect in predictive inference. That they continue to fall short of normative levels, however, requires further explanation. The purpose of Experiment 3 was to test the second explanation for apparent alternatives neglect; namely, that the noisy-OR rational model of predictive inference underestimates the impact of disablers on human predictive reasoning.

Predictive inference requires reasoning from cause to effect, and ample evidence indicates that such reasoning strongly activates disablers. This suggests that disablers contribute twice to predictive judgment, once when determining causal power and again when calculating predictive estimates. As a result, they overshadow alternative causes in predictive inference so that even

Table 5
*Model Fitting Results in Experiment 2*

| Judgment | Model | Model mean | Observed mean | $t(df = 113)$ | *p* |
|---|---|---|---|---|---|
| Diagnostic | CBN-D | 56.8 | 54.3 | −1.09 | .28 |
| Standard predictive | CBN-P | 80.0 | 62.5 | −11.44 | .001 |
| Standard predictive | $W_c$ | 63.1 | 62.5 | −0.49 | .63 |
| And/or predictive | CBN-P | 80.0 | 70.2 | −6.0 | .001 |
| And/or predictive | $W_c$ | 63.1 | 70.2 | 4.69 | .001 |

*Note.* *df* = degrees of freedom; CBN-D = causal Bayes network–diagnostic; CBN-P = causal Bayes network–predictive; $W_c$ = causal power.

when the predictive query invites consideration of alternative causes, resulting predictive judgments still fall short of normative values. In Experiment 3, decision makers were given the opportunity to explicitly indicate which factors they perceived to be relevant when making predictive and diagnostic judgments. If the second alternative explanation is correct, they should indicate that alternative causes are perceived to be relevant to diagnostic inference, but both disablers and alternative causes are relevant to predictive inference.

## Method

**Participants.** Two hundred fifty Amazon Mechanical Turk workers served as participants. Participation was restricted to fluent speakers of English, residing in the United States, who had previously completed over 50 Amazon tasks and whose task approval rate was greater than 80%. Nineteen were eliminated for failing to complete the task. Of the remaining 231, 42.8% ($n = 99$) were female. The age distribution of participants was as follows: 18–25 ($n = 86$), 26–35 ($n = 94$), 36–45 ($n = 21$), 46–55 ($n = 20$), 56 and older ($n = 10$). Participants were paid $.15.

**Design.** The same two between-subject variables used in Experiment 1 were used here, judgment type (predictive or diagnostic) and causal story (learn French, quit smoking, or lose weight).

**Materials and procedure.** Participants accessed the study materials through a Survey Monkey link posted on Amazon Mechanical Turk. The same cause–effect relationships used in Experiment 1 were used here. Each participant made either a predictive judgment or a diagnostic judgment about one scenario. An example of a predictive judgment is as follows:

> *Predictive judgment: Imagine you wear a nicotine patch. How likely is it that you will quit smoking?*

A likelihood scale appeared beneath this question that ranged from 0 to 100, in 10 unit increments. The word Impossible appeared above 0, and the word Definite appeared above 100. On the next page, the following questions appeared:

> *Alternatives question: Which of the following did you consider when making your decision?*
>
> *a. other factors that can cause you to quit smoking*
>
> *b. other factors that can prevent you from quitting smoking even though you wear a nicotine patch*
>
> *c. both*
>
> *d. neither*

Participants could not go back to the previous page, and the order of choices was randomized across participants. An example of a diagnostic judgment is as follows:

> *Diagnostic judgment:* Imagine you quit smoking. How likely is it that you wore a nicotine patch?

The same scale appeared under this scenario, and the same alternatives question appeared under that. The questions appeared without the labels.

## Results and Discussion

All $t$ tests are two-tailed. All $\chi^2_{(1)}$ tests include correction for continuity.

**Frequency data.** Selections of alternative causes, disablers, both, and neither are presented in Table 6. (Story contributed no theoretical significance to the overall pattern of results. See the Appendix for details.)

There was a statistically significant contingency between choice and judgment, $\chi^2_{(3)} = 37.70$, $p < .0001$, Cramer's $V = .40$. The most frequent selections for predictive judgments were disablers only ($n = 34$, 28.3%) and both disablers and alternative causes ($n = 44$, 36.9%), and these did not differ, $\chi^2_{(1)} = 1.04$, $p = .31$. For diagnostic judgments, the most frequent selection was alternative causes only ($n = 58$, 50.9%); participants made this selection more than twice as frequently as both causes and disablers ($n = 22$, 19.8%), $\chi^2_{(1)} = 15.32$, $p < .001$. The most reasonable interpretation of the contingencies reported in Table 6 is that participants believed both disablers and alternative causes are relevant to predictive inferences, while only alternative causes were relevant to diagnostic inferences.

**Likelihood estimates.** Mean predictive and diagnostic estimates for the choice selections are illustrated in Figure 2.

Although mean predictive estimates for people who selected "disablers only" ($M = 52.6$) was lower than mean predictive judgments for people who selected "both" ($M = 60.0$), the difference was only marginally significant, $t(76) = -1.51$, $p = .14$. Mean diagnostic estimates for participants who chose "causes only" ($M = 53.1$) were also marginally lower than diagnostic estimates of those who chose "both" ($M = 63.2$), $t(78) = -1.72$, $p = .09$.

**Summary.** These results indicate that different sets of alternatives are considered relevant to predictive and diagnostic inference. For predictive judgments, disablers and alternative causes are considered relevant; for diagnostic judgments, only alternative causes are. These results have two implications. The first is that alternative causes are not neglected during predictive inference. Instead, they are overshadowed by strongly activated disablers. The second is that predictive judgments are more demanding of processing resources than diagnostic judgments because they involve processing two types of relevant information simultaneously (alternative causes and disablers) while diagnostic judgment involve processing only one (alternative causes). Given the marginal results of the predicted differences between people who believed two sources of variability were relevant and those who believe only one was, however, it was decided to investigate the impact of disablers and alternative causes more directly.

