Psychonomic Bulletin & Review
2007, 14 (4), 635-639

The effect of causal knowledge on judgments of the likelihood of unknown features

Caroline Proctor and Woo-kyoung Ahn
Yale University, New Haven, Connecticut

People frequently infer unknown aspects of an entity based on their knowledge about that entity. The current study reports a novel phenomenon, an inductive bias people have in making such inferences. Upon learning that one symptom causes another in a person, both undergraduate students (Experiment 1) and clinicians (Experiment 2) judged that an unknown feature associated with the cause-symptom was more likely to be present in that person than an unknown feature associated with the effect-symptom. Thus, these findings suggest a specific mechanism in which causal explanations influence one's representation of and inferences about an entity. Implications for clinical reasoning and associative models of conceptual knowledge are discussed.

People frequently use known features of an entity (e.g., Jan has insomnia) to infer unknown properties of that instance (therefore she must be tired). Such inferences are exceedingly common and likely to be incorporated into our representations of entities. Yet, any inductive biases[1] people may have in such cases have been rarely studied. Previous literature on induction has instead focused on what is often known as "category-based induction," in which people judge whether a feature seen in one category (e.g., robins have sesamoid bones) can be generalized to another category (e.g., therefore sparrows have sesamoid bones) (e.g., Rips, 1975).

Although both tasks are inductive inferences, different mechanisms are likely at work. When generalizing a known feature to another instance in category-based induction (e.g., given that robins have sesamoid bones, what other animals have sesamoid bones?) factors such as the coverage, similarity, or relevant ecological, thematic, or causal relations between categories can have a large effect (e.g., Medin, Coley, Storms, & Hayes, 2003). However, these factors pertain to multiple categories, and thus they are clearly not at work in the prediction of novel features within the same instance (e.g., given that robins have sesamoid bones, what else do robins have?).

To study inductive biases in inferring unknown features, the current work examines the role of causal explanations. We hypothesize that unknown features are judged to be more likely when associated with causes than when associated with effects. Consider an entity with two features (X and Y); the hypothesis is that those who believe that X causes Y would be more likely to draw inferences about the entity from X than from Y, whereas those who believe that Y causes X would be more likely to draw inferences from Y than from X. For example, suppose Helen feels her life is empty, which causes her to be excessively devoted to her job. In this case, it is not difficult to infer that she is lonely and sad. However, if one believes that Helen is excessively devoted to her work, which causes her to feel a void in her life, one might be more inclined to infer that she is ambitious and single-minded.

This hypothesis that we are more willing to draw inferences from causes than effects has not been previously tested. Nonetheless, literature from other domains highlights the salience of causes in a number of contexts and illustrates how they are often weighed more heavily than effects. In the domain of physical reasoning, people overestimate the strength and importance of cause objects and underestimate or neglect forces exerted by effect objects (White, 2006). In inference tasks, Tversky and Kahneman (1980) found that people infer effects from causes (e.g., estimating a son's height from his father's) with greater confidence than causes from effects, even when each gives the same amount of information about the other. Cause features have also been shown to be more central to categorization judgments than effect features across various domains (Ahn, 1998).

Causes may appear more important than effects because they constrain the meaning and interpretation of their outcomes. To return to the example of Jan, if her tiredness is caused by her having stayed up all night talking with her date, her fatigue is interpreted differently than if it is caused by insomnia or having run a marathon the day before. A cause thus has epistemological power as it provides an explanation for why a phenomenon occurs, whereas an effect is less likely to serve such a role.

Indeed, the extensive literature on attribution theory shows how knowledge about different types of causes has differing impacts on people's reactions and behavior (e.g., Weiner, 1986; Ahn, Novick, & Kim, 2003). For example, attributing someone's crime to internal traits as opposed to situational

C. Proctor, caroline.proctor@yale.edu

Table 1
Stimuli Used in Experiments 1 and 2

	To-Be-Inferred Features	
Known Symptoms (X and Y)	X-Associated Feature	Y-Associated Feature
J.M. has large mood swings (X) and excessive social anxiety (Y).	engages in reckless behavior	tendency to blush
S.S. has recurrent suspicions about her husband's fidelity (X) and requires excessive attention (Y).	has doubts about the loyalty of her friends	fishes for compliments about her appearance
B.L. frequently lies (X) and eats in binges (Y).	is manipulative of others	is dissatisfied with her body weight
W.R. reads malicious meanings into benign remarks (X) and fears being left to take care of himself (Y).	has difficulty trusting others	engages in submissive behavior
C.H. always chooses solitary activities (X) and has a lack of empathy (Y).	is shy	exploits others
H.V. devotes herself to work to the exclusion of friendships and leisure (X) and has chronic feelings of emptiness (Y).	pays extraordinary attention to checking for possible mistakes	performs impulsive, harmful actions such as self-mutilation
L.S. fails to plan ahead (X) and has difficulty doing things on her own (Y).	lacks a realistic concern about future problem	engages in clinging behavior

influences drastically changes perceptions of the nature of their actions and subsequent reactions and penalties.

