METHODOLOGICAL ISSUES IN EPISTEMOLOGY AND MORAL PSYCHOLOGY

BY

ZACHARY HORNE

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Philosophy in the Graduate College of the University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Jonathan Waskan, Chair
Associate Professor Daniel Korman
Assistant Professor Jonathan Livengood
Professor John Hummel
ABSTRACT

Between 1960 and 1999, it was quite common for philosophers to rely almost completely on *a priori* methods to advance their arguments (Knobe, 2015); in a recent study by Knobe, the majority of papers sampled from this period used strictly *a priori* methods. In contrast, in the last decade and a half, many philosophers’ strategy for making progress on philosophical questions has changed. Philosophers are now relying more heavily on empirical data—including running their own observational and experimental studies—in order to support their arguments. Without a doubt, part of this shift was due to the rise of experimental philosophy—roughly a methodological approach to addressing philosophical questions wherein the researcher uses psychological methods to investigate the parameters that affect the deployment of philosophically significant concepts. This dissertation raises methodological problems with work in experimental philosophy and then proposes solutions to these problems. My focus in this dissertation is on two subfields in philosophy—epistemology and moral psychology—both of which have witnessed increased uses of psychological methods to investigate philosophical questions.
# TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION ...........................................................................................................1

CHAPTER 2: SEMANTIC INTEGRATION AND EXPERIMENTAL PHILOSOPHY.............4

CHAPTER 3: EVIDENCE FOR ANTI-INTELLECTUALISM ABOUT KNOW-HOW FROM A SEMANTIC INTEGRATION TASK ..................................................................................27

CHAPTER 4: SYNTACTIC CUES INDUCE DIVERGENT EPISTEMIC INFERENCES ........................................................................................................................................41

CHAPTER 5: HOW LARGE IS THE ROLE OF EMOTION IN JUDGMENTS OF MORAL DILEMMAS? ..............................................................................................................................66

CHAPTER 6: A SINGLE COUNTEREXAMPLE LEADS TO MORAL BELIEF REVISION ........................................................................................................................................93

CHAPTER 7: ORDERING EFFECTS AND UPDATING EFFECTS .....................................113

REFERENCES ...................................................................................................................................155
CHAPTER 1
INTRODUCTION

*Philosophers are doing something different now* – Knobe

Between 1960 and 1999, it was quite common for philosophers to rely almost completely on *a priori* methods to advance their arguments (Knobe, 2015); in a recent study by Knobe, the majority of papers sampled from this period used strictly *a priori* methods. In contrast, in the last decade and a half, many philosophers’ strategy for making progress on philosophical questions has changed. Philosophers are now relying more heavily on empirical data—including running their own observational and experimental studies—in order to support their arguments. Without a doubt, part of this shift was due to the rise of experimental philosophy – roughly a methodological approach to addressing philosophical questions wherein the researcher uses psychological methods to investigate the parameters that affect the deployment of philosophically significant concepts. This dissertation raises methodological problems with work in experimental philosophy and then proposes solutions to these problems. My focus in this dissertation is on two subfields in philosophy – epistemology and moral psychology – both of which have witnessed increased uses of psychological methods to investigate philosophical questions.

1. Epistemology

Some studies in experimental philosophy have taken the shape of observational survey studies examining people’s knowledge ascriptions without manipulating people’s knowledge ascribing behavior (e.g., Bengson, Moffett, & Wright, 2007). In contrast, some experimental studies has sought to understand the psychological factors at play in epistemological debates by manipulating the parameters (e.g., justification, truth) thought to affect knowledge attribution
(e.g., Powell, Horne, Pinillos, & Holyoak, in press). Although many epistemologists have welcomed the use psychological methods, others have taken issue with this strategy (e.g., Kauppinen, 2007; Sosa, 2007). Epistemologists in the latter camp suggest that, for example, current methods do not help us make significant progress in understanding the parameters that affect knowledge attribution because responses to surveys may not be truly reflective of people’s knowledge concept.

In Chapters 2 and 3, I argue that critics of experimental epistemology are right to be skeptical of the results from survey studies, but that there are other empirical methods that do not succumb to the criticisms that plague surveys. As one example, I show that memory tasks can be used to implicitly measure conceptual activation of knowledge (Chapter 2) and know-how (Chapter 3). Consequently, I argue we should not conclude that empirical data is irrelevant to traditional epistemological questions full stop just because some psychological methods are ill-suited for the task. However, I also demonstrate that when sufficient care is taken to rule out competing hypotheses regarding effects found through the use of survey tasks, these data can also have important implications for traditional epistemological debates. For example, in Chapter 3, I discuss evidence showing how domain-general psychological biases have significant implications for ongoing debates in traditional epistemology based on data drawn from directly asking people questions about knowledge.

