Cognition 139 (2015) 92–104

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

A Bayesian framework for knowledge attribution: Evidence from semantic integration ☆

Derek Powell [a,*], Zachary Horne [b], N. Ángel Pinillos [c], Keith J. Holyoak [a]

[a] Department of Psychology, University of California, Los Angeles, United States
[b] Department of Psychology, University of Illinois at Urbana-Champaign, United States
[c] Department of Philosophy, Arizona State University, United States

ARTICLE INFO

Article history:
Received 1 November 2013
Revised 9 February 2015
Accepted 1 March 2015

Keywords:
Knowledge
Bayesian reasoning
Implicit memory
Semantic integration
False memory

ABSTRACT

We propose a Bayesian framework for the attribution of knowledge, and apply this framework to generate novel predictions about knowledge attribution for different types of "Gettier cases", in which an agent is led to a justified true belief yet has made erroneous assumptions. We tested these predictions using a paradigm based on semantic integration. We coded the frequencies with which participants falsely recalled the word "thought" as "knew" (or a near synonym), yielding an implicit measure of conceptual activation. Our experiments confirmed the predictions of our Bayesian account of knowledge attribution across three experiments. We found that Gettier cases due to counterfeit objects were not treated as knowledge (Experiment 1), but those due to intentionally-replaced evidence were (Experiment 2). Our findings are not well explained by an alternative account focused only on luck, because accidentally-replaced evidence activated the knowledge concept more strongly than did similar false belief cases (Experiment 3). We observed a consistent pattern of results across a number of different vignettes that varied the quality and type of evidence available to agents, the relative stakes involved, and surface details of content. Accordingly, the present findings establish basic phenomena surrounding people's knowledge attributions in Gettier cases, and provide explanations of these phenomena within a Bayesian framework.

1. Introduction

In everyday life, it is often vital that we draw accurate distinctions between what we know, and what we merely believe. Whereas knowing may license action (Hawthorne & Stanley, 2008), lack of knowledge calls for caution and consideration of more evidence. Moreover, we continually need to evaluate *other people's* knowledge. For example, when someone harms us, assignment of blame often involves assessing whether that person knew that their actions would have harmful consequences (e.g., Young & Saxe, 2011). Evaluating knowledge requires an understanding of what it means to know, raising an important psychological question: what is people's concept of knowledge?

Recent psychological research on this question (Nagel, San Juan, & Mar, 2013; Starmans & Friedman, 2012; Turri, Buckwalter, & Blouw, 2014) has taken inspiration from philosophical analyses of knowledge. Philosophers once commonly accepted that knowledge is *justified true belief* (JTB; Ayer, 1956; Plato, 1961). But recently, many epistemologists have rejected this analysis in light of a

☆ A preliminary report of Experiments 1 and 2 was presented at the 35th Annual Conference of the Cognitive Science Society (Berlin, August 2013).

* Corresponding author at: Department of Psychology, University of California, Los Angeles, 1285 Franz Hall Box 951563, Los Angeles, CA 90095, United States.
E-mail address: derekpowell@ucla.edu (D. Powell).

http://dx.doi.org/10.1016/j.cognition.2015.03.002
0010-0277/

class of thought experiments now known as *Gettier cases* (Cohen, 1998; Greco, 2003; Lewis, 1996; Sosa, 2007; Turri, 2011; Williamson, 2002; Zagzebski, 1996). Gettier cases (named after their originator; Gettier, 1963) are situations in which an agent holds a justified true belief, but unexpected elements of the situation (allegedly) prevent the agent from truly "knowing." Psychological investigations of people's evaluations of Gettier cases may serve as a revealing window into people's concept of knowledge and the basis for their knowledge attributions.

A number of different types of Gettier cases have been discussed in the philosophy literature (e.g., Fantl & McGrath, 2009; Goldman, 1976; Sturgeon, 1993; Turri, 2011). In the present paper we focus on two major classes of such cases, which respectively concern (1) the replacement of items or evidence, and (2) the presence of counterfeit objects.

In *replaced evidence* cases, an agent encounters what appears to be direct evidence for a belief, but which is actually a copy of the original evidence or a similar substitute.[1] For example, suppose a young man named Will commits a crime, and then covers his tracks by destroying the evidence that would have implicated him. Unfortunately for Will, his enemy Beth is aware of his crime and plants evidence to ensure that he is caught anyway. A detective investigates the crime and finds the planted evidence, which leads him to believe that Will committed the crime. The detective's belief is true and is justified by the evidence he found. Consequently, on the JTB account of knowledge, the detective knows Will is guilty. However, many philosophers claim that the detective does not know that Will is guilty (for discussion of an analogous case, see Fantl & McGrath, 2009).

Another type of replacement case involves the replacement of the subject of an agent's belief, which Turri et al. (2014) have labeled "replacement-by-backup" cases. For instance, suppose a woman places a pen on a table in her apartment and then steps into the shower. Then, a burglar silently steals the pen and replaces it with another identical pen. After the burglar leaves, the woman still (correctly) believes there is a pen on her table, yet most philosophers conclude she does not know this fact (e.g., Sturgeon, 1993; Turri, 2011; Williams, 1978).

The second class of Gettier case we will consider are those due to *counterfeit objects*.[2] For example, imagine a mother and her young son are driving along a country road. As they drive, the mother is pointing out the window and labeling the things they see for her child's benefit. At one point she sees a barn and says, "That's a barn." Unbeknownst to the mother, the residents along this strip of highway have erected several facades that look exactly like real barns. The barn she is looking at is actually the only real barn for miles, and from the road she would have no way of distinguishing between it and the facades. In fact, it was by sheer luck that she ended up pointing at a real barn. Her belief is true, and it is justified by her perceptual experience of the barn. Yet, it has been claimed that she does not know that she is pointing at a barn (e.g., Goldman, 1976).

Though many philosophers have argued that agents in Gettier cases do not have knowledge (e.g., Fantl & McGrath, 2009; Goldman, 1976; Sturgeon, 1993; Turri, 2011; Williams, 1978), psychological investigations of laypeople's judgments about such cases have produced intriguing, if sometimes inconsistent, results (e.g., Colaço, Buckwalter, Stich, & Machery, 2014; Nagel, San Juan, et al., 2013; Starmans & Friedman, 2012; Turri et al., 2014; Wright, 2010). Some of these findings stand in contrast to philosophers' intuitions. Starmans and Friedman (2012) found that participants tended to attribute knowledge in "replacement-by-backup" Gettier cases almost as readily as in standard cases of justified true belief. Similarly, Turri et al. (2014) found that people also attributed knowledge to agents in counterfeit-object cases. However, Turri et al. also report an experiment in which participants distinguished between a replacement-by-backup case and a standard JTB case, contradicting the findings reported by Starmans and Freidman. Unlike Turri et al., Colaço et al. (2014) did observe differences between rates of knowledge attribution in counterfeit-object and JTB cases (although participants' ratings appear to have weakly favored knowledge attribution for both types of cases). Finally, Nagel, San Juan, et al. (2013) examined a variety of cases, including "replacement-by-backup", replaced-evidence, and counterfeit-objects cases. Averaging across these different cases, they found that people tended to deny that agents knew (although this claim is disputed by Starmans & Friedman, 2013). Altogether, there seem to be few points of agreement among these findings: both replacement and counterfeit-object Gettier cases have been found to elicit knowledge attributions in some experiments (Starmans & Friedman, 2012; Turri et al., 2014), but not in others (Colaço et al., 2014; Nagel, San Juan, et al., 2013; Turri et al., 2014).

Setting aside these inconsistencies, there are at least two problems with extant research on knowledge attribution. First, there is an unresolved methodological debate over the best way to probe participants' knowledge attributions (Nagel, Mar, et al., 2013; Starmans & Friedman, 2013). Existing research has relied on explicit survey-like questions for assessing knowledge attributions, but the reliability and validity of these measures have not been established. Methodological issues thus offer one possible explanation for the lack of agreement among the findings of different researchers. Later in the present paper, we discuss these methodological issues further, and report three experiments that begin to address these concerns.