Whereas the aforementioned works suggest that induction would be more influenced by causes than effects, theories and models based on purely associative mechanisms of cognition (e.g., Allan, 1993; Shanks, Lopez, Darby, & Dickinson, 1996; Rogers & McClelland, 2004) would not predict such effects. Associative models argue that causal knowledge is acquired through a domain-general learning mechanism, and represented in terms of patterns of covariation (but see Waldmann & Holyoak, 1992, for an argument that the direction of causal relations influences the learning of associations between events). For example, although previous work has shown that people weigh cause features more than effect features (e.g., Ahn, 1998), Rogers and McClelland (2004) argue that these findings were obtained because the specific cause features used in these studies were consistently associated with more features of the target concept than the effect features. According to this account, if background knowledge and associations are controlled, there should be no difference in inductive potency between cause and effect features.

To the contrary, we argue that people *are* inherently sensitive to the direction of causal relations among known properties, and that people are more willing to make inductive inferences from causes than effects within the same entity. Since the literature reviewed above suggests that cause features are more central to an object's representation than effect features, it follows that they are therefore better bases of inference. In fact, studies on category-based induction have shown central features of a category are more likely to be generalized to other categories (Hadjichristidis, Sloman, Stevenson, & Over, 2004). However, no previous studies have examined how such causal centrality can influence inferences about other unknown features, which we investigate here with two experiments. We took special care to control the associations between known features and to-be-inferred features in order to demonstrate that the phenomenon is due to causal links per se, and not from the differences in associations.

EXPERIMENT 1

To examine how causal relations between features affect induction of unknown properties, we presented participants with a series of scenarios in which a person had two causally related mental disorder symptoms (X and Y). Half the participants were told that Symptom X caused Symptom Y, and the other half were told that the Symptom Y caused Symptom X. Participants in both conditions then made two judgments: they inferred the likelihood of the target person having a new feature associated with Symptom X but not Symptom Y, and they also judged the likelihood of the target having a new feature associated with Symptom Y but not Symptom X. It was predicted that participants would rate these novel features more likely when they were associated with what they had been told was the cause than the symptom they had been told was an effect.

Pretest 1

In order to identify causally reversible symptom pairs, twenty initial pairings of diagnostic criteria or descriptions of mental disorders that seemed bidirectionally plausible were taken from different disorders in the DSM–IV (American Psychiatric Association [APA], 1994). A group of 10 undergraduate participants evaluated the plausibility of the causal relations in two counterbalanced blocks on a 1 (*very implausible*) to 7 (*very plausible*) scale. Ten symptom pairings that received a mean plausibility rating of greater than 4.5 in both causal directions were identified. The overall mean plausibility was 5.4, significantly greater than the midpoint [$t(19) = 8.80$, $p < .01$]. Seven of these 10 pairings were selected for the main experiment through a second pretest described below.

Pretest 2

In order to identify a set of to-be-inferred features associated with the X or Y symptoms, a second pretest was performed. For each of the 10 X–Y pairs chosen above, eight candidate features were initially generated. Features were intended to be intuitively associated with one symptom from the pair *but not the other*. A group of 14 undergraduates rated on 160 trials how strongly a candidate feature was associated with the relevant Symptom X and Symptom Y on a 21-point pretrained scale, spanning from −10 (*extremely negatively*) through 0 (*not at all*) to +10 (*extremely positively*). From these data, we selected seven X–Y pairings in which a candidate feature was judged to be associated only with X, and another only with Y. The overall mean ratings for the to-be-associated features ($M = 7.2$, minimum = 5.5) was significantly higher than the overall mean ratings for the not-to-be-associated features ($M = 0.8$, maximum = 3.2) [$t(13) = 18.29$, $p <$ .01]. The final set of symptoms (Symptom X, Symptom Y, X-associated feature, and Y-associated features) are shown in Table 1.[2]

Method

Forty undergraduate participants at Yale University completed the study for experimental credit or $5. Participants were given seven scenarios, each one describing a person who had two mental disorder symptoms (X and Y), shown in Table 1. For example, all participants were given a scenario "H.V. has chronic feelings of emptiness and devotes herself to work to the exclusion of friendships and leisure." Half were then told that X causes Y, (i.e., "her chronic feelings of emptiness cause her to devote herself to work to the exclusion of friendships and leisure") while the other half were told that Y causes X (i.e., "her devotion to work to the exclusion of friendships and leisure causes her to have chronic feelings of emptiness.")