2. Moral Psychology

Paralleling the increasing use of experimental methods in epistemology, empirical methods and results are now playing a larger role in mainstream ethics and moral psychology (Greene, 2013). Several ethicists have argued that not only does psychological research offer insight into normative ethics and meta-ethics by elucidating how laypeople reason about ethical
issues (e.g., Prinz, 2007; Sinnott-Armstrong, 2008), but also normative ethical truths can be read off, as it were, from empirical discoveries (e.g., Greene, 2007; Prinz, 2007; Singer, 2005). Here too, skeptics have argued that empirical data is irrelevant to normative ethical considerations (e.g., Moore, 1993; 1903), while others suggest that popular arguments in which empirical data is meant to demonstrate the unreliability of certain moral intuitions begs the question against the views that these data are aiming to impugn (e.g., Berker, 2009).

The last three chapters in my dissertation concern ongoing debates in moral psychology. Although I do not argue that ethical truths can be derived from psychological findings, I show how understanding the psychological processes involved in moral judgment can help adjudicate between competing moral psychological theories (e.g., Greene et al., 2001; Prinz, 2007), and allow us to understand the role of thought experiments in ethical theorizing. In chapter 5, I argue that while moral judgments are emotion-laden, emotion does not primarily drive moral judgments about moral dilemmas. In chapter 6, I argue that people treat their judgments about moral dilemmas as evidence for or against some of their moral beliefs. In turn, my research has significant implications for philosophical issues like moral belief revision. In the final chapter, I show how understanding the psychological issues at play in moral judgment affects how we understand the reliability of moral judgments and moral intuitions. In sum, my research on moral psychology speaks to the claim that empirical results offer important insights into central philosophical questions.
CHAPTER 2  
SEMANTIC INTEGRATION AND EXPERIMENTAL PHILOSOPHY

The last decade in philosophy has been marked by widespread interest in experimental philosophy (see Knobe, 2015 for a review). Much of this work has been carried out by philosophers, who aim to better understand the contours of philosophical concepts and intuitions by importing methods from the social sciences. Many experimental philosophers hope is that a better understanding of the psychology of philosophically important concepts such as KNOWS, INTENTION, FREE WILL, and many others, will allow them to better assess philosophical arguments in which these concepts are deployed. Experimental philosophers have amassed a number of interesting results, but compelling concerns have been raised about the survey-based experimental methods that they typically use. Here we argue, on the basis of these concerns and our own, that the possibility of experimental artifacts is good reason to adopt new, more reliable experimental paradigms. In this chapter, we argue that one new experimental paradigm that we call semantic integration is one promising method for empirically examining people’s concepts. There are two desirable advantages to this method. First, because the method uses a memory task as an implicit measure of the degree to which different situations instantiate concepts, it avoids the methodological problems with surveys. Second, and perhaps more importantly, the method more directly taps into people’s concepts, allowing experimental philosophers to address several concerns philosophers have raised about the philosophical import of empirical data. Both of these points will be explained in further detail in sections 3 and 4 of the chapter.

The chapter will proceed as follows: After discussing survey methodology (section 1), we consider some challenges to the methodology (Section 2) and then describe how semantic integration tasks can be used to implicitly examine people’s concepts (Section 3). Next we argue that, by investigating concepts implicitly, semantic integration offers important advantages over
more explicit survey methods. Finally, we discuss caveats regarding semantic integration methods, and variations on these methods (Section 4).

**Section 1: The methods of experimental philosophy**

Experimental philosophers investigate philosophical concepts by presenting participants with short passages and then asking them to make judgments about a series of questions on the basis of this passage. These passages, which are often derived from philosophical thought experiments, are designed to test whether certain features of a situation are parameters for instantiating a philosophical concept. Although studies using this survey methodology have improved substantially when compared to early research that lacked proper control conditions, the methodology is still limited in important ways. In this section, we review some of the challenges faced by researchers that rely on the standard method to investigate people’s concepts. We will focus on two issues in with survey methodology – *pragmatic cues* and *demand characteristics*. First, we will discuss issues raised by Simon Cullen (2010) about how pragmatic features can influence survey responses. Second, we will address longstanding methodological limitations of survey methodology.