Second, there is no clear theoretical context within which to interpret empirical findings regarding laypeople's reactions to Gettier cases, or from which specific predictions can be generated about their expected behavior. For example, only Turri et al. (2014) have drawn a clear distinction between replacement cases and counterfeit-object cases, but even these researchers have not examined how this distinction might be explained, or what this distinction implies about the lay concept of knowledge. Without any overarching theoretical framework, it is unclear how

[1] Elsewhere these types of cases have sometimes been referred to as "false lemma" cases (e.g., Nagel, Mar, and San Juan, 2013; Nagel, San Juan, et al., 2013).

[2] Many philosophers have referred to cases of this sort as "fake barn" Gettier cases (after Goldman, 1976), but we believe it is useful to introduce more general terminology that is less dependent on an incidental example.

people's knowledge attributions in Gettier cases should inform our understanding of their knowledge concept. In the next section, we attempt to address this concern by developing a theory of knowledge attribution based on a Bayesian framework.

1.1. A Bayesian analysis of knowledge attribution

In studies to date, researchers have generally sought to compare laypeople's knowledge attributions in Gettier cases to philosophers' intuitions about these cases (Nagel, Mar, et al., 2013; Nagel, San Juan, et al., 2013; Starmans & Friedman, 2012; Turri et al., 2014). In spirit, this approach accords well with a common approach in cognitive science, where human behavior in a task is compared with a computational account that establishes the ideal or optimal behavior in that task (Marr, 1982). However, it is unclear whether philosophers' intuitions about these cases actually reflect some well-defined optimal evaluation.

The model we propose distinguishes between the viewpoint of an *agent* (typically a character described in a vignette, who receives information and makes inferences) and the *observer* (typically the participant in an experiment, who reads the vignette and receives additional information not available to the agent). In the cases considered in the literature on knowledge attribution, agents' beliefs are justified inductively (rather than deductively) from evidence they observe. Thus, agents' beliefs and observers' evaluations of those beliefs might be understood within a probabilistic or Bayesian framework. In recent years, Bayesian models have been applied across a wide a variety of psychological tasks, including visual perception (Yuille & Kersten, 2006), causal inference (Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Powell, Merrick, Lu, & Holyoak, 2013, 2014), informal argumentation (Hahn & Oaksford, 2007, 2012), moral reasoning (Rai & Holyoak, 2014), and various estimation tasks (Griffiths & Tenenbaum, 2006). Within this framework, inferences are "ideal" insofar as they accords with the axioms of probability. This sense must be carefully distinguished from being "ideal" in the normative sense of correctly identifying cases of knowledge.

We assume that three factors jointly determine whether an observer views the agent described in a story as "knowing" that a hypothesis is true: (1) *truth*: the observer must be told or infer that the agent's hypothesis is true; (2) *agent's posterior*: the agent's posterior probability must be assessed by the observer to be higher than some contextually-determined threshold; and (3) *consistency*: the amended posterior probability that the observer computes based on their additional information must also meet or exceed the contextually-determined threshold. In calculating the agent's and the observer's posteriors, we assume the observer screens off direct knowledge that the conclusion is true or false (usually stated directly in the vignette). Thus the posterior probabilities will typically involve some degree of uncertainty, rather than taking on values of 0 or 1. Of course, observers may not entirely succeed in screening off direct knowledge of truth when they calculate the agent's posteriors and their own amended posteriors given phenomena such as hindsight bias (Fischhoff, 1975) and the so-called "curse of knowledge" (e.g., Birch & Bloom, 2007). However, such errors are unlikely to alter the relative order of knowledge attribution across conditions, on which we focus in deriving predictions.

The requirement of truth is assumed because verbs such as "know" are by definition factive. The Bayesian component of our model involves the second and third constraints (an agent's posterior and consistency), which reflect the inferential processes used to assess whether or not an inference is accepted as true. Importantly, all three requirements are assumed to be necessary for knowledge attribution. Thus meeting the thresholds of certainty required by agent's posteriors and consistency does not result in knowledge attribution if the conclusion is known to be false. Accordingly, the Bayesian model is applicable to the constraints of agent's posteriors and consistency, but the model as a whole (including the requirement of truth) is not fully Bayesian.

Within a Bayesian framework, an agent's confidence in a belief should be expected to vary with the posterior probability of the relevant hypothesis. The posterior probability in turn depends on the integration of (1) the likelihood of the observed evidence given the target hypothesis and each alternative hypothesis, and (2) the prior probability of each hypothesis, as proscribed by Bayes' rule:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

We assume that the observer (i.e., a participant in an experiment who is presented with a vignette describing evidence and an agent's conclusion from it) will use Bayesian reasoning to evaluate the beliefs of the agent (i.e., the cognizer in the vignette who interprets the stated evidence). First, the observer must consider the evidence available to the agent, and how that evidence drives the agent's belief. This process will require some assessment of the likelihood and prior probability functions that the agent employs, respectively $P_A(d|h)$ and $P_A(h)$. Typically, we expect that the observer assumes the *agent's posterior* $P_A(h|d)$ would be the same as their own, given the same evidence, although this assumption is likely defeasible. The agent can only be considered to "know" if this posterior exceeds their contextually-determined threshold. Having evaluated the agent's posterior, the observer can then revaluate the situation by taking into account the additional information provided in the vignette (of which the agent is unaware) in order to calculate the *observer's amended posterior*, $P_O(h|d)$. Finally, they can assess whether the agent "knows" by ensuring that both the agent's posterior and their own amended posterior fall above the contextually determined threshold.

Applying this Bayesian framework to the Gettier cases we discussed earlier yields a set of novel hypotheses about when people will attribute knowledge to an agent. Knowledge attribution in Gettier cases hinges on the consistency condition (since by the very nature of such cases the conditions of truth and of agent's posterior will be

satisfied). Whether or not the consistency condition will be met can be evaluated by examining how each term in Bayes' equation is affected by the events in the Gettier case in comparison to a matched case of justified true belief that has not been Gettiered.

First, let us consider a Gettier case due to the presence of counterfeit objects. Recall the case of the mother and son traveling down a country road, pointing at barns. We assume that agents and observers restrict their hypothesis space to two mutually exclusive and exhaustive hypotheses. That is, they calculate both the probability of their hypothesis (h_1) and of a relevant alternative hypothesis (h_0). In this case, h_1 would represented the hypothesis that the object the mother points to is a real barn, and h_0 would represent the hypothesis that the object is instead a fake barn. Bayes' rule can therefore be expanded as:

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d|h_1)P(h_1) + P(d|h_0)P(h_0)}$$

When the mother points out the barn to her son, the likelihood of her evidence (the visual appearance of the barn) might be high under both h_1 and h_0. However, she believes that the prior probability of the alternative hypothesis (that locals have erected elaborate barn facades) is low. Hence, she computes a high posterior probability that she is pointing at a barn. Although this is true (the condition of truth is satisfied), and her posterior is high (the condition of agent's posterior is satisfied), her calculation is wrong given what the observer recognizes (and thus the consistency criterion is not met). The observer receives additional information not available to the mother: in the context of this particular stretch of road the prior probability of the "fake" hypothesis is very high and the probability of the "real" hypothesis is quite low (i.e., $P_A(h) \neq P_O(h)$). Hence, the observer's amended posterior probability that the mother is looking at a barn is much lower than the agents' posterior, and should not meet the threshold for knowledge. In this case of counterfeit objects, the Bayesian framework predicts that the exceptional circumstances in the Gettier case will lead the ideal observer to deny that the agent knows she is looking at a real barn.[3]

Next, we consider Gettier cases involving the intentional replacement of evidence. Recall the case in which Beth, the enemy of the perpetrator Will, plants evidence to implicate him in the crime he committed. Again there are two relevant hypotheses: Will is either guilty (h_1) or he is innocent (h_0). In the Gettier case, Beth plants evidence similar to the evidence Will left behind in the standard justified true belief case. Accordingly, the likelihood of the data (the incriminating evidence at the scene) under the hypothesis that Will is guilty is similarly high for both the agent and the observer (i.e., $P_A(d|h_1) \approx P_O(d|h_1)$). In addition, Beth only plants the evidence after Will commits the crime. Therefore, the likelihood of the data under the hypothesis that Will is innocent is also unaffected (i.e., $P_A(d|h_0) \approx P_O(d|h_0)$). Similarly, the prior probability that Will committed the crime is unaffected by whatever actions Beth takes after-the-fact (i.e., $P_A(h) \approx P_O(h)$). As every term in the equation remains unchanged, the inductive strength of the detective's belief is unchanged relative to a case of standard justified true belief. Therefore, the consistency condition is satisfied and the Bayesian analysis predicts that observers will typically attribute knowledge to agents in Gettier cases due to intentionally-replaced evidence.