Table 6
*Frequency of "What Did You Consider?" Selections for Predictive and Diagnostic Judgments in Experiment 3*

| Judgment | $n$ | Alternative causes | Disablers | Both | Neither |
|---|---|---|---|---|---|
| Predictive | 120 | 19 | 34 | 44 | 23 |
| Diagnostic | 111 | 58 | 12 | 22 | 19 |

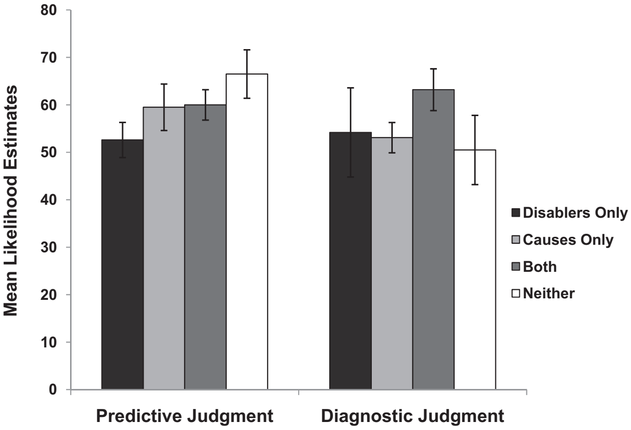

*Figure 2.* Mean likelihood estimates for predictive and diagnostic inference for participants who considered disablers only, alternative causes only, both, or neither when making their decisions in Experiment 3. Error bars are standard errors of the mean. See Table 6 for sample sizes.

## Experiment 4

The purpose in Experiment 4 was to test directly the impact of disablers and alternative causes on diagnostic and predictive inference. People were shown causal scenarios that explicitly mentioned (a) a strong cause and a weak alternative cause, (b) a strong cause and a disabler, or (c) a strong cause and irrelevant information. Because disablers strongly influence reasoning from cause to effect but are less influential when reasoning from effect to cause (e.g., Cummins et al., 1991; Fernbach & Erb, 2013), explicit mention of a disabler should reduce predictive estimates but have little impact on diagnostic judgments. Because the likelihood of an effect is presumed to be higher when multiple causes are present than when only a single cause is present, explicit mention of an additional cause should raise predictive estimates. Explicit mention of an alternative cause, however, cause should lower diagnostic judgments because it casts doubt on whether the cited cause was indeed responsible for the elicitation of the effect.

### Method

**Participants.** One hundred forty-five Amazon Mechanical Turk workers served as participants. Participation was restricted to fluent speakers of English, residing in the United States, who had previously completed over 50 Amazon tasks and whose task approval rate was greater than 80%. Two participants were eliminated for failing to complete the task. Of the 143 remaining, 79 (55.2%) were female. Age distribution was as follows: 18–25 (*n* = 48), 26–35 (*n* = 54), 36–45 (*n* = 12), 46–55 (*n* = 17), 56 and older (*n* = 12). Participants were paid $.25.

**Design.** The experiment employed one between-subject factor and two repeated measures. The between-subject measure was type of additional information (disabler, alternative cause, or irrelevant event). The first repeated measure was judgment type (predictive and diagnostic). The second was causal scenario. Participants made predictive judgments about four causal stories, and diagnostic judgments about a different set of four causal stories. Their estimates were then averaged within each set so that each participant contributed two values for analysis: a mean predictive estimate and a mean diagnostic estimate.

**Materials and procedure.** Table 7 lists the causal story, the alternative cause, the disabler, and the irrelevant event cited for each causal relationship used in Experiment 4. The stories were chosen from materials used by Fernbach et al. (2010, supplemental material). The target cause was always the strong cause, and the additional cause was the weaker cause. (Weak causes were selected through pilot work.) Using the nicotine patch example again, the resulting scenario appeared as follows (sans labels):

***Predictive judgments.***

> *Alternative weak cause cited: Imagine Tom wears a nicotine patch. Tom also begins seriously dating a nonsmoker. How likely is it that Tom will quit smoking?*
>
> *Disabler cited: Imagine Tom wears a nicotine patch. Tom also begins seriously dating a woman who is a heavy smoker. How likely is it that Tom will quit smoking?*
>
> *Irrelevant event cited: Imagine Tom wears a nicotine patch. Tom also begins wearing a baseball cap. How likely is it that Tom will quit smoking?*

***Diagnostic judgments.***

> *Alternative weak cause cited: Imagine Tom quit smoking. Before he quit, Tom had begun seriously dating a nonsmoker. How likely is it that Tom used a nicotine patch?*

Table 7
*Alternative Causes, Disablers, and Irrelevant Events Cited for Each Causal Story Used in Experiment 4*

| Effect | Cause | Alternative | Disabler | Irrelevant |
|---|---|---|---|---|
| Quit smoking | Nicotine patch | Dates nonsmoker | Dates heavy smoker | Wears baseball cap |
| Learn French | Rosetta Stone | Dates French speaker | Doesn't practice | Plays tennis |
| Lose weight | Exercises hard | Eats less | Eats more | Wears T-shirt |
| Complete marathon | Trains hard | Endurance supplements | Sustains injury | Lives in apartment |
| Millionaire by age 40 | Inherits $750,000 | High salary | Spendthrift | Likes coffee |
| Good grade | Studies hard | Took class before | Difficult course | Likes folk music |
| Romance | Dating service | Friends introduce | Social anxiety | Watches TV |
| Play guitar | Lessons | Natural talent | Doesn't practice | Has a sister |

*Note.* Causal stories selected from Fernbach et al. (2010, supplemental materials).

> *Disabler cited: Imagine Tom quit smoking. Before he quit, Tom had begun seriously dating a heavy smoker. How likely is it that Tom used a nicotine patch?*
>
> *Irrelevant event cited: Imagine Tom quit smoking. Before he quit, Tom had begun wearing a baseball cap. How likely is it that Tom used a nicotine patch?*

Two different causal scenario sets were constructed. The first consisted of the predictive inferences for smoking, French, weight, and marathon and diagnostic inferences for millionaire, grade, romance, and guitar. The second switched the judgment (i.e., diagnostic judgments for smoking, French, weight, and marathon; predictive judgments for millionaire, grade, romance, and guitar). Order of judgment presentation (whether predictive judgments block was presented first or second) was randomized across participants.

Participants accessed the study by clicking on a Survey Monkey link posted on Amazon Mechanical Turk.

## Results and Discussion

All *t* tests are two-tailed. Each participant's mean predictive and diagnostic likelihood estimates were calculated, and they were analyzed via ANOVA using judgment (predictive and diagnostic) and information (alternative cause, disabler, or irrelevant) as between-subject factors. The main effects of judgment and information were significant, $F(1, 140) = 6.4$, $MSE = 154.7$, $p < .025$, $\eta^2 = .05$, and $F(2, 140) = 10.3$, $MSE = 229.7$, $p < .0001$, $\eta^2 = .13$. These factors interacted, $F(2, 140) = 73.5$, $MSE = 154.7$, $p < .0001$, $\eta^2 = .51$. The results are illustrated in Figure 3.