In a recent critique, Cullen (2010) argues that researchers conducting surveys need to take into account both the semantic and the pragmatic features of their experimental materials. Grice (1975) observed that when people attempt to comprehend a natural language utterance, they do not attempt to comprehend the exact meaning of the words as spoken or as written on the page. Rather, they attempt to comprehend the speaker’s meaning. Cullen argues that the participants in an experiment behave similarly, attempting to comprehend the experimenter’s meaning, a consideration that many experimental philosophers have ignored. According to Grice (1975), people make assumptions about the requirements for rational communication. These
assumptions, often referred to as Gricean norms, allow listeners or readers to grasp what a speaker means to convey or what they think a speaker means to convey. People assume that speakers are “cooperative communicators” — that their utterances are true, orderly, relevant, and nonredundant. Typically, speakers are themselves aware that their interlocutors make these assumptions and so they exploit these assumptions to help conversational participants understand them. For example, sometimes the best way of making sense of someone’s communications, given that they are following Gricean norms, is by inferring that they mean something that goes beyond what they said or stated with their utterance.

If John asks, “Has the number two bus come by yet?” a listener, Bill, can infer that there is a number two bus, that its route passes by this location, and that John is hoping to catch the bus. Of course, none of these facts are explicitly stated in John’s question. Bill can infer these things because he assumes that John is following Gricean norms. For example, John would not be following the norm of relevance if he was not planning to get in the bus. Moreover, John can count on Bill to infer these things about his utterance because John knows that Bill will assume he is following the Gricean norms. Roughly, the propositions that the speaker means to convey go beyond what is said and are inferable in a conversation applying the Gricean norms are called “conversational implicatures” (as opposed to propositions that are conventionally associated with the words used). It is widely accepted that the deployment and computing of conversational implicatures is pervasive in human communication. For this reason, Cullen (2010) argues that if experimental philosophers ignore conversational implicatures, then their instructions, stimuli, response options, and other experimental materials may not convey the meanings they intend.

For an illustration of how conversational implicatures can affect survey results, consider research on base-rate neglect, which is the tendency for people to ignore relevant statistical base
rates when judging the probabilities of events (for a review, see Nisbett & Ross, 1980). In one study, Kahneman and Tversky (1973) presented people with a description of a fictional college student and asked them to estimate the probability that the student majored in various fields. If the descriptions included traits that seem stereotypical of an engineering student (e.g., introverted, enjoys solving problems), then people estimated the probability that the student was an engineer was high. Interestingly, people make similar probability estimates even when they were told that only a small percentage of students study engineering. Kahneman and Tversky concluded that people ignore base-rate information in their probability estimates and instead employ a representativeness heuristic: since the student resembles an engineer, they judge that it is probable he is one, and they ignore the base-rate information which would suggest that any individual student is most likely not an engineer.

However, more recent research suggests that base-rate neglect may be due, at least in part, to conversational processes rather than to decision processes. If participants assume that experimenters are cooperative communicators, then they assume that the information they’ve been given is the most relevant to the task at hand. This may lead them to place a greater weight on the descriptions given than they would have otherwise. Schwarz and colleagues (1991) examined this by manipulating the guarantee of relevance. Participants in one condition were told that the descriptive information presented to them had been compiled by psychologists (as in the original experiments of Kahneman and Tversky), and in another condition, they were told that the same description had been compiled by a computer that randomly sampled from a database of information. Whereas communication from another person comes with an implied guarantee of relevance, people may not think that computer-generated text has the same guarantees. As predicted, researchers found that participants were significantly less influenced
by computer-generated descriptions than by human-generated descriptions (Schwarz et al., 1991). Even relatively subtle pragmatic cues can have important effects on people’s responses to survey questions. For instance, people seem to place greater weight on the last source of evidence they are shown: Krosnick and colleagues (1990) found that base rates had a larger effect on participants’ judgments when they were the last piece of information participants read before making their response. The guarantees of relevance and nonredundancy imply that if experimenters present an apparently sufficient source of evidence (e.g., base-rate information), and then present another source of evidence (e.g., the description of a person), then this second source should be interpreted as nonredundant and highly relevant to the task at hand.