Thus, in contrast to accounts offered by philosophers who have argued that agents generally lack knowledge in Gettier cases (e.g., Fantl & McGrath, 2009; Sturgeon, 1993; Turri, 2011; Williams, 1978), on our account observers should not attribute knowledge to agents in counterfeit-object cases, but *should* attribute knowledge to agents in replaced evidence cases. The present account thus generates novel predictions about people's knowledge attributions.

However, a similar prediction might be derived by a simpler account that assumes observers are sensitive to the operation of *luck*, and refrain from attributing knowledge when they detect that an agent has gotten lucky. The beliefs of agents in counterfeit-object cases happen to be true simply by luck; but in the replaced-evidence case discussed above, one agent intentionally leads another to the truth by replacing evidence. Within philosophy there has been considerable discussion of the role of luck in Gettier cases (Pritchard, 2004; Turri, 2013; Zagzebski, 1996). It is therefore important that we distinguish the predictions of the Bayesian framework we have proposed from this alternative account.

One way this might be accomplished is by examining Gettier cases involving accidental or lucky replacement of evidence. For example, imagine a woman named Sharon is a vegan who has a craving for chocolate chip cookies. Her husband Mark has recently gone grocery shopping and restocked their cookie jar, but Sharon worries that Mark may have purchased non-vegan cookies. Sharon looks for evidence that the cookies are vegan, and finds some vegan cookie packaging in the trash that matches the cookies in the jar. However, this wrapper is not actually from the cookies her husband purchased. In reality, Mark couldn't remember what brand of cookies to purchase, so before shopping he dug through the trash to find an old wrapper, which he then left on top of the trash. Fortunately for Sharon, he did correctly purchase the brand's vegan cookies, so the cookie she is considering eating really is vegan.

Let us compare the probabilities from the agent's perspective to the observer's amended probabilities. First, the cookie is either vegan (h_1) or not (h_0) and the prior probability that each is true will be unaffected by adding the extra information that the cookie wrapper belonged to a different cookie. Therefore, as in the case of intentional replacement, $P_A(h) \approx P_O(h)$. Next, from Sharon's perspective, the probability of finding a vegan cookie wrapper is high under h_1, but low under h_0. However, from the observer's perspective, the probability of Sharon finding a wrapper is high under h_1 but also high under h_0. This last point is crucial. The cookie wrapper Sharon found was left on top

[3] A possible boundary condition would arise if the probability of the agent's data under the alternative hypothesis, $P(d|h_0)$, is sufficiently low. In this case, the appropriate posterior could be relatively unaffected by the presence of counterfeit objects. However, in the case described here it seems reasonable to expect that $P(d|h_0)$ should remain relatively high—that is, it is fairly likely that a barn facade would look like a barn.

of the trash before the cookies were purchased, so it would have been found whether or not Mark made the correct purchase. Thus, in the case of accidental replacement, we conclude that $P_A(d|h_0) \neq P_O(d|h_0)$. Instead, the data is just as likely under h_0 as under h_1—that is, $P_O(d|h_1) \approx P_O(d|h_0)$. In this case, an ideal observer would compute the correct posterior as:

$$\begin{aligned} P(h_1|d) &\approx \frac{P(d|h_1)P(h_1)}{P(d|h_1)P(h_1) + P(d|h_1)P(h_0)} \\ &\approx \frac{P(d|h_1)P(h_1)}{P(d|h_1)[P(h_1) + P(h_0)]} \approx \frac{P(h_1)}{P(h_1) + P(h_0)} \approx \frac{P(h_1)}{1} \\ &\approx P(h_1) \end{aligned}$$

—or simply the prior probability of h_1. To summarize, because the data are equally likely under either hypothesis, the data are uninformative and the posterior probability the observer assigns to the hypothesis should to be equal to the prior probability of the hypothesis.

Examining a case of accidental replacement may allow us to dissociate the predictions of the Bayesian account from those generated by an alternative account focused solely on luck. This is because the degree to which the operation of luck in evidence replacement impugns the agent's knowledge will depend on the prior probability that was assigned to the hypothesis. That is, if the appropriate prior probability is sufficiently high, observers might still conclude that the agent knows—even though the agent has gotten lucky. We test this prediction in Experiment 3.

1.2. Semantic integration paradigm

Testing the Bayesian framework and alternative accounts of knowledge attribution requires a paradigm that is sensitive to differences in observers' assessments of an agent's knowledge, while minimizing extraneous influences on responses. As we have discussed, Starmans and Friedman (2012), Nagel, San Juan, et al. (2013), Colaço et al. (2014), and Turri et al. (2014) have reported conflicting findings regarding people's concept of knowledge. These discrepancies among empirical findings have sparked a methodological debate about the best way to probe participants' intuitive judgments about knowledge (Nagel, Mar, et al. 2013; Starmans & Friedman, 2013). Whereas a number of researchers have measured knowledge ascription (Starmans & Friedman, 2012; Turri et al., 2014), Nagel, San Juan, et al. (2013) measured knowledge denials using a two-stage procedure. First, participants were asked whether the agent knew, didn't know, or whether it was unclear. If (and only if) participants said the agent knew, they were asked a follow-up question that gave them another opportunity to deny the agents' knowledge. Starmans and Friedman (2013) have argued that this procedure biased participants' responses toward knowledge denial, as participants were asked follow-up questions after they ascribed knowledge but not after they denied knowledge. This debate highlights more general limitations of survey-based methods for assessing people's concepts (e.g., various types of demand characteristics), which we discuss in more detail when we discuss the advantages of our approach.

To overcome the limitations of survey methods involving explicit questions about knowledge, we applied an experimental methodology based on *semantic integration,* which provides an implicit measure of conceptual activation. Semantic integration refers to the cognitive process by which units of semantic information are combined to form larger structured representations. Research has shown that sentences read within a larger piece of discourse are not encoded in isolation; rather, their meanings appear to be combined to form a coherent whole (Franks & Bransford, 1974), and that people's memory for meaning tends to be more robust than their memory for specific episodic details (e.g., Sachs, 1967). Thus, when recalling a passage of text after a delay, participants' memories often reflect their semantic interpretations rather than the actual sentences they read. For example, Bransford and Franks (1971) found that after exposure to several interrelated sentences (e.g., "The frog was on the log" and "The fish swam under the log"), participants falsely recognized sentences that expressed a situation that could be inferred from their combination (e.g., "The fish swam under the frog").

Later investigations used a variety of false memory paradigms to examine different aspects of semantic integration (e.g., Flagg, 1976; Owens, Bower, & Black, 1979; Sulin & Dooling, 1974; Thorndyke, 1976). Gentner (1981) explored whether semantic integration occurs within smaller units of meaning, such as individual verbs. To illustrate, consider the relationship between the general verb 'give' and the more specific verb 'pay'. To give some item is to take some action that transfers ownership of that item to a recipient; to pay is a more specific form of giving, in which the giver owes the recipient. In a context where one agent owes another, giving might be incorrectly remembered as paying. Gentner asked participants to read paragraph-length stories that each included a critical sentence with a generic target verb. For instance, one of her stories contained the sentence, "Max finally gave Sam the money." She created two versions of each story, one that contained additional context explaining that Max owed Sam money, and a control story that did not include the additional context. After reading one version of the story, participants performed a recall task in which they were shown the sentence with the target verb "gave" removed, and were asked to fill in the word that had appeared in the story. Gentner found that participants who read the additional contextual information were more likely to falsely recall "paid," as having appeared in the critical sentence than were participants who read the story that did not contain the additional contextual information.