Two planned comparisons were conducted for both types of judgments using mean likelihood estimates in the irrelevant information condition as a control. Turning first to predictive judgments, the mean likelihood estimate was significantly higher following the insertion of an alternative cause ($M = 70.5$) than in the control condition ($M = 63.4$), indicating that the explicit mention of alternative causes raised predictive judgments, $t(94) = 2.47$, $p < .02$, Cohen's $d = .5$. Disablers, on the other hand, significantly lowered predictive judgments ($M = 40.9$, $SD = 12.04$), $t(94) = 8.07$, $p < .0001$, Cohen's $d = 1.66$. More important, the impact of disablers (Cohen's $r^2 = 40.9\%$) was seven times greater the impact of alternative causes (Cohen's $r^2 = 5.7\%$).

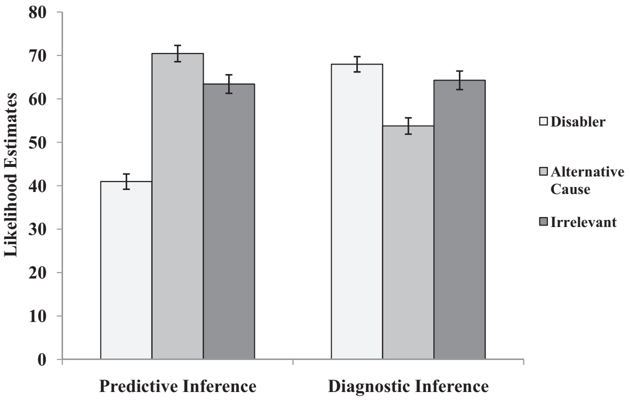

*Figure 3.* Mean likelihood estimates for predictive and diagnostic judgments when the problem description cited a strong generative cause and (a) a weak alternative generative cause, (b) a disabler, or (c) an irrelevant factor in Experiment 4. Compared to the irrelevant information condition, disablers lowered likelihoods for predictive inference and had no effect on diagnostic inference. A weak alternative cause lowered likelihoods for diagnostic judgments while raising predictive estimates. The impact of disablers on lowering predictive judgments was seven times greater than the impact of alternative causes on raising them.

Turning now to diagnostic judgments, the introduction of an alternative cause had the opposite effect, lowering likelihood estimates ($M = 53.8$) compared to the control condition ($M = 64.3$), $t(94) = -3.52$, $p < .001$, Cohen's $d = .72$. Disablers had no significant impact on diagnostic judgments ($M = 67.9$), $t(94) = 1.29$, $p = .20$, Cohen's $d = .26$. The impact of alternative causes on diagnostic judgment was twice their impact on predictive judgment (Cohen's $r^2 = 11.6\%$ vs. $5.7\%$).

**Summary.** These results show that alternative causes compete with the candidate cause as explanations of the effect, and explicit mention of them lowers people's diagnostic estimates. Disablers had virtually no impact on diagnostic judgment. In contrast, alternative causes raised predictive judgments while disablers lowered them, which constitutes a rational (rather than biased) decision profile. But the fact that the impact of disablers on predictive likelihood judgments was seven times greater than the impact of alternative causes on these judgments suggests that disablers greatly overshadow alternative causes in their contributions to predictive judgment.

# Experiment 5

The results of Experiments 1–4 showed that people interpret standard predictive inference queries as requests to estimate the likelihood that the cause will bring about the effect (causal power), that they believe disablers are relevant to this decision, and that the impact of disablers is substantial when they are explicitly presented. The purpose of Experiment 5 was to more directly test the impact of disablers on causal power judgments by allowing participants the opportunity to list the disablers they themselves consider relevant to such a judgment. Accordingly, participants made a causal power judgment, then listed the disablers considered relevant to causal judgments and provided prior probability and strength information for each disabler retrieved.

Two models were then tested using these data. The first model tested was the model of causal power proposed by Fernbach and Erb (2013), in which causal power estimates are based on an aggregate disabling probability. The purpose of the model is to account for variability in modus ponens inferences that are based on causal conditionals (i.e., Cause -> Effect/Cause/Therefore, Effect). As pointed out by Cummins (1995, 1997, 2010; Cummins et al., 1991), when evaluating such arguments, people perform a causal analysis in which disablers loom large, rather than relying on simple truth functions (which cannot capture the meaning of causal conditionals). As pointed out by Fernbach and Erb (2013), this means that people assess causal power when evaluating such arguments, and their decisions can be considered causal power estimates, MP = P(Effect/Cause, No Alternatives) = $W_c$. Fernbach and Erb (2013) propose that such estimates are derived as follows: Each disabler has some prior likelihood of being present ($P_d$) and, when present, has likelihood of preventing the effect from occurring, which constitutes its strength ($W_d$). The disabling probability of any given disabler (Ai) is equal to the product of its prior probability and its strength,

$$A_i = P_{di} * W_{di}. \tag{3}$$

The likelihood that the cause will successfully bring about an effect is the aggregate of these individual disabling probabilities:

$$A' = \sum_{i=1}^{n} Ai - \sum_{i,j:i<j} AiAj + \sum_{i,j,k:i<j<k} AiAjAk - \ . \ . \ . \ + (-1)^{n-1} \prod_{i=1}^{n} Ai. \tag{4}$$

Causal power, $W_c$, is the complement of this aggregate disabling probability, which means that it expresses the likelihood that the cause will bring about the effect when there are no disablers to prevent it:

$$W_c = 1 - A'. \tag{5}$$

For the duration of this paper, A′ will be referred to as *disabler impact,* as it more clearly indicates what this value captures—the impact of disablers on causal power estimates.