Cullen (2010) demonstrated that pragmatic cues can also affect people’s responses to philosophical thought experiments. However, he argues that researchers can overcome these challenges if they are sensitive to the context in which participants interpret their experimental materials, and the norms that govern these interpretations. Following Schwarz (1994), he argues that experimenters and participants are engaged in a conversation governed by the norms of cooperative communication (Cullen, 2010; Grice 1975; Schwarz, 1994). Since participants abide by these norms, and expect researchers to abide by them as well, experimental materials must be constructed with pragmatic cues in mind.

We agree that addressing the pragmatic features of experimental materials would improve the conclusions that can be drawn from surveys. However, overcoming these challenges might prove quite difficult. In practice, researchers still need to determine exactly how materials and questions ought to be phrased, and what implicatures they ought to contain. To make matters more difficult, this would need to be determined for each concept that experimental philosophers intend to examine.
To illustrate the difficulty of designing appropriate questions and materials, consider the challenges faced by researchers studying causal learning: An important construct in research on causal learning is *causal strength*, defined as the probability that some cause produces an effect (Cheng, 1997). Although people often make judgments about causal strength, researchers can ask participants to report such a judgment in any number of ways, and it is not obvious which way is optimal. In one experiment Buehner et al. (2003) asked their participants to make a causal strength rating on a scale from 0 (X does not cause Y at all) to 100 (X causes Y every time). They found that participants’ judgments tended to cluster into two groups: one group of participants made judgments consistent with Cheng’s (1997) probabilistic definition of causal strength, and the other participants made judgments consistent with competing associative models. Because causal learning is assumed to be a fundamental cognitive mechanism (Cheng, 1997), it would be quite remarkable if people learned causal relationships through different cognitive mechanisms. However, Buehner and colleagues investigated whether ambiguities in the question they used to probe participants’ judgments were responsible for the divergent pattern of responses. Indeed, they noted that the causal strength question they initially used can be interpreted as applying in one of two different contexts: (1) the experimental learning context where the effect is also produced by other background causes or (2) a counterfactual context where only the cause of interest is present. The clustering of participants’ diverging responses appeared to follow these two interpretations. The upshot of this research is that resolving ambiguities and constraining participants’ interpretations of questions and materials is feasible, but can require systematic investigation for each concept at issue.
Section 2: Demand characteristics

The results of surveys can also be affected by demand characteristics (Orne, 1962). Crudely put, demand characteristics are features of an experimental task that lead participants to perform a task other than what the task researchers intended them to. Demand characteristics can occur when participants are apprehensive about being evaluated (Weber & Cook, 1972). Apprehension can lead participants to respond in ways they perceive as either socially desirable, or “correct,” irrespective of their actual attitudes or intuitions. Demand characteristics can also occur when participants assume the role of a faithful participant, eschewing all pragmatic cues and following instructions exactly to the letter (Weber & Cook, 1972). Survey materials in experimental philosophy studies are particularly likely to exhibit demand characteristics because experimental philosophers often present naive participants with bizarre thought experiments. Although the uniqueness of thought experiments may be harmless in professional philosophy, there is evidence that survey participants are more likely to assume a faithful role, ignoring pragmatic and contextual cues, when experimental materials are particularly unrealistic (Weber & Cook, 1972). In other words, if experimental materials are convoluted or strange, then participants are more likely to ignore the contextual cues in experimental materials, or to interpret them under different assumptions. Additionally, if participants are apprehensive about being evaluated, then they are more likely to try to guess at desirable or “correct” response. For instance, participants may engage in a kind of amateur philosophizing, diverging from the aims of experimental philosophers. If demand characteristics cannot be ruled out, then it is unclear how to interpret the results of surveys.¹

¹ Although we have focused on demand characteristics in experimental philosophy, it is also common for experimental demands to affect experiments in psychology. For instance, in a recent paper, Firestone and Scholl (2013) argue that the effects reported in a large literature in social psychology, more than likely, owes to demand
Section 2.1: Semantic Integration

In this section, we propose a new methodology for investigating concepts that we call semantic integration. First, we introduce research on memory and language processing that inspired the experimental methodology we propose. Then, we describe the components of a semantic integration task, and two toy experiments in which we use this method.

Semantic integration uses a memory task as an implicit measure of how concepts are activated by reading different passages. As we discuss, this method has important advantages over survey-based research: it minimizes the influence of pragmatic cues and greatly reduces the possibility of demand characteristics. What’s more, we argue that semantic integration provides a more direct measure of conceptual activation. In contrast, participants’ responses to survey questions in experiments typically constitute their judgments about whether a particular concept applies in a given situation. These judgments may be the products of either people’s concepts or other downstream decision processes.