We applied a similar paradigm to examine people's concept of knowledge. We constructed stories containing the verb "thought", and used false recall of the more specific verb "knew" (and near-synonyms) as a measure of the extent to which different contexts instantiate the concept of knowledge. In earlier pilot work we asked participants to read either a vignette about an agent who formed a

justified true belief or a vignette in which the agent's belief was unjustified (Powell, Horne, Pinillos & Holyoak, 2013). When the context in the vignette indicated that the agent's belief was justified and true, 39.7% of participants recalled having read that the agent 'knew', significantly more than when this contextual information was omitted (19.7%). This initial study served as a proof-of-concept for the use of semantic integration tasks to investigate the knowledge concept.

The semantic integration paradigm affords several advantages over traditional survey-based methods (Powell, Horne, & Pinillos, 2014). Although participants' responses to surveys may be reflective of their concepts, they are also likely to reflect downstream decision processes involved in interpreting and responding to the questions under consideration. These processes can be affected by factors specific to the experimental context, such as demand characteristics (Orne, 1962; Weber & Cook, 1972) that can lead participants toward socially-desirable responses. Similarly, the particular phrasings of questions may contain pragmatic cues that unintentionally bias participants' responses (Cullen, 2010; see Schwarz, 1994, for a review). By contrast, in the semantic integration task participants are not asked to make any explicit judgments at all; rather, they simply read and attempt to remember a story. The task is naturalistic, yet its true aims remain concealed from participants. Thus, there is very little chance that the experimental context might bias or influence participants' responses, as there are no explicit questions to provide pragmatic cues and there are no "socially desirable" responses other than correctly recalling the words presented in the story.

Because the semantic integration paradigm is fundamentally a memory task, findings based on this method must be interpreted somewhat differently than those produced by explicit questions. In particular, the semantic integration paradigm only supports comparisons between closely-matched conditions in a between-subjects design. No conclusions can be drawn on the basis of absolute recall rates, as these may be influenced by a variety of factors (e.g., length and complexity of a particular story). Moreover, the responses of an individual participant cannot be treated as a measure of that person's attitudes or evaluations. It is quite possible that a participant in a semantic-integration experiment may think the agent "knows", yet nonetheless correctly recall "thought," simply because they have retained a detailed episodic memory for the specific wording used in the story. Thus, in all experiments we compare the pattern of recall across groups of participants, where each group has read a closely-matched variant of the same basic story.

We now report a series of experiments using the semantic integration paradigm to evaluate our proposed Bayesian framework for knowledge attribution. We examined laypeople's reactions to a Gettier case due to counterfeit objects (Experiment 1) and a different Gettier case due to replaced evidence (Experiment 2). Then, we compared these two types of cases using two new sets of matched vignettes, while also examining Gettier cases due to accidentally replaced evidence (Experiment 3).

2. General methods

2.1. Participants

Participants were recruited online from Amazon's Mechanical Turk (mTurk) work distribution website, and then redirected to a Qualtrics survey software webpage at which experimental procedures were administered. Participants were paid $0.50 for their participation.

2.2. Materials and procedures

Each of the studies reported here followed a similar procedure. Each experiment comprised three main components: an experimental story, a distractor story, and a recall task. The experimental stories described a situation in which an agent formed a belief, and each included a critical sentence stating that the agent "thought" their belief was true. The content of these stories was manipulated between conditions, and participants were each randomly assigned to read one story. The texts of all stories are provided in the Supplemental Online Materials.

After providing basic demographic information and reading instructions, participants read the experimental story corresponding to their assigned condition. After reading, participants completed a distractor task, which involved reading and answering questions about a distractor story. This was an approximately 1000-word selection from a fictional article on gamma ray bursts (Waskan, Harmon, Horne, Spino, & Clevenger, 2014). Importantly, this distractor story did not include the word "knew" or any near synonyms. Timing controls ensured that participants spent an adequate amount of time attending to instructions and the vignettes at each stage of the experiment.

Next, participants advanced to the recall task. Here they were shown four sentences from the experimental story, one sentence at a time. Each sentence was missing one word that was replaced with an underscored blank space. Participants were instructed to type in the word that originally appeared in the story. The critical sentence was always presented first, followed by three non-critical sentences.

2.3. Response scoring

Participants' responses during the recall task were corrected for typos and other minor syntactic errors, and were then classified as either *thought-type* responses or *knew-type* responses. Two scoring procedures were used, and separate analyses were conducted using both sets of scores. Responses that fell outside these categories were excluded from the analyses. In the strict scoring system, only the word "knew" was scored as a knew-type response, and only "thought" and its close synonym "believed" were scored as thought-type responses. A second,

Table 1
Summary of words categorized as knew-type and thought-type responses under the strict and liberal coding systems.

Coding system	Scored response	Participant's response
Strict	Knew-type	knew
	Thought-type	thought, believed
Liberal	Knew-type	knew, was sure, was certain, verified, confirmed
	Thought-type	thought, believed, assumed, felt, decided

more liberal, scoring system was also used, in which responses were categorized as described in Table 1. This liberal coding system represents a compromise, as some of the responses coded as "knew-type" do not strictly entail knowledge (e.g., "was sure"). More generally, the two scoring systems reflect a trade-off between specificity and coverage, with the strict system lending itself to clearer interpretations and the more liberal system making more effective use of all of the data. Fortunately, analyses under the two scoring systems ultimately support the same conclusions.

3. Experiment 1

3.1. Method

In Experiment 1 we examined a Gettier case due to counterfeit objects, adapting our materials from the jewelry-store scenario tested by Nagel, Mar, et al. (2013) and Nagel, San Juan, et al. (2013). We predicted that people would behave according to our Bayesian account. Thus, we predicted that the counterfeit-object Gettier case would activate knowledge concepts significantly less than a standard justified true belief case.

In addition to examining Gettier cases, Nagel, San Juan, et al. (2013) were interested in how "skeptical pressure" could affect knowledge attribution. On their characterization, a situation presents skeptical pressure if it mentions or raises the possibility of error. For example, in a story where a person is shopping for a diamond necklace, skeptical pressure could be introduced by the narrator mentioning that this shopper wouldn't be able to tell a real diamond from a fake just by simply looking at or touching the diamond. In their study, Nagel et al. found that skeptical pressure cases were treated very much like Gettier cases. Just as with Gettier cases, participants were less likely to attribute knowledge in the skeptical pressure cases than in cases of justified true belief, but more likely to attribute knowledge than in cases of false belief. Although we are primarily concerned with examining how Gettier cases might inform our understanding of people's concept of knowledge, we wished to follow up on Nagel, San Juan, et al.'s (2013) examination of skeptical pressure to determine whether this factor is orthogonal to those that lead people to attribute or fail to attribute knowledge in Gettier cases.

In terms of our Bayesian analysis, applying skeptical pressure might have several effects. The language used may indicate to the reader that the likelihood of the agent's evidence would be the same (or similar) under either h_1 or h_0 (since the suggestion is that agent cannot tell real diamonds from fake), thus impacting judgments in much the same way as predicted in the case of evidence replaced by luck. In addition, in a JTB scenario, it is likely that skeptical pressure would also raise the observer's prior on h_0 by suggesting the possibility that fake diamonds might be present. Accordingly, we could expect that skeptical pressure will affect knowledge attributions in both Gettier and JTB scenarios.

3.1.1. Participants

A total of 531 participants (299 female) were recruited for Experiment 1. The mean age of participants was 32 years (SD = 12.1 years).

3.1.2. Materials and design

The vignettes used in Experiment 1 were about a character named Emma who goes shopping for a diamond necklace. We created six versions of this story by crossing three scenarios (justified true belief, Gettier, and false belief) with the skeptical pressure variable (present or absent) introduced by Nagel, San Juan, et al. (2013), creating a 3×2 factorial design. Participants were each randomly assigned to one of these six conditions. This factorial design allowed us to test and compare the effects of these two variables independently, overcoming some of the limitations of prior studies.

In the justified true belief (JTB) conditions, Emma sees a diamond necklace on a tray marked "Diamond Necklaces" and picks it out. She admires the sparkle and luster of the stone, and decides to buy the necklace. In the false belief (FB) scenarios, a dishonest store clerk has secretly replaced nearly all the diamonds in the shop with cubic zirconia fakes, and Emma picks a fake. Thus, her belief that the stone is a diamond is false. In the Gettier scenarios, the store clerk is similarly dishonest, but Emma (by mere chance) picks one of the few real diamond necklaces remaining in the store. Her belief is justified and true, but true only by luck.