The second model tested is the retrieval-driven model of causal judgment proposed by Cummins (2010), which is summarized in Equation 6 and defines how decision makers arrive at causal power estimates:

$$W_c = B * (\alpha / (\alpha + \text{disablers})). \tag{6}$$

B is a parameter that reflects the believability of the causal mechanism underlying the purported causal relationship. The inclusion of this parameter is a theoretical commitment that people treat causal contingency differently than simple covariation. If they do not believe the two events are causally related (B = 0), disablers are irrelevant and hence not activated in memory. Only when they believe a causal mechanism exists that empowers one event to evoke another (B = 1) do disablers become relevant. This is consistent with the results of Fernbach and Erb (2013), in which acceptance levels for arguments based on noncausal conditionals were found to differ from those of causal conditionals, despite similarity of their conditional probabilities. When evaluating arguments with embedded noncausal conditionals, acceptance levels hovered around uncertainty. In contrast, when evaluating arguments with embedded causal conditionals, acceptance levels varied as a function of disablers retrieved. It is also consistent with research showing that people ignore or discount covariation information if no they can think of no plausible causal mechanism whereby the purported cause can bring about the effect (e.g., Ahn & Kalish, 2000; Ahn, Kalish, Medin, & Gelman, 1995; Fugelsang & Thompson, 2001; Fugelsang, Thompson, & Dunbar, 2006).

The term (α/(α+disablers) is a memory activation function—a positively accelerated curve—in which the first few disablers retrieved from memory have greater impact on judgment than those retrieved later. Activation spreads throughout the network of associated disablers, and likelihood estimates drop off significantly the farther it spreads. This is because stronger disablers are presumed to be activated earlier than weaker ones, and therefore have greater impact on judgment outcomes. In other words, the psychological difference between 0 and 3 items is greater than the psychological difference between 4 and 7. α is a free parameter; it simply expresses the steepness of the curve, and its value is determined empirically. (See also Rehder, 2003 for a similar account for categorization judgments.) Figure 4 depicts causal power likelihood estimates for different disabler and α values when B = 1.

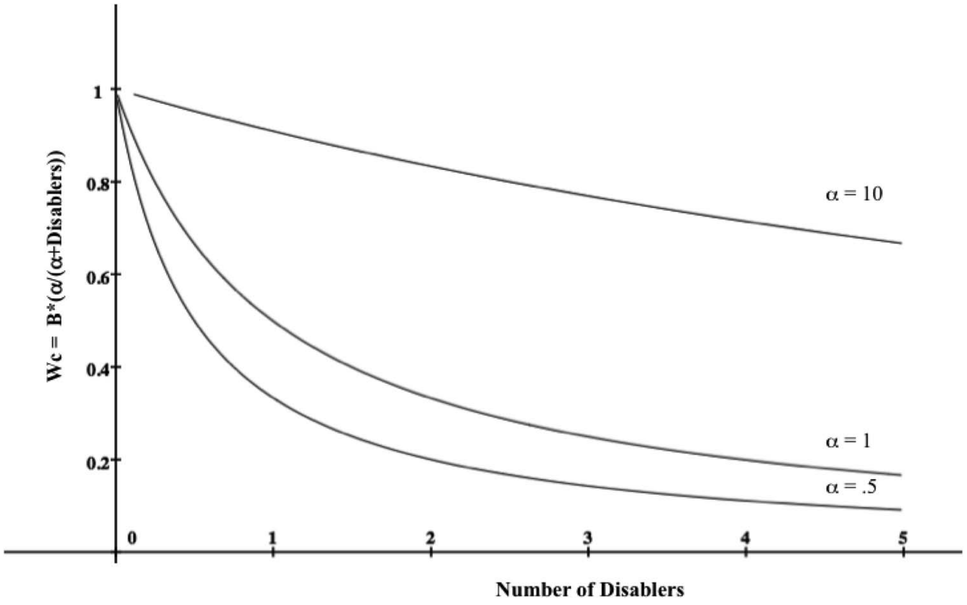

*Figure 4.* A model of causal power values ($W_c$) as a function of belief that a causal mechanism underlies the empirically observed regularity. In the graph, B = 1, meaning that the decision maker believes the contingency reflects a causal relationship. The function shows that the first few disablers retrieved have greater impact on causal power estimates than ones retrieved later.

The model captures the likelihood of an effect occurring when a cause is present and disablers are absent, and its crucial prediction is that both number of disablers and order of disabler retrieval matter. This memory retrieval characteristic of the model is what distinguishes it from the model proposed by Fernbach and Erb (2013), where such information is lost once individual disabler impact information is aggregated.

The legitimacy of this model is supported by studies of causal conditional reasoning. De Neys et al. (2003) presented arguments in which the number of explicitly mentioned disablers ranged from 0 to 4. They found a significant negative linear trend on acceptance ratings. De Neys et al. also reported that while "thinking aloud," reasoners did not halt the retrieval process upon retrieving a single counterexample. Instead, they continued to retrieve disablers or alternatives until a final judgment was made, and willingness to accept causal argument conclusions declined as more disablers were activated in memory. Their results suggested a nonlinear retrieval function, however, in which a threshold occurred at about 3 retrieved items, after which argument acceptance ratings changed very little. Nonlinear functions are common among other cognitive and perceptual phenomena (e.g., sensory thresholds, serial position curves). Negatively decelerating ones such as this are typical of spreading activation functions in associative memory models.

To summarize, according to Cummins (2010) (a) causal power likelihood estimates diminish as the number of disablers retrieved increases, and (b) earlier retrieved disablers have greater impact than later ones. According to Fernbach and Erb (2013), causal power likelihood can be captured by aggregate disabler impact, and value is not affected by order of disabler retrieval.

## Method

**Participants.** Eighty-six Amazon Mechanical Turk workers served as participants in the experiment. Participation was restricted to fluent speakers of English, residing in the United States, who had previously completed over 50 Amazon tasks and whose task approval rate was greater than 80%. Seven were eliminated for failing to provide all requested estimates. Of the remaining 79, 47.8% ($n = 37$) were female. Age categories (and sample sizes) were as follows: 18–25 ($n = 15$), 26–35 ($n = 37$), 36–45 ($n = 13$), 46–55 ($n = 5$), 56 and older ($n = 7$); Two participants declined to reveal their ages. Participants were paid $.25 for their participation.

**Design.** The experiment employed one between-subject variable and four repeated measures. The between-subject variable was causal story chosen from Fernbach et al. (2010) (using Rosetta Stone to learn French, using a nicotine patch to quit smoking, or exercising hard to lose weight). The repeated measures were causal power estimate, disabler count, disabler prior probability, and disabler strength.

**Materials and procedure.** Participants read one causal story, gave a causal power estimate, and then listed disablers that could prevent the effect in the presence of the cause. They then went back and, for each disabler listed, estimated its prior probability and causal strength.