Section 2.1.1: Memory and Language Processing Research

People unfamiliar with memory research tend to think of errors in memory as errors of omission — they acknowledge that we sometimes forget things that have happened to us, but

characteristics (Firestone & Scholl, 2013). As an example, consider one study purporting to demonstrate that people who think about unethical (rather than ethical) actions will literally see the world as though it is darker (Banerjee, Chatterjee, & Sinha, 2012). The experiment proceeded as follows: Participants first thought about an unethical or ethical action they had previously performed. After reflecting on this action, participants were asked to rate how bright the room they were in was using a 7-point Likert scale. In line with the “darkness” hypothesis, participants in the unethical condition rated the room they were in as less bright than participants that thought about an ethical action. However, Firestone and Scholl demonstrated that this effect owes to the demands of the experiment rather than how participants literally perceive the world. Thus, even in an experiment that is quite subtle compared to the typical manipulations in experimental philosophy, participants guessed the aim of the experiment and answered accordingly.
assume that we can only form memories for events that we have experienced. Yet, psychologists have amassed a large body of evidence that people sometimes remember events that never actually occurred (for a review, see Schacter 1995), indicating that memory is not entirely dependent on external inputs. Bartlett (1932) is often credited with reporting the first experimental evidence for the formation of false memories. In his research, he had participants read a story and then recall it several times after subsequent delays. He reported that memories grew increasingly distorted after each recall. Since Bartlett, researchers have found evidence for the formation of false memories in list-learning paradigms (Deese 1959; Roediger & McDermot, 1995), as well as in retention of sentences (Bransford & Franks, 1971), longer prose passages (Sulin & Dooling, 1974), image sequences (Loftus et al. 1978), and videos (Loftus & Palmer, 1974). Important for our purposes, researchers have leveraged false memory to investigate the nature of mental representation and language comprehension.

Psychological research indicates that people’s memory is better for semantic information than for specific episodes or verbatim utterances (Anderson et al. 1994; Anderson & Ortony, 1975; Deese 1959; Loess 1967; Roediger & McDermott, 1995; Sachs, 1967). Even in simple experimental contexts (e.g., learning lists of words), experiences are semantically represented. In a now classic study, Roediger and McDermott (1995; also see Deese 1959) asked participants to memorize lists composed of different words that were semantically related to a single target word. When participants were asked later to recall the words they had been presented with, they were often just as likely to falsely recall the target word, which had never been presented, as often as any of the other words that actually appeared in the list. For example, when presented with a list made up of words like “glass,” “pane,” and “shade,” people recalled the target word “window,” even though the word never appeared in the list. To introduce some terminology, the
words in the list semantically activate the word “window” — which is to say that they cause people to form or retrieve stored mental representations associated with this word.

Psychologists have leveraged the relationship between false memories and semantic activation to examine language processing (e.g., Bransford & Franks, 1971; Flagg, 1976; Gentner, 1981). In particular, prior research investigated how semantic information is combined to form structured representations of discourse, what we will call discourse meaning. This process, sometimes called semantic integration (Bransford & Franks, 1972), enables people to comprehend complex ideas communicated through connected discourse. Early research by Sachs (1967) found that memory for the meanings of sentences are more robust than memory for their specific wordings. To test this, Sachs asked participants to read passages and then tested their recognition for sentences either immediately or after they had read different amounts of intervening material. Some of the tested sentences had actually appeared in the text, but others were altered semantically or syntactically. When the meanings of the sentences were changed, participants made few errors; even after substantial distraction, participants rarely reported memory for sentences that had not appeared in the passage. However, when the changes were syntactic (e.g., a shift from active to passive voice), participants often reported that the new sentences had appeared in the passage they read. After distraction, their recognition performance was near chance. Sachs concluded that during language processing, the original form of presented material is stored temporarily, only long enough to be comprehended, whereas the material’s meaning is encoded into long-term memory. If semantic information is integrated during language processing and it is the meaning of a passage that is encoded into memory, then memory ought to exhibit productivity. That is, it should be possible for exposure to several basic, interrelated sentences to produce false memory for a sentence that expresses the integrated
representation. Subsequently, several studies have confirmed this prediction, indicating that
people integrate simple sentences to form representations for more complex sentences during
language comprehension (Bransford & Franks, 1971; Cofer, 1973; Flagg, 1976). Additionally,
people have been found to integrate information from text passages read during an experiment
with their general background knowledge, leading to false recall for additional information that
was not experimentally presented (Owens et al. 1979; Sulin & Dooling, 1974; Thorndyke, 1976).
To explain these findings, Gentner (1981) examined a model of language processing in which
sentences are considered both individually and in the broader context in which they appear. Her
model states that when a sentence is read within the context of a larger passage, the discourse
meaning that a reader forms may incorporate information not contained in the original sentence.
She focused her investigation on an examination of the integration of verb meanings in context.