We manipulated skeptical pressure by including or omitting an additional passage in the story, stating that "Emma loved jewelry, but she was no expert. For instance, she could not tell the difference between a real diamond and a cubic zirconia fake just by looking at it or touching it."

The critical sentence in each story was, "The price was high, but Emma *thought* the stone was a diamond, so the price seemed fair" (italics added here). The structure and wording of each story was identical, save for the relevant manipulations.

3.2. Results and discussion

Applying our strict coding system, 491 of the original 531 responses were scored. An additional 13 responses were coded under the liberal coding system, for a total of 504. The results of our analyses under the two coding systems were generally in agreement, and are reported together.

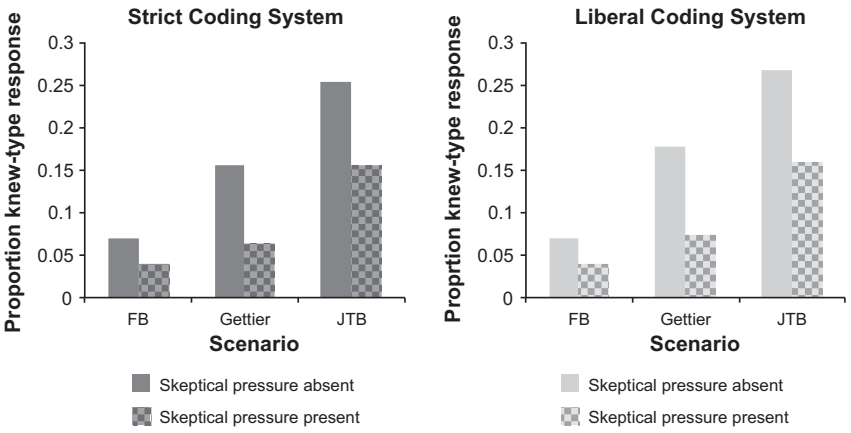

Fig. 1. Proportion of knew-type responses for each type of scenario under both strict (left) and liberal (right) coding systems (Experiment 1). FB = false belief; Gettier = counterfeit-object Gettier; JTB = justified true belief.

Fig. 1 shows the proportion of knew-type responses across the six conditions. Corresponding to the 3 × 2 factorial design and binary response measure, we conducted multi-way frequency analyses using log-linear modeling.[4] This analysis revealed a significant main effect of scenario (strict: $G^2(2) = 17.82$, $p < .001$; liberal: $G^2(2) = 19.9$, $p < .001$). There was also a significant main effect of skeptical pressure, indicating that participants were less likely to recall "knew" when skeptical pressure was present (strict: $G^2(1) = 6.52$, $p = .011$; liberal: $G^2(1) = 7.71$, $p < .01$). There was no reliable interaction between these two variables (strict: $G^2(2) = .364$, $p = .834$; liberal: $G^2(2) = .340$, $p = .844$), although inspection of Fig. 1 suggests that the influence of skeptical pressure was minimal in the FB condition, presumably due to a floor effect. This pattern is consistent with the Bayesian interpretation of the impact of skeptical pressure on knowledge attribution.

As the interaction between scenario type and skeptical pressure was non-significant, we then pooled responses across the conditions of skeptical pressure (present versus absent) for each of the three scenarios. Fisher's exact tests revealed that participants were significantly more likely to give knew-type responses in the JTB condition than in the FB condition (strict: $p < .001$, Cramer's $V = .212$; liberal: $p < .001$, $V = .225$) or Gettier condition (strict: $p = .018$, $V = .131$; liberal: $p = .034$, $V = .118$). The significant difference in knew-type responding among the JTB and Gettier cases confirms the predictions of our Bayesian account. Knew-type responses tended to be more frequent in the Gettier condition than in the FB condition, a difference that was non-significant under the strict coding system, ($p = .136$, $V = .094$) but that was significance under the liberal coding system ($p = .043$, $V = .122$). Though not specifically predicted by our Bayesian account, this last result is also consistent with it (since the condition of truth is satisfied in the Gettier case, but not in the FB case).

[4] For the log-linear analysis, we label our results according to their analogs in the ANOVA framework, which we expect will offer more familiar terminology. Thus, first-order interactions in the log-linear model are referred to as "main effects", second-order interactions as "interactions", and so forth.

Finally, a 3 × 2 ANOVA was conducted on participant's recall scores for non-critical sentences to rule out the possibility that the observed differences in knew-type response rates might be attributable to some versions of the story being harder to remember than others. There were no significant main effects of condition ($F(2,525) = .580$, $p = .560$) nor of skeptical pressure ($F(1,525) = .382$, $p = .463$), nor was there any interaction between these variables ($F(2,525) = 1.887$, $p = .153$). Thus, differences between conditions in rates of intrusions were unique to the critical sentence.

4. Experiment 2

In Experiment 2 we examined a Gettier case resulting from replaced evidence. For this situation we predicted that participants would differentiate between a Gettier case resulting from replaced evidence and a false belief case, but would not differentiate between the Gettier case and a standard case of justified true belief.

4.1. Method

4.1.1. Participants

A total of 304 participants (164 female) were recruited in Experiment 2. The mean age of participants was 31.1 (SD = 10.34) years old.

4.1.2. Materials, design, and procedure

Experiment 1 revealed that the effects of skeptical pressure and the structure of Gettier cases operate orthogonally. As our investigation was primarily focused on Gettier cases, we did not manipulate skeptical pressure in subsequent experiments (i.e., none of vignettes introduced skeptical pressure). In Experiment 2 we used three stories, adapted from the detective stories described in the Introduction. In the first story, a character named Will is guilty of a crime and Dempsey, the detective in the story, finds evidence of his guilt, forming the justified true belief that he is guilty (JTB condition). Meanwhile, another character, Beth, who is Will's girlfriend, observes

the sequence of events that unfold and result in Dempsey thinking that Will is guilty. In the second story, Will is innocent of the crime, but is framed by his girlfriend Beth because she suspects that he is cheating on her. Dempsey finds evidence planted by Beth, and as a result forms the false belief that Will is guilty of the crime (FB condition). Finally, in the third story, Will is guilty of the crime, but he has eliminated all the authentic evidence of his crime. Beth, as part of a ploy to seek reprisals against Will, plants evidence that implicates him in the crime. Dempsey finds this evidence and forms the belief that Will is guilty. In this case, Dempsey's belief is both justified and true, but is only true in the context of unusual circumstances (Gettier condition). In each of these stories, the critical sentence was, "Whatever the ultimate verdict would be, Dempsey *thought* Will was guilty" (italics added here). Importantly, the evidence Dempsey uncovered was of the same quality in each version of the story. Again, the structure and wording of the stories was identical, save for the relevant manipulations.

4.2. Results and discussion

After coding participants' recall responses according to our strict coding system, 260 of the original 304 participants remained in the final analysis. Under the liberal coding system, 23 additional responses were scored, for a total of 283 out of 304 responses.

Fig. 2 shows the proportions of knew-responses across the different conditions. We observed more knew-type responses in the JTB (Fisher's exact test, strict: $p = .036$, Cramer's $V = .161$; liberal $p = .043$, $V = .157$) and Gettier conditions (strict: $p = .016$, $V = .188$; liberal: $p = .009$, $V = .198$) than in the FB condition. Rates of knew-type responses in the JTB and Gettier conditions were similar, and did not differ statistically (strict: $p = .760$, $V = .028$ liberal: $p = .655$, $V = .042$). Participants thus seemed to believe that agents in the Gettier case "knew" the accused was guilty, apparently drawing no distinction between the Gettier case of the replaced-evidence type and non-Gettier cases of justified true belief. These findings accord well with the predictions we derived from our Bayesian analysis of knowledge attribution.

We again followed our primary analyses with an ANOVA examining participant's recall scores for non-critical sentences. Unlike in Experiment 1, we found a significant difference in participants' non-critical recall performance in the three conditions, $F(2,301) = 3.864$, $p = .022$. Most importantly, non-critical recall performance was lower in the Gettier condition (mean = 1.02, SD = .750) than in the FB condition (mean = 1.27, SD = .763), $t(200) = 2.309$, $p = .022$.