Using the nicotine patch scenario as an example, the questions were as follows:

> *John is an average male between the ages of 18 and 25 who wants to quit smoking.*
>
> *Imagine John wears a nicotine patch. How likely is it that John will quit smoking?*
>
> *Enter your decision on a scale of 0 to 100, where 0 means "Impossible" and 100 means "Definite."*

Participants indicated their decision by typing a number between 0 and 100 in a box.

On the next page, the protocol continued as follows:

> *Disabler list query: Imagine that John wore a nicotine patch exactly according to instructions. Think about factors that might prevent him from quitting smoking anyway. Please write down as many factors as you can think of that could make this happen.*

Participants then listed factors by typing them on the screen.

> *Now we are interested in more information about the factors you listed. Answer the following questions for each factor you listed by entering a number from 0 to 100 in the boxes, where 0 means "Impossible" and 100 means "Definite."*

| | *How likely is it that the factor will occur?* | *If the factor occurred, how like is it to prevent John from quitting smoking?* |
|---|---|---|
| *First Factor* | *[____]* | *[____]* |
| *Second Factor* | *[____]* | *[____]* |
| *Third Factor* | *[____]* | *[____]* |

and so on.

Participants entered their decisions by typing a number between 0 and 100 in the boxes. On the next page, they were asked to indicate their sex and their age group.

Participants accessed the study materials through a Survey Monkey link posted on Amazon Mechanical Turk. Story was randomized across subjects. Participants could not go back to previous pages.

## Results and Discussion

All *t* tests are two-tailed. Participants listed a total of 239 items. Disabler lists were reviewed by two independent checkers for duplicates (e.g., "John was depressed because he lost his job; John was depressed because his girlfriend left him") or illegitimate items (e.g., "John did not follow instructions"). Twenty-one items (8.9%) were eliminated using these criteria, leaving a total of 218 items listed. Disabler counts ranged from zero to six. Table 8 displays the number of participants logging each disabler count, corresponding mean likelihood estimate, and corresponding mean normative causal power estimate calculated using Equation 5.

Table 9 displays mean prior probability, mean disabler strength, and mean impact (the product of these two) for disablers as a function of retrieval order. Only seven participants retrieved more than four disablers, and variability was quite high among their probability estimates. For this reason, they were excluded from subsequent analyses. Sample sizes for disabler count zero and one were also small ($n = 4$ and 9, respectively), but variability was comparable to the other categories. For this reason, these two categories were retained or collapsed in subsequent analyses, as indicated.

Table 9
*Mean Prior Probability ($P_d$), Mean Disabler Strength ($W_d$), and Mean Disabler Impact (A′) According to Retrieval Order From Experiment 5*

| Retrieval order | *n* | Prior probability ($P_d$) | Strength ($W_d$) | Disabler impact (A′)[a] |
|---|---|---|---|---|
| First | 68 | 58.8 (2.7) | 72.2 (2.6) | 44.9 (2.8) |
| Second | 59 | 50.2 (2.9) | 64.6 (2.9) | 33.8 (2.8) |
| Third | 40 | 41.4 (3.8) | 56.2 (4.2) | 24.9 (3.2) |
| Fourth | 13 | 36.0 (4.2) | 60.8 (5.6) | 22.4 (3.2) |
| Fifth | 7 | 39.7 (5.4) | 70.0 (11.1) | 26.7 (6.0) |
| Sixth | 2 | 40.0 (35.0) | 87.5 (7.5) | 37.6 (3.4) |

[a] $A' = 100 * (P_d/100) * (W_d/100)$.

The two main theoretical commitments of the model proposed by Cummins (2010) are that (a) causal power estimates diminish as the number of disablers retrieved increases, and (b) earlier retrieved disablers should have greater impact than later ones. As a test of the first hypothesis, a stepwise regression was conducted on causal power estimates using disabler count and aggregate causal power as predictor variables. Recall that disabler impact of any given disabler is equal to the product of its prior probability and its strength (Equation 3) and that the likelihood that the cause will successfully bring about an effect is the aggregate of these individual disabling probabilities (Equation 4). Aggregate causal power, $W_c$, is the complement of this aggregate disabling probability, which means that it expresses the likelihood that the cause will bring about the effect when there are no disablers to prevent it (Equation 5). Only disabler count satisfied the criterion for inclusion in the resulting equation, $F(1, 70) = 9.13$, $MSE = 400.0$, $p < .004$, $R^2 = 11.5\%$. The standardized regression coefficient was $-.34$, indicating that the larger the number of disablers retrieved, the lower the causal power estimate. The standardized regression coefficient for aggregate causal power was .14 ($p = .28$).

The data trends apparent in Table 9 are consistent with the second hypothesis: Prior probability, disabler strength, and disabler impact all decline as a function of retrieval order. In order to test this prediction, two sets of analyses were conducted. First, three planned comparisons were conducted: When one disabler was retrieved, the mean causal power estimate was 84.4. This declined significantly to 65.6 when an additional disabler was retrieved, $t(26) = 2.51$, $p < .02$, Cohen's $d = .98$, Cohen's $r = .44$. However, no significant decline was observed between two disablers and three ($M = 64.7$), $t(44) < 1$, or between three disablers and four ($M = 63.1$), $t(38) < 1$. Next, the mean disabler impacts of the first, second, and third disablers retrieved were compared for participants who retrieved three or four disablers (using $P_d * W_d$ as a measure of disabler impact). Mean disabler impact for each disabler and its correlation with causal power estimates are displayed in Table 10.

Planned comparisons indicated that high-impact disablers were retrieved prior to weaker ones. Mean impact of the first retrieved disabler ($M = 44.3$) was significantly higher than the mean impact of the second disabler retrieved ($M = 31.4$), $t(39) = 3.38$, $p < .01$, Cohen's $d = 1.08$, Cohen's $r = .48$, and the mean impact of the second disabler retrieved was greater than the mean impact of the third retrieved ($M = 24.9$), $t(39) = 2.70$, $p < .01$, Cohen's $d = .86$, Cohen's $r = .40$. Correlations with causal power, however, showed the reverse pattern. Causal power estimates correlated significantly with the impact of last disabler retrieved ($r = -.31$, $p < .05$), while the impact of the first two retrieved did not ($r = .07$ and $-.13$ for the first and second disabler retrieved, respectively).