For each scenario, participants carried out two tasks. First, participants were asked to generate an explanation and elaborate as to why one symptom would cause the other as specified, and rate how believable their own explanation was on a 1 (*very implausible*) to 7 (*very plausible*) scale. This was done to ensure that participants deeply processed the appropriate causal direction between the symptoms, and to check whether they believed this relation.

Second, participants made separate judgments about two features isolated in Pretest 2: one about the likelihood of a feature associated with Symptom X but not Symptom Y, and one about the likelihood of a feature associated with Symptom Y but not Symptom X. For example, for a feature associated with feelings of emptiness but not devotion to work, they were asked, "How likely do you think it is that H.V. performs impulsive harmful actions like self-mutilation?" For half the participants who learned that H.V.'s feelings of emptiness were caused by devotion to work, this feature was coded as effect associated, whereas for the other half who learned the opposite causal relation, it was coded as cause associated. For a feature associated with devotion to work and not feelings of emptiness, they were asked, "How likely do you think it is that H.V. pays extraordinary attention to checking for possible mistakes?" Again, for half the participants, this question was coded as cause associated, and for the other half, it was effect-associated. They made judgments on a 0 (not at all) to 10 (completely certain) scale.

For each version of the questionnaire (X-caused-Y version and the Y-caused-X version) there were four random orders of items. Within these two versions the order in which the symptoms were initially mentioned and the order in which the two judgments were made was counterbalanced.

Results and Discussion

First, examining participants' plausibility ratings of their own explanations, only two explanations of the 280 solicited in the experiment were given a rating of "very implausible," and 83% of the explanations were rated four or higher. The mean rating was 4.86 indicating that, on the whole, people believed the manipulated causal relationships and were able to construct plausible explanations of them.

Next, the likelihood ratings from the induction task showed that, as predicted, causal status did play a role. The same features were rated more likely when they were cause-associated ($M = 7.75$) than when they were effect-associated ($M = 7.19$) [$t(39) = 3.09$, $p < .01$, $d = .49$] (see Figure 1). Since across the conditions, the same feature appeared both as a cause-associated and an effect-associated feature, all potential effects of background knowledge were controlled and associative strengths held constant. For instance, when devotion to work was described as a cause of feelings of emptiness, the mean ratings that participants gave for the likelihood judgment that this person would pay attention to checking for possible mistakes was 9.10, whereas it was only 7.25 when feelings of emptiness were described as a cause for devotion to work. Also, experimental demand was minimal in this study, as although participants were told explicitly that one symptom caused the other and to elaborate on this relationship, it was in no way obvious how this should translate to the induction task, especially as participants were not directly told the induced features were related to the initial symptoms.

EXPERIMENT 2

The current experimental task roughly captures a situation that professional mental health clinicians encounter in dealing with clients. They observe symptoms, notice or assume the causal relations between them (Kim & Ahn, 2002), and then they may infer unobserved symptoms based on this information. Prior work with experts has shown that they tend to use considerable domain-specific causal knowledge, at least with category-based induction tasks (Proffitt, Coley, & Medin, 2000; López, Atran, Coley, Medin, & Smith, 1997). Our second experiment therefore examined whether expert clinicians who commonly make such inferences in this domain would also be inclined to rely more on a cause-associated feature.

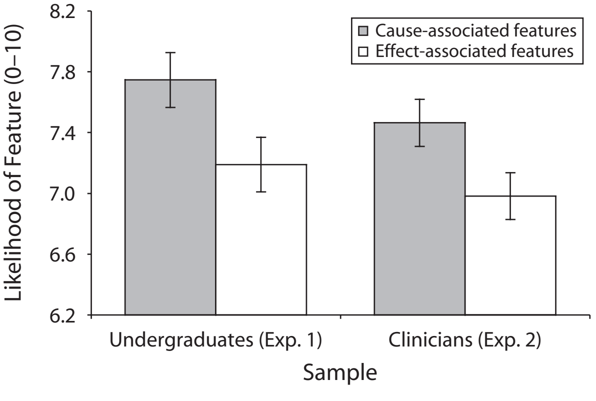

Figure 1. Judged likelihood of features associated with the cause and effect symptoms in undergraduates (Experiment 1) and professional mental health clinicians (Experiment 2).