Given the modest size of this effect ($d = .33$), it seems unlikely that our observation of differences in critical knew-type recall could be explained by a general tendency to make intrusions in the Gettier condition. However, we sought to rule out this possibility more clearly. We observed that a greater number of participants failed to recall any non-critical words in the Gettier condition than in either of the other two conditions. As such low performance might indicate a failure to attend to the case, we ran a second set of analyses after dropping data from these participants. We found no systematic differences in non-critical recall performance after these participants were removed, $F(2,251) = 1.96$, $p = .143$. However, we again found differences in rates of knew-type responses across conditions. Knew-type responses were observed more often in the JTB (strict: 40.7%; liberal: 41.2%) and Gettier (strict: 40.6%; liberal: 41.1%) conditions as compared with the FB (strict: 22.7%; liberal: 23.3%) condition (strict: $p = .017$, $V = .194$; liberal: $p = .014$, $V = .192$, and strict: $p = .030$, $V = .193$; liberal: $p = .017$, $V = .191$, respectively). Again, no significant differences were found between the frequencies of knew-type recall in the JTB and Gettier conditions (strict: $p > .99$, $V = .002$; liberal: $p > .99$, $V = .001$).

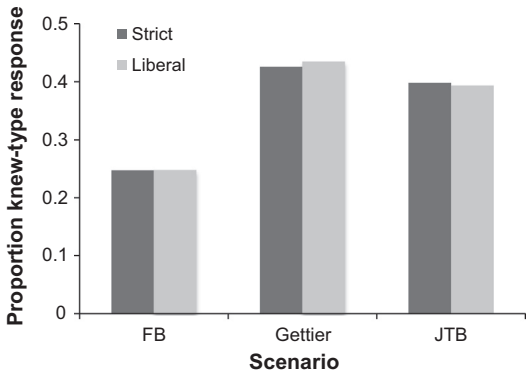

Fig. 2. Proportion of knew-type responses across conditions and under both strict and liberal coding systems (Experiment 2). FB = false belief; Gettier = replaced-evidence Gettier case; JTB = justified true belief.

5. Experiment 3

Experiments 1 and 2 revealed different patterns of results for counterfeit-object and replaced-evidence Gettier cases, confirming the predictions of our Bayesian account. However, these experiments also used very different cases that differed in many other respects (e.g., the types of evidence available to the agents and the costs of error). Experiment 3 was conducted to confirm that the different patterns of results observed for replaced-evidence and counterfeit-object cases are not specific to the particular stories used in Experiments 1 and 2. To this end, we examined two new sets of cases, thus increasing the generalizability of our findings. We examined counterfeit-object and replaced-evidence Gettier cases that occurred within the same general story, ensuring that differences in our findings could not be attributed to extraneous differences among the stories.

In addition, in order to rule out an alternative to our Bayesian account, we also examined Gettier cases due to the accidental replacement of evidence. The different patterns of results in our two previous experiments might be attributable to the presence of luck in the counterfeit-object case and its absence in the replaced-evidence cases. For the stories used in Experiment 3, we attempted to ensure that the agent's priors would be relatively high. Accordingly, we predicted that the observer's amended posterior will remain relatively high even in a case of

accidental replacement, and thus luck should not impugn the agent's knowledge as much as does a similar manipulation in a counterfeit-object case.

5.1. Method

5.1.1. Participants

A total of 1598 participants (952 female) were recruited in this experiment. The mean age of participants was 33.91 (SD = 12.05) years old. Experiment 3 examined a greater number of cases and conditions, necessitating a larger sample than those used in the previous experiments.

5.1.2. Materials, design, and procedure

Participants were assigned to one of six experimental conditions in Experiment 3: counterfeit-object false belief (FB-C), replaced-evidence false belief (FB-RE), non-Gettiered justified true belief (JTB), counterfeit-object Gettier (Gettier-C), replaced-evidence Gettier (Gettier-RE), and accidental-replaced-evidence Gettier (Gettier-ARE).

We created two sets of stories, each with six versions corresponding to the conditions described above. The different versions of both of these stories are summarized in Table 2. In the "conflict diamonds" stories (801 participants), a woman named Emma is shopping for a diamond and is taking steps to ensure that her diamond is conflict-free. Each diamond has a serial number that can be used to identify its origins. Emma contacts a friend at Amnesty International who can look up the diamond in a database and send her a certificate of the diamond's conflict-free status. The critical sentence in these stories read: "Emma didn't feel one bit guilty about purchasing the diamond, since she *thought* it was conflict-free" (italics added here).

In the "vegan cookies" stories (797 participants), a woman named Sharon has a craving for chocolate chip cookies. Her husband has recently gone grocery shopping and restocked their cookie jar, but Sharon is vegan and worries that her husband may have purchased the wrong cookies. Sharon looks for evidence that the cookies are vegan, and finds some vegan cookie packaging in the trash that matches the cookies in the jar. In these stories, the critical sentence read, "Although she was worried earlier, now she *thought* the cookie was vegan" (italics added here). By using two entirely different cover stories for the vignettes, we sought to further establish that the pattern of results is not specific to any one particular scenario.

Finally, Experiment 3 introduced "catch" questions during the distractor task, meant to identify participants who were not paying attention to the task. For these questions, participants were simply instructed to enter a particular response (e.g., "true").

Table 2
Summary of conditions for the two stories used in Experiment 3.

	Conflict diamonds stories
JTB	Emma is considering buying a conflict-free diamond and her friend sends her the correct certificate showing that the diamond is conflict-free.
Gettier-C	A dishonest jeweler has falsely labeled many of his diamonds with serial numbers from legitimate conflict-free diamonds. Emma happens to pick one of the only truly conflict-free diamonds in the store.
FB-C	Emma is unlucky and chooses a conflict diamond.
Gettier-ARE	Emma's friend makes a typo while looking up the serial number and just happens to find a conflict-free certificate. Fortunately, Emma's diamond really is conflict-free.
Gettier-RE	Emma's friend looks up the correct diamond and sees that it is conflict-free but his computer crashes before he can send the certificate. To save time, he later looks up a different diamond and sends Emma that certificate instead.
FB-RE	Emma's friend makes a typo while looking up the serial number and finds a conflict-free certificate, but Emma's diamond is not actually conflict-free.
	Vegan cookies stories
JTB	Mark purchased vegan cookies and threw away the wrapper in the trash, which Sharon then found.
Gettier-C	Mark made the correct purchase but a mistake at the factory led non-vegan cookies to be placed into vegan cookie packages. Sharon happened to pick one of the only truly vegan cookies in the jar.
FB-C	Mark made the correct purchase but a mistake at the factory led non-vegan cookies to be packaged into vegan cookie packages. Sharon chose a non-vegan cookie.
Gettier-ARE	Before shopping, Mark dug through the trash to find an old wrapper so he could be sure to purchase the right brand of cookies. He happened to leave the old packaging on top of the trash, which Sharon found. Fortunately, he not only purchased the right brand, but also vegan cookies.
Gettier-RE	Mark purchased vegan cookies and recycled the original wrapper. Then, he wondered if Sharon would trust that the cookies were vegan. To avoid having an argument about the groceries, he dug up the old packaging from the trash bin and planted it on top for her to find.
FB-RE	Before shopping, Mark dug through the trash to find an old wrapper so he could be sure to purchase the right brand of cookies, but still made a mistake while shopping. Mark purchased the correct brand, but accidentally purchased their non-vegan cookies.

5.2. Results and discussion

Before coding participants' responses, participants were screened based on their responses to catch questions, which resulted in removal of data for 21 participants.[5] Using our strict coding system, 1201 of the 1557 eligible responses were scored. Using our more liberal coding system, 1414 responses were scored.

Data were pooled across the two story conditions, as a log-linear analysis found no reliable differences in the pattern of results among the conflict diamonds stories compared with the vegan cookies stories ($G^2(5) = 7.147$, $p = .210$). Fig. 3 shows the pattern of results under each coding system. These results were largely in agreement, and analyses of these the two coding systems are reported together. First, we tested for differences between the JTB and FB conditions. As expected, rates of knew-type responding were higher in the JTB condition than in the FB-C (Fisher's exact test, strict: $p < .001$, Cramer's $V = .262$; liberal: $p < .001$, $V = .245$) and FB-RE conditions (strict: $p < .001$, $V = .232$; liberal: $p < .001$, $V = .237$).