**Summary.** The results of Experiment 5 indicated that number of disablers matter—the more retrieved, the less likely the effect. But order of retrieval matters as well. As predicted by Cummins (2010) (Equation 6 above), stronger items are retrieved first, but, contrary to the model, the ultimate judgment is more strongly influenced by later retrieved items than by earlier ones. That is why aggregate impact scores do not capture final likelihood judgments well. The impact of later items may constitute a recency effect in causal judgment.

## General Discussion

The purpose in this work was to test a two-part explanation of apparent alternative neglect in predictive inference; namely, that (a) standard predictive inference queries are frequently interpreted as requests to estimate the likelihood that a particular cause will bring about the effect and (b) the impact of disablers on human predictive inference is underestimated in current probabilistic models. The results of five experiments were consistent with this two-part explanation The results of Experiment 1 showed quite clearly that when given the opportunity to select the meaning of the standard predictive inference query from among alternatives, the most frequent interpretations were (a) a request to estimate the likelihood of E given C alone, and (b) a request to estimate the

Table 8
*Mean Observed Causal Power ($W_c$) Estimates and Mean Normative Aggregate Causal Power (A′) Estimates as a Function of Number of Disablers Retrieved in Experiment 5*

| Disabler count | *N* | Mean observed causal power estimate ($W_c$) | Mean normative causal power estimate (A′) |
|---|---|---|---|
| 0 | 4 | 90.0 (5.8) | 100 (0) |
| 1 | 9 | 84.4 (5.9) | 51.5 (5.4) |
| 2 | 19 | 65.6 (4.3) | 35.6 (5.1) |
| 3 | 27 | 64.7 (4.2) | 32.3 (4.9) |
| 4 | 13 | 63.1 (5.6) | 25.9 (5.7) |
| 5 | 5 | 54.0 (3.7) | 26.8 (9.5) |
| 6 | 2 | 45.0 (5.0) | 30.0 (14.0) |
| Total | 79 | | |

Table 10
*Mean Disabler Impact (A′) and Correlation With Causal Power Estimates ($W_c$) as a Function of Retrieval Order in Experiment 5*

| Retrieval order | A′ | *r*(A′, $W_c$) | *p* |
|---|---|---|---|
| 1 | 44.3 (3.8) | .07 | .68 |
| 2 | 31.4 (3.2) | −.13 | .41 |
| 3 | 24.9 (2.2) | −.31 | .05* |

*Note.* $n = 40$.
* $p \leq .05$.

likelihood that C can cause E. Rarely did participants interpret it as a request to estimate the likelihood of E given C and/or other causes. Diagnostic inference queries were overwhelmingly interpreted as requests to estimate the likelihood that the effect was due to the cited cause as opposed to alternative causes. Experiment 2 directly tested the sufficiency of the noisy-OR model proposed by Fernbach et al. (2011a) to capture predictive and diagnostic inferences. The results indicated that the model constituted an excellent fit to diagnostic inference, but greatly overestimated predictive inference relative to observed estimates when this type of inference was queried in the standard way. But when predictive inference was queried using "C and/or alternatives" (which more accurately expresses the meaning of noisy-OR), likelihood estimates approached normative likelihood estimates. In Experiment 3, people overwhelmingly selected alternative causes as relevant for diagnostic inferences; for predictive inference, they selected "disablers only" and "both disablers and alternatives causes" with equivalent frequency. Experiment 4 showed that the impact of disablers on predictive judgments was seven times greater than the impact of alternative causes on predictive inference, while having negligible impact on diagnostic judgments. Experiment 5 allowed participants to explicitly list relevant disablers and to provide estimates of their prior probability and strength. The results showed quite clearly that both number of disablers and order of disabler retrieval matter in causal power estimates. The larger the number of disablers retrieved, the lower the causal power estimate. Disablers with strong impact were retrieved early in the process, but later retrieved items exerted the greatest influence on ultimate estimates.

More important, these results indicate that the probabilistic models proposed by Fernbach and colleagues do not constitute adequate descriptive models of human predictive inference. Disablers overshadow alternative causes in ways that are not captured by causal Bayes networks. This implies that human predictive inference is not purely Bayesian. As well documented by Tversky and Kahneman (1973; Kahneman, 2011), the source of the discrepancy appears to lie in the way knowledge retrieval transacts with probability estimations. Automatic (e.g., Cummins, 1995, 2010) or deliberate (e.g., Johnson-Laird, 2006) activation of relevant counterexamples is a hallmark of human reasoning, and this characteristic must be accommodated in descriptive models of causal inference. The results reported here indicate that causal judgments are strongly tempered by memory retrieval/activation processes that are not adequately captured by the models: Predictive inferences are made within a cognitive context that is populated by strongly activated disablers and weakly activated alternative causes, while diagnostic inferences occur within a cognitive context that is populated by strongly activated alternative causes and weakly activated disablers. This result is consistent with results reported by Fernbach and Rehder (2012), in which difficulty of retrieving alternative causes from memory was not sufficient to explain their paltry contribution to predictive inference. Across three experiments, participants were instructed on the causal links as part of the experimental session and then provided a diagram of those relations during the inference task. Nonetheless, alternative causes did not contribute significantly to predictive estimates—even when the focal cause was described as particularly weak. Put another way, research on predictive inference consistently shows that people focus on what they believe to be relevant when making causal judgments. They believe disablers to be highly relevant to predictive inference and alternative causes to be less so.

The results of Experiment 5 again underscore that knowledge retrieval effects prove problematic for the descriptive power of causal Bayes networks as currently conceived. Order of disabler retrieval was found to be systematically related to disabler impact, and disabler impact was systematically related to causal power estimates. This information is lost when aggregating across disablers, thereby reducing modeling accuracy. The noisy-OR model predicts that disablers have equivalent impact on predictive and diagnostic inference, yet the data presented here quite arguably prove otherwise. Disablers influenced predictive inference far more strongly than they did diagnostic inference in Experiment 3, while causal power estimates in Experiment 1 did not differ between diagnostic and predictive inference. One interpretation of these results is that disablers are considered twice when making predictive inference (once when estimating causal power and again when making a predictive judgment), thus magnifying their impact. The result of this knowledge-retrieval process is the underestimation of the impact of disabler impact by the model. Moreover, it also suggests that predictive inference is more resource demanding than diagnostic inference because of the need to consider the impact of disablers twice during the judgment process. The impact of knowledge-retrieval effects and differential processing demands may be sufficient to explain the failure of human predictive inference to match normative standards even when the "and/or" query form is used.