Method

Mailing addresses of mental health professionals were obtained from state licensing boards of Arizona, Florida, Maryland, Virginia, and Washington, DC. Clinicians who had been licensed for at least 10 and at most 30 years were invited to participate in an online study. Forty clinicians (10 psychiatrists, 18 clinical psychologists, and 12 clinical social workers) completed the online questionnaire for $40. The materials and procedure were identical to Experiment 1.

Results

The mean plausibility rating of clinicians' explanations was 5.25, even higher than for students. Additionally, only four explanations of the 280 were given a rating of "very implausible," and 88.2% of the explanations were rated above the midpoint showing that our clinician participants had considered and believed the manipulated causal relationships.

As with undergraduates, clinicians judged cause-associated features ($M = 7.46$) to be more likely than effect-associated features, ($M = 6.98$) [$t(39) = 3.12, p < .01, d = .49$] (see Figure 1). A two-way ANOVA examining the effect across subject populations (i.e., comparing with Experiment 1) yielded only the expected main effect of causal status [$F(1,78) = 19.14, p < .001, \eta^2_{\text{p}} = .20$] with no significant interaction [$F(1,78) = .10, p = .75, \eta^2_{\text{p}} = .001$] and no significant effect of expertise [$F(1,78) = 1.17, p = .28, \eta^2_{\text{p}} = .02$].

GENERAL DISCUSSION

Two experiments found a novel inductive phenomenon: People are more likely to rely on a feature serving as a cause than one serving as an effect when inferring unknown features of an entity. Experiment 1 found that features associated with symptoms serving as causes in person descriptions were judged more likely than those associated with effect symptoms. This effect was found even though the same symptoms were employed in the roles of causes and effects, and the same features were used as being cause- and effect-associated.

These results are consistent with other findings indicating the salience and weight given to causes in reasoning, as discussed in the introduction. They also demonstrate the insufficiency of purely associative network model accounts of semantic cognition (e.g., Allan, 1993; Shanks et al., 1996), which are not sensitive to the directions of causal associations and could not have predicted these results. Moreover, this work in particular addresses recent arguments of Rogers and McClelland (2004) that previous demonstrations of causal influences on reasoning (e.g., Keil, 1989; Gopnik & Sobel, 2000) are due to differing background knowledge.

For example, in their account of a study by Ahn (1998) in which features described as causes were found to be more important in determining category membership than those same features when described as effects, they point out that different knowledge and associations could have been brought to bear when causal relations were reversed which could account for these results. In the case of one item involving a flower, they noted that when a chemical in the flower was described as a cause for attracting bees, this chemical compound was construed as a product of that flower, namely something that would be strongly associated with many features of the flower. However, in the other condition, where the chemical compound was described as an effect (a residue left by visiting insects), the chemical compound would not be associated with many features of the flower. Thus, they argued that different background knowledge, or associations could thus be brought to bear in the two conditions, and the associative mechanism could account for such differences.

The current study, however, is highly unlikely to have suffered from these problems. First of all, features that served as causes and effects were all symptoms of disorders, and not different in kind as in some previous studies (e.g., internal vs. appearance features in Keil, 1989; compositional vs. functional features in Ahn, 1998). Second, in our experiments we reversed the causal relations among identical symptoms *without providing additional mechanisms* that could have changed the number of associated features. Causal status was the only thing that differed between the two conditions. Knowing a feature is a cause must therefore force us to weigh it more heavily in inferences than when not in this causal role; this is precisely the inductive bias we are capturing that is difficult to account for by associative mechanisms alone.

However, a possible alternative explanation for our results is a phenomenon known as causal discounting: when informed that a given cause of behavior is present, people tend to see an alternative cause as less likely (Kelley, 1972). Thus, if the effect-associated feature happened to be an alternative cause of the effect, it would have been discounted. That is, this feature would be judged less likely than the cause-associated feature due to causal discounting and not the diminished inductive potency of effects. A follow-up experiment with 18 undergraduates did find that four of the 14 induced features in Experiment 2 were viewed as plausible causes of the effect symptoms; however, causal discounting cannot account for our results because removing the affected inductions from the analysis did not affect the difference between cause-associated and effect-associated features [$t(39) = 4.99, p < .01, d = .79$], nor did removing the items completely in order to balance the design [$t(39) = 2.22, p < .05, d = .35$].