[5] A separate set of analyses indicated that excluding these participants did not materially change the results of this experiment.

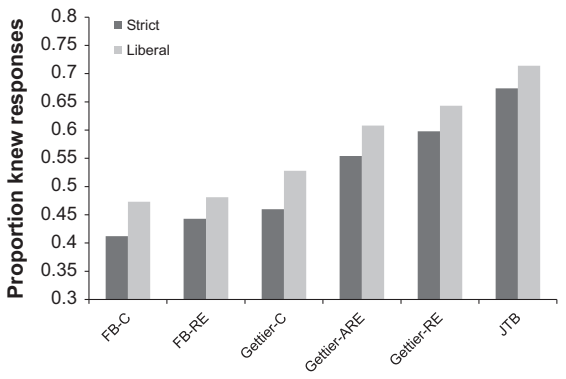

Fig. 3. Proportion of knew-type responses across conditions and under both strict and liberal coding systems (Experiment 3). C = counterfeit object; RE = replaced evidence; ARE = accidental replaced evidence; FB = false belief; JTB = justified true belief.

Next, we tested the predictions derived from our Bayesian analysis of knowledge attribution. We predicted that the circumstances of the counterfeit-object Gettier case would have the greatest effect on activation of participants' knowledge concepts. As predicted, the Gettier-C condition produced fewer knew-type responses than the JTB condition (strict: $p < .001$, $V = .216$; liberal: $p < .001$, $V = .192$), and did not differ from its corresponding FB-C condition (strict: $p = .369$, $V = .048$; liberal: $p = .271$, $V = .055$).

Next, we tested our predictions for the replaced-evidence cases. Compared with the FB-RE condition, participants were also more likely to give knew-type responses in the Gettier-RE condition (strict: $p = .002$, $V = .155$; liberal: $p < .001$, $V = .163$), and in the Gettier-ARE condition (strict: $p = .035$, $V = .110$; liberal: $p = .006$, $V = .127$). Knew-type response rates did not differ between the Gettier-RE and JTB conditions (strict: $p = .143$, $V = .078$; liberal: $p = .111$, $V = .076$), but were significantly lower in the Gettier-ARE condition than in the JTB condition (strict: $p = .016$, $V = .123$; liberal: $p = .018$, $V = .113$). Still, the Gettier-RE and Gettier-ARE conditions did not differ significantly (strict: $p = .418$, $V = .045$; liberal: $p = .449$, $V = .036$). We also tested whether rates of knew-type responding were lower in the Gettier-C condition than in the Gettier-RE and Gettier-ARE conditions. The Gettier-RE condition produced significantly more knew-type responses than the Gettier-C condition (strict: $p = .007$, $V = .139$; liberal: $p = .012$, $V = .117$). Knew-type responses were also more frequent in the Gettier-ARE condition, although this difference only approached significance (strict: $p = .070$, $V = .094$; liberal: $p = .094$, $V = .081$). As predicted, the intentional replacement of evidence did not reduce rates of knew-type responses below those observed in the JTB condition. Evidence replaced by luck did reduce knew-type responding, but not so severely as in the counterfeit-object Gettier cases. Altogether, the observed pattern of results accords with the predictions of our Bayesian account of knowledge attribution.

Finally, ANOVAs were conducted to examine memory intrusions in the non-critical sentences in the recall task. No differences were observed between conditions among the conflict diamonds stories ($F(5,786) = .778$, $p = .566$), nor among the vegan cookies stories ($F(5,779) = .230$, $p = .949$), indicating that differences in recall performance for each story were specific to the critical sentence.

6. General discussion

We proposed a Bayesian framework for knowledge attribution, and applied this framework to generate novel predictions about knowledge attribution for different types of Gettier cases. Then, we tested these predictions using a paradigm based on semantic integration, coding the frequencies with which participants falsely recalled the word "thought as "knew" (or a near synonym) to yield an implicit measure of conceptual activation. Using this semantic-integration method, we confirmed the predictions of our Bayesian account of knowledge attribution across three experiments. Based on participants' recall performance, we conclude that Gettier cases due to counterfeit objects were less likely to be treated as cases of knowledge (Experiment 1). However, Gettier cases due to intentionally-replaced evidence do appear to have been treated as cases of knowledge (Experiment 2), as they elicited false recall of "knew" (and near synonyms) at rates similar to standard cases of justified true belief. Our findings are not well explained by an alternative account based only on luck, because Gettier cases due to accidentally-replaced evidence activated the knowledge concept more strongly than did similar false belief cases (Experiment 3). We observed a consistent pattern of results across a number of vignettes that varied the quality and type of evidence available to agents, the relative stakes involved, and surface details of each story. Importantly, findings based on the implicit semantic-integration paradigm are not subject to the criticisms that have been directed at explicit survey-based measures of knowledge concepts (e.g., Powell, Horne, et al., 2014; Starmans & Friedman, 2013). Accordingly, the present findings establish basic phenomena surrounding people's knowledge attributions in Gettier cases, and explain these phenomena within a Bayesian framework.

Our findings suggest the surprising conclusion that attributions of knowledge do not require that the agent have an accurate causal model of the situation. A wide range of evidence shows that people's reasoning is often guided by causal models (for a review see Holyoak & Cheng, 2011). But in the replaced-evidence Gettier case we have discussed, the detective Dempsey is mistaken about the causal process that produced the evidence convincing him of Will's guilt—rather than being the direct effect of the crime that Will committed, the evidence is the effect of Beth's intervention. The agent's erroneous causal model may underlie the intuition held by many philosophers that the agent does not "know" in this situation. However, our Bayesian model, as well as our lay participants, attribute knowledge to the detective in this case. What appears critical is not the accuracy of the agent's causal model *per se*, but simply the consistency of the posterior probabilities computed by the agent and those amended by the observer.

Crucially, the model generates these predictions by assuming that observers consider only the most immediately relevant hypotheses, without reasoning too far beyond the specifics of the case. It should be emphasized that this analysis does not in any way entail that observers *ought* to reason in this way. Continuing with the same replaced-evidence Gettier case as our example, an observer could reason beyond the facts of the case, and consider whether or not Beth might have framed Will if he had not committed the crime (affecting the probability of the data under h_0). It might be argued that, normatively, an observer ought to consider this hypothesis and adjust their amended posterior accordingly. In this case, they would likely refrain from attributing knowledge to agents in replaced evidence cases. However, our findings suggest that observers consider these cases in the simplest and most direct way possible, focusing only on the events actually described in the case.

Finally, we should consider what the present computational account of knowledge attribution can actually tell us about people's concept of knowledge. Our Bayesian account is not meant to describe the actual semantic content of this concept. Rather, we argue that the semantic content of the knowledge concept—whatever that may be—leads people to exhibit behavior that is at least qualitatively consistent with this Bayesian account. There are potentially many different representations, algorithms, or semantic contents that could produce such a pattern of responding, so our account leaves many questions open. Still, an accurate account at the computational level would constitute a major step toward understanding and explaining people's knowledge attribution behavior, and their concept of knowledge.

6.1. Issues for further investigation

There are a number of important issues that our account does not yet fully address, but that deserve investigation. First, knowledge attributions depend on accurate estimates of the likelihood of evidence and of the prior probabilities of hypotheses, both for the agent and the observer. Making these estimates is a task that appears to require a fairly sophisticated "theory of mind", which could pose a variety of challenges. For example, some agents might be experts who are particularly capable in their evaluations of evidence; others, such as children, might be less capable (e.g., O'Neill, Astington, & Flavell, 1992).