Fernbach and Rehder (2012) acknowledged that the need to reduce computational complexity may underlie observed differences between normative and observed causal inference. They argued that augmenting the causal Bayes networks with cognitive shortcuts that reduce processing demands may provide a more complete account of causal inference.

Other researchers have pointed out the computational intractability of Bayesian inference in real time, a drawback that renders their wholesale use as models of human causal inference questionable. Rottman and Hastie (2014) pointed out that performing the full Bayesian calculations requires reasoning about many nodes simultaneously, combining causes in complex ways, and using multiple parameters for a single inference. As a result, the chief challenge human reasoners face is finding a way to reduce the complexity of the computation, such as focusing primarily on one piece of information and then sequentially adjusting for other pieces of information.

Another way to distinguish among the various explanations that have been offered for causal inference is to view them as addressing different explanatory goals. Marr (1982) introduced the distinction between computational, algorithmic, and implementational levels of explanation. The computational level theory specifies what is computed (e.g., the sum of inputs). The algorithmic level theory specifies a representational scheme for the inputs and outputs of a system, and the algorithm(s) that transform the former into the latter (e.g., Arabic numerals that represent numbers, and a function that maps sets of input numerals into an output that represents the sum of the input numbers). The implementational level theory specifies a physical realization of the algorithmic level (e.g., calculator, abacus, second-grade student). Using this classification scheme, causal Bayes networks can be understood as computational theories while the model offered by Cummins (2010) can be understood as an algorithmic theory.

Another view, however, may be informative in the current context, and that is the distinction between unbounded and bounded rationality. Bayesian models of rationality assume infinite processing time and resources. Bounded rationality, a term introduced by Herbert Simon (1957, 1991) is the notion that rational decision making is limited by the information available, the cognitive resources available to devote to the task, and the finite amount of time within which the decision needs to be made. From this perspective, human decision-makers are not optimizers, they are satisficers: They seek to obtain the most accurate decision possible given limited time and processing resources. It is important to note that the notion of rationality cannot be abstracted away from resource and processing constraints. Consider the question of what is the best car for a consumer to buy. The rational choice in such a case does not depend simply on the quality of vehicles under consideration but also on the consumer's bank account and time available for car-comparison shopping. A Lexus may be a rational choice for someone with unlimited financial resources but an irrational choice for someone on a limited budget. Similarly, it may be rational for a consumer with unlimited free time to take time out to test drive and compare all possible choices, but it would be an irrational waste of time for a consumer with limited free time to engage in such industrious car shopping. The same argument can be made regarding decision windows, cognitive resources, and memory capacity: They are finite in nature, and the complexity of Bayesian calculations can make demands that rapidly exceed decision processing time, effort, and memory limits. In other words, given infinite time, processing capacity, and memory capacity, anyone can be Bayesian. In the real world, animals are afforded no such luxury. Instead, they are imbued with decisional biases and heuristics that constitute decisional satisficers: They do an sufficient job most of the time when operating in normal environments (Todd & Gigerenzer, 2007). The moral of the story is that the nature of what constitutes a rational choice changes as a function of processing and time constraints in which the decision must be made.

From this perspective, three possibilities present themselves. The first is that causal Bayes networks, as currently formulated, underestimate the true impact of disablers on causal inference. In the model proposed by Fernbach and colleagues, the impact of disablers is implicit in causal power estimates. The results of Experiment 1 suggest that people distinguish between causal power as a measure of the causal plausibility ("can cause") and the ability of the cause to overcome disablers ("will cause"). This is not unreasonable, so the question becomes whether causal Bayes net computations should separate these computations when modeling predictive inference. The second possibility is that the overweighting of disablers relative to alternative causes in predictive inference constitutes a genuine error-producing bias in ordinary human causal inference. The third is that human predictive likelihood estimates constitute evidence of bounded rationality (i.e., values that are good enough to guide action while still being computationally tractable in real time).

The take-home message of this work is twofold. First, human causal inference cannot be adequately modeled without taking into consideration the ways in which knowledge is activated and applied in the decision process. Second, making decisions in real time with limited computational resources necessitates employing cognitive shortcuts that allow for realistic management of time and processing constraints while preserving a satisficing degree of accuracy.

## References

Ahn, W. K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199–225). Cambridge, MA: MIT Press.

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition, 54,* 299–352. doi:10.1016/0010-0277(94)00640-7

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition, 53,* 217–238. doi:10.1016/0010-0277(94)90049-3

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction, 8,* 261–289. doi:10.1207/s1532690xci0803_2

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23,* 646–658. doi:10.3758/BF03197265

Cummins, D. D. (1997). Reply to Fairley and Manktelow's comment on "Naïve theories and causal deduction." *Memory & Cognition, 25,* 415–416. doi:10.3758/BF03211297

Cummins, D. D. (2010). How memory processes temper causal inferences. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals* (pp. 207–218). New York, NY: Oxford University Press.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438. doi:10.1016/0010-0285(88)90011-4

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274–282. doi:10.3758/BF03211151

Dellarosa, D. (1986). A computer simulation of children's arithmetic word-problem solving. *Behavior Research Methods, Instruments & Computers, 18,* 147–154. doi:10.3758/BF03201014

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition, 30,* 908–920. doi:10.3758/BF03195776

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581–595. doi:10.3758/BF03196099

Elio, R. (1998). How to disbelieve $p \geq q$: Resolving contradictions. *Proceedings of the Twentieth Meeting of the Cognitive Science Society* (pp. 315–320). Mahwah, NJ: Erlbaum.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21,* 329–336. doi:10.1177/0956797610361430

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011a). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140,* 168–185. doi:10.1037/a0022100

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011b). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition, 119,* 459–467. doi:10.1016/j.cognition.2011.01.013

Fernbach, P. M., & Erb, C. D. (2013). A quantitative theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1327–1343. doi:10.1037/a0031851

Fernbach, P. M., & Rehder, B. (2012). Cognitive shortcuts in causal inference. *Argument and Computation, 4,* 64–88. doi:10.1080/19462166.2012.682655

Fugelsang, J. A., & Thompson, V. (2001). Belief-based and covariation-based cues affect causal discounting. *Canadian Journal of Experimental Psychology, 55,* 70–76. doi:10.1037/h0087354

Fugelsang, J., Thompson, V., & Dunbar, K. (2006). Examining the representation of causal knowledge. *Thinking & Reasoning, 12,* 1–30. doi:10.1080/13546780500145678

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines, 8,* 39–60. doi:10.1023/A:1008234330618

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32. doi:10.1037/0033-295X.111.1.3

Griffiths, T. L., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. doi:10.1037/a0017201

Janveau-Brennan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology, 35,* 904–911. doi:10.1037/0012-1649.35.4.904

Johnson-Laird, P. N. (2006). *How we reason*. New York, NY: Oxford University Press.