Finally, Experiment 2 found that the inductive impact of causes generalizes beyond undergraduates to professionals trained to reason about mental disorder symptoms. These results have practical implications as they illustrate how clinicians' causal inferences about patients' symptoms will affect their inferences about other unknown features they may have. Although clinicians are trained using the DSM-IV which takes an atheoretical approach to mental disorders and describes them in terms of combinations of equally weighted symptoms (APA, 1994), the effect of causal status we observed fits well with previous work showing that clinicians' diagnoses are influenced by rich causal theories they have about mental disorders (Kim & Ahn, 2002).

That we induce unknown features more readily from causes than from effects is consistent with a wealth of research emphasizing the importance of causes to concepts and categorization (Ahn, 1998; Keil, 1989). Murphy (2002) summarized the large body of work on category-based induction as illustrating how inductions form a continuum of causal explanations, from general to more specific ones, with similarity as a fallback strategy when specific explanations are lacking. The current experiments investigating induction within an entity build on this literature and suggest that the use of causal information is a widespread strategy in human reasoning.

One remaining question, however, is what particular relationship drives the greater likelihood of cause-associated features. Although we pretested our to-be-inferred features to be simultaneously associated with one symptom and not associated with the other, this double manipulation leaves open the possibility that only one of these relationships is necessary to produce this effect. It may be that the reason effect-associated features were rated less likely was not due to their association with the effect symptom, but because of their disassociation from the cause symptom. We leave this question open for future work to address.

In conclusion, the present study found a novel inductive phenomenon that people exhibit in making inferences about an entity based on one's knowledge about that entity. Although it seems quite intuitive that people would frequently make inferences about an entity based on known information about it, the inductive bias behind such inferences has been rarely studied. Future study can further examine consequences of cause-based induction, as well as revealing other inductive tendencies that people might exhibit in inferring from the known.

AUTHOR NOTE

Support for this research was provided in part by National Institute of Mental Health Grant MH57737 to W.-K.A. The authors thank Christian Luhmann and Jessecae Marsh for helpful comments and suggestions. Correspondence concerning this article should be addressed to C. Proctor, Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520-8205 (e-mail: caroline.proctor@yale.edu).

REFERENCES

Ahn, W.-K. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, **69**, 135-178.

Ahn, W.-K., Novick, L., & Kim, N. S. (2003). "Understanding it makes it normal": Causal explanations influence person perception. *Psycho nomic Bulletin & Review*, **10**, 746-752.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative*? Psychological Bulletin*, **114**, 435-448.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychological Association.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, **71**, 1205-1222.

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science*, **28**, 45-74.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kelley, H. H. (1972). *Causal schemata and the attribution process*. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151-174). Morristown, NJ: General Learning Press.

Kim, N. S., & Ahn, W.-K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, **131**, 451-476.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, **32**, 251-295.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, **10**, 517-532.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 811-828.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning & Verbal Behavior*, **14**, 665-681.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, D. L. Medin, & K. J. Holyoak (Eds.), *Psychology of learning and motivation* (Vol. 34, pp. 265-311). San Diego: Academic Press.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49-72). Hillsdale, NJ: Erlbaum.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, **121**, 222-236.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, **113**, 132-147.

NOTES

1. In this article, we use the term "bias" in a neutral way without implying that it is necessarily irrational. The rational basis of the phenomenon studied in this paper needs further investigation.

2. Given the high associative ratings between the to-be-induced features and the known (X and Y) symptoms, one might argue that the to-be-induced features were not necessarily *induced* in the main experiment because they were merely restatements or examples of the known features. For instance, one might argue that the feature "being shy" is essentially the same as the symptom "chooses solitary activities." However, although features were selected to be highly associated with one of the symptoms, all of the mean ratings were significantly less than the maximum value of 10 (all p values $< .01$). Second, symptoms and features were taken from diagnostic criteria of disorders in the DSM-IV manual, and thus there seems to be some professional clinical consensus that these represent meaningfully different descriptions of people's behavior. For example, many shy people are not necessary solitary and are quite happy with close friends and family. Similarly, those who are socially capable often pursue solitary activities for relaxation or concentration.

(Manuscript received July 17, 2006;
revision accepted for publication October 19, 2006.)