Second, knowledge attribution requires that observers set appropriate thresholds for what constitutes knowledge, or how large a posterior probability is necessary for a hypothesis to be known. As contextualists in epistemology have argued (Cohen, 1986; DeRose, 1991; Lewis, 1996), different contexts may call for different thresholds, which will in turn influence an observer's propensity to attribute knowledge to an agent. It is perhaps vital that epistemic thresholds be contextually dependent in order for a probability-threshold account of knowledge such as ours to avoid the "lottery paradox" (Kyburg, 1961; also see Pinillos, 2011). Although the chances that a given ticket-holder will win the lottery are often vanishingly small, it cannot be rightly said that any ticket-holder knows they will lose the lottery. This thought experiment may suggest that the only acceptable threshold for knowledge is absolute certainty, thereby leading to skepticism about knowledge. Contextualism offers a solution to this paradox. On such an account, the threshold for knowledge differs according to one's present epistemic context: when extreme probability values are involved, as in lottery cases, the threshold for knowledge is higher than in most ordinary cases (Hawthorne, 2003). This assumption would enable threshold accounts to at once explain why the ticket-holder doesn't know they will lose, and to allow for successful knowledge attributions under more ordinary circumstances.

In addition to extreme probability values, many other contextual factors may influence the appropriate threshold for knowledge. For instance, the issue of how thresholds for knowledge are established is especially salient in legal contexts where standards range from "by the preponderance of evidence" to "beyond a reasonable doubt." These issues raise important questions for our theory and for future research, which could examine how context affects epistemic thresholds, and what happens when observers' and agents' thresholds differ.

Lastly, there are discrepancies between our findings and some prior findings reported in the literature. For example, Turri et al. (2014) found that participants attributed knowledge in counterfeit-object cases, but refrained from attributing knowledge in a case involving replaced evidence—the opposite of the findings reported here. One possibility is that these discrepant findings reflect differences in implicit and explicit knowledge attributions. That is, explicit questions may prompt participants to reflect on agents' and their own causal models of a situation, leading them to different judgments. Further research is needed to explore this possibility. Of course, alternative explanations are also possible. For example, these explicit questions may also have introduced demand characteristics, or otherwise biased participants' responses.

6.2. Conclusions

The present investigation introduced a new theoretical framework for understanding knowledge attribution, and a novel experimental paradigm (semantic integration) that provides an implicit measure of laypeople's knowledge attribution. From our theoretical account we generated novel predictions about people's knowledge attributions in different types of Gettier cases, and confirmed these predictions experimentally using a semantic integration paradigm. The Bayesian framework we have proposed provides an explanation for the knowledge attribution phenomena we uncovered, while also providing a means to generate further predictions about knowledge attribution and people's knowledge concept. In addition, the present experiments demonstrate the efficacy of the semantic integration paradigm as an implicit measure of conceptual activation. Together, these theoretical and methodological innovations establish a foundation for future investigation of people's knowledge concept.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.03.002.

References

Ayer, A. (1956). *The problem of knowledge*. London: Macmillan.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*(5), 382–386. http://dx.doi.org/10.1111/j.1467-9280.2007.01909.x.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*, 331–350.

Cohen, S. (1986). Knowledge and context. *The Journal of Philosophy, 83*, 574–583.

Cohen, S. (1998). Contextualist solutions to epistemological problems: Skepticism, Gettier and the lottery. *Australasian Journal of Philosophy, 76*(2), 289–306.

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme, 11*(02), 199–212. http://dx.doi.org/10.1017/epi.2014.7.

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology, 1*, 275–296.

DeRose, K. (1991). Epistemic possibilities. *Philosophical Review, 100*(4), 581–605.

Fantl, J., & McGrath, M. (2009). Advice for fallibilists: Put knowledge to work. *Philosophical Studies, 142*(1), 55–66.

Fischhoff, B. (1975). Hindsight does not equal foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299.

Flagg, P. (1976). Semantic integration in sentence memory? *Journal of Verbal Learning and Verbal Behavior, 15*, 491–504.

Franks, J. J., & Bransford, J. D. (1974). A brief note on linguistic integration. *Journal of Verbal Learning and Verbal Behavior, 13*(2), 217–219.

Gentner, D. (1981). Integrating verb meanings into context. *Discourse Processes, 4*, 349–375.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*, 121–123.

Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy, 73*(20), 771–791.

Greco, J. (2003). Knowledge as credit for true belief. In M. DePaul & L. Zagzebski (Eds.), *Intellectual virtue: Perspectives from ethics and epistemology*. New York: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9), 767–773.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*, 704–732.

Hahn, U., & Oaksford, M. (2012). Rational argument. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 277–298). New York: Oxford University Press.

Hawthorne, J. (2003). *Knowledge and lotteries*. New York: Oxford University Press.

Hawthorne, J., & Stanley, J. (2008). Knowledge and action. *Journal of Philosophy, 105*, 571.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135–163.

Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Middletown, CT: Wesleyan University Press.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy, 74*(4), 549–567.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*(4), 955–984.

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

Nagel, J., Mar, R., & San Juan, V. (2013). Authentic Gettier cases: A reply to Starmans and Friedman. *Cognition, 129*(3), 8–11.

Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition, 129*(3), 652–661.

O'Neill, D., Astington, J., & Flavell, J. (1992). Young children's understanding of the role that sensory experiences play in knowledge acquisition. *Child Development, 63*(2), 474–490.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776–783.

Owens, J., Bower, G., & Black, J. (1979). The "soap opera" effect in story recall. *Memory & Cognition, 7*(3), 185–191.

Pinillos, N. Á. (2011). Some recent work in experimental epistemology. *Philosophy Compass, 6*(10), 675–688.

Plato (1961). Theaetetus (F. M. Cornford, Trans.). In E. Hamilton & H. Cairns (Eds.), *The collected dialogues of Plato* (pp. 200d–201d). Princeton, NJ: Princeton University Press.

Powell, D., Horne, Z., & Pinillos, Á. (2014). Semantic integration as a method for investigating concepts. In J. Beebe (Ed.), *Advances in experimental epistemology*. London: Bloomsbury Academic.

Powell, D., Horne, Z., Pinillos, Á., & Holyoak, K. J. (2013). Justified true belief triggers false recall of "knowing". In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1151–1156). Austin, TX: Cognitive Science Society.

Powell, D., Merrick, A., Lu, H., & Holyoak, K. J. (2014). Generic priors yield competition between independently-occurring preventive causes. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 2793–2798). Austin, TX: Cognitive Science Society.

Powell, D., Merrick, A., Lu, H., & Holyoak, K. J. (2013). Generic priors yield competition between independently-occurring causes. In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1157–1162). Austin, TX: Cognitive Science Society.

Pritchard, D. (2004). Epistemic luck. *Journal of Philosophical Research, 29*, 193–222.

Rai, T. S., & Holyoak, K. J. (2014). Rational hypocrisy: A Bayesian analysis based on informal argumentation and slippery slopes. *Cognitive Science*. http://dx.doi.org/10.1111/cogs.12120.

Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2*, 437–442.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology, 26*, 123–162.

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. I). New York: Oxford University Press.

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition, 124*, 272–283.

Starmans, C., & Friedman, O. (2013). Taking "know" for an answer: A reply to Nagel, San Juan, and Mar. *Cognition, 129*(3), 662–665.

Sturgeon, S. (1993). The Gettier problem. *Analysis, 53*, 156–164.

Sulin, R. A., & Dooling, D. J. (1974). Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology, 103*, 255–262.

Thorndyke, P. (1976). The role of inferences in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior, 15*, 437–446.

Turri, J. (2011). Manifest failure: The Gettier problem solved. *Philosophers Imprint, 11*(8), 1–11.

Turri, J. (2013). A conspicuous art: Putting Gettier to the test. *Philosopher's Imprint, 13*(10), 1–16.

Turri, J., Buckwalter, W., & Blouw, P. (2014). Knowledge and luck. *Psychonomic Bulletin & Review*.

Waskan, J., Harmon, I., Horne, Z., Spino, J., & Clevenger, J. (2014). Explanatory anti-psychologism overturned by lay and scientific case classification. *Synthese, 191*(5), 1013–1035.

Weber, S., & Cook, T. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*, 273–295.

Williams, M. (1978). Inference, justification, and the analysis of knowledge. *Journal of Philosophy, 75*, 249–263.

Williamson, T. (2002). *Knowledge and its limits*. New York: Oxford University Press.

Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition, 115*(3), 491–503.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.

Zagzebski, L. (1996). *Virtues of the mind*. Cambridge, UK: Cambridge University Press.