Kahneman, D. (2011). *Thinking: Fast and slow*. New York, NY: Penguin Books.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856–876. doi:10.1037/0278-7393.30.4.856

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. doi:10.1037/a0013256

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Thinking & Reasoning, 6,* 335–347. doi:10.1080/135467800750038166

Markovits, H., Fleury, M., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69,* 742–755. doi:10.1111/j.1467-8624.1998.tb06240.x

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition, 29,* 736–744. doi:10.3758/BF03200476

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54,* 33–61. doi:10.1016/j.cogpsych.2006.04.004

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. doi:10.1037/0033-295X.101.4.608

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Kaufmann.

Pearl, J. (2000). *Causality*. New York, NY: Cambridge University Press.

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: Strength of association as a predictive factor for content effects. *Cognition, 68,* B93–B101. doi:10.1016/S0010-0277(98)00053-5

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1141–1159. doi:10.1037/0278-7393.29.6.1141

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140,* 109–139. doi:10.1037/a0031903

Simon, H. (1957). *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*. New York, NY: Wiley.

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science, 2,* 125–134. doi:10.1287/orsc.2.1.125

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 35–42). Cambridge, MA: MIT Press.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science, 16,* 167–171. doi:10.1111/j.1467-8721.2007.00497.x

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. doi:10.1016/0010-0285(73)90033-9

Verschueren, N., Schacken, W., De Neys, W., & d'Ydewalle, G. (2004). The difference between generating counterexamples and using them during reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 57*(A), 1285–1308. doi:10.1080/02724980343000774

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 453–484). Oxford, England: Oxford University Press.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Deny (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Erlbaum.

(*Appendix follows*)

# Appendix

## Preliminary Analyses That Included Story as a Variable in Experiments 1–3

### Experiment 1

A log-linear analysis was conducted on choice frequency using story (French, nicotine, or exercise), judgment (predictive or diagnostic), and choice (causal power, alone, or and/or) as variables. The three-way interaction was not significant, $G^2_{(12)} = 18.5$, $p = .08$, nor did story significantly interact with the other two variables, Story × Judgment $G^2_{(4)} = 1.88$, $p = .39$; Story × Choice $G^2_{(4)} = 3.26$, $p = .51$. As predicted, the interaction of judgment and choice was significant, $G^2_{(2)} = 10.26$, $p < .01$.

### Experiment 2

Four preliminary omnibus mixed ANOVAs were conducted to evaluate the impact of the story variable on judgments. The first ANOVA assessed predictive likelihood ratings using story (learn French, quit smoking, or lose weight) as a between-subject variable and query (standard predictive likelihood and and/or predictive likelihood) as repeated measures. The main effect of query was significant, $F(1, 111) = 28.57$, $MSE = 121.80$, $p < .0001$, as was the main effect of story, $F(2, 111) = 13.58$, $MSE = 404.62$, $p < .001$. The interaction of the two variables was not significant ($F < 1$), and subsequent planned $t$ tests indicated that the impact of query wording was significant for each story (i.e., and/or estimates > standard estimates for French, quit, and weight, $t(39) = 2.51$, $p < .02$); $t(42) = 3.22$, $p < .01$; $t(30) = 3.78$, $p < .01$, respectively). For the second mixed ANOVA, predictive model likelihood estimates were subtracted from observed predictive likelihood estimates, and these difference scores were analyzed using story as a between-subject variable and query as a repeated measure. The main effect of story was again significant, $F(2, 111) = 6.24$, $MSE = 421.46$, $p < .003$, $\eta = .10$, as was the main effect of query, $F(1, 111) = 28.57$, $MSE = 121.80$, $p < .001$, $\eta = .21$. Tukey's pairwise comparisons indicated that the mean difference for the story about quitting smoking ($M = -19.2$) was significantly larger than the mean difference for story about losing weight ($M = -7.3$), $HSD_{(05)} = 3.40$, $p < .01$. The two variables did not interact, however ($F < 1$).

For the third, a one-way ANOVA was conducted on diagnostic judgments using story as a between-subject factor. The main effect of story again was significant, $F(2, 111) = 4.08$, $MSE = 498.4$, $p < .02$, $\eta^2 = .07$. Tukey's pairwise comparisons indicated that the mean estimate for the story about learning French ($M = 49.3$) was significantly lower than the mean estimate for the story about losing weight ($M = 63.9$), $HSD_{(05)} = 12.3$, $p < .05$. In the fourth and final ANOVA, story (French, nicotine, weight) was used as a between-subject variable and estimate (observed and predicted) was used as a repeated measure. The main effect of story was significant, $F(2, 111) = 4.13$, $MSE = 906.9$, $p < .05$, $\eta^2 = .07$. More important, the main effect of estimate was not significant, $F(1, 111) = 1.2$, $MSE = 291.5$, $p = .27$; nor did story interact with this variable, $F(1, 111) = 1.2$, $p = .29$.

### Experiment 3

A three-way log-linear analysis was calculated using story (French, nicotine, or exercise), judgment (predictive or diagnostic), and choice (causes, disablers, both, or neither) as between-subject variables. As predicted, a statistical dependency between judgment and choice was observed, $G^2_{(3)} = 47.4$, $p < .001$. This dependency was qualified by an interaction with story, $G^2_{(17)} = 62.7$, $p < .001$, and a significant simple interaction of story with choice at each judgment, $G^2_{(27)} = 25.0$, $p < .05$. The source of these statistical qualifications was the dearth of "neither" selections for the exercise story ($n = 5$) compared to the nicotine ($n = 18$) and French ($n = 19$) stories. Because this result carries no theoretical significance and the pattern of distribution of other selections were the same across the three stories, the story factor was collapsed in subsequent analyses.

Received September 30, 2013
Revision received March 27, 2014
Accepted April 6, 2014 ■