Following research in linguistics (e.g., Chafe 1970), artificial intelligence (e.g., Schank,
1972, 1973), and psychology (e.g., Miller & Johnson-Laird, 1976; Stillings, 1975), Gentner
hypothesized that complex verb meanings can be represented by networks of subpredicates that
express semantic relationships. Crudely put, a verb’s subpredicates are simpler verbs that
function as components of the more complex verb’s meaning. To illustrate, consider the
relationship between the verb “give” and the more specific verb “pay.” On Gentner’s analysis,
“giving” some item is to take some action that transfers ownership of that item to a recipient.
“Paying” is a more specific form of giving, in which the giver owes the recipient. Thus, a
representation of “gave” would include subpredicates like “caused,” “changed,” and
“possession,” and a representation of “paid” would add the subpredicate “owed.” Gentner tested
this hypothesis by asking her participants to read paragraph-long stories that each included a
sentence with the verb of interest — what we will call the critical sentence. For instance, one
story contained the critical sentence, “Max finally gave Sam the money.” In the experimental condition, additional context explained that Max owed Sam money, whereas the control condition lacked this context. After reading one version of the story, participants performed a recall task in which they were shown the critical sentence with the word “gave” removed, and they were asked to fill in the word that had appeared in the story. In support of Gentner’s predictions, participants who had been provided with the additional context were more likely to falsely recall the verb “paid” than participants in the control condition. Gentner (1981) used a false recall paradigm to examine how verbs with known meanings are integrated during language processing. We now will discuss how we can extend this paradigm in order to investigate philosophically significant concepts.

2.2: Semantic Integration as a method for examining concepts

On a traditional view, many philosophical concepts are complex mental entities constituted by simple concepts. The simple concepts jointly provide a “definition” of the complex concept. This means that the constituent concepts express properties that provide necessary and jointly sufficient conditions for the instantiation of the concept. In the terms of semantic integration research, the traditional view makes the prediction that a concept C will have subpredicates that are the constituents of C. For example, a view which says that KNOWLEDGE is a complex concept constituted by JUSTIFIED TRUE BELIEF will make the prediction that these constituent concepts, expressing necessary conditions, will be subpredicates for KNOWLEDGE. Thus, researchers can test whether including these subpredicates in a passage leads to false recall for words picking out KNOWLEDGE, offering evidence that these subpredicates were integrated to produce the concept. When this integration occurs, researchers can infer that the concepts JUSTIFIED TRUE BELIEF are constituents of KNOWLEDGE. This
application of semantic integration is straightforward because we are able to propose a jointly sufficient set of constituents for the concept. However, there may be other situations where this is not possible. For instance, some constituents of a complex concept might be unknown. Alternatively, some concepts may be simple, or may be sensitive to nonconstituent parameters, or may be context-sensitive. Fortunately, Gentner (1981) showed that concepts could play the role of subpredicates for a target concept even when they are not jointly sufficient or necessary for instantiating the target concept. For example, she shows that people falsely recall “painting” when they integrate “working” and “workers are carrying brushes, whitewash, and rollers.” Strictly speaking, these features are not jointly sufficient for painting. The workers might have carried the whitewash but ended up working on something unrelated to painting. Of course, we expect participants in semantic integration tasks to understand the story in a plausible way. As a result, there is no requirement that the items which are integrated to yield a concept actually form a jointly sufficient set for that concept in some strong metaphysical sense. Nor is there a requirement that the items playing the role of subpredicates in integration correspond to necessary conditions for the concept at issue. Recall Gentner’s example: “Carrying brushes, whitewash, and rollers” is not a necessary condition for painting (e.g., think of spray painting). All that is required is that the context makes it more or less likely that the target concept is instantiated. This is good news for three reasons:

1. Complex concepts can be examined even if some of their constituents are unknown. If we are interested in studying a complex concept, we can examine participants’ integration of a set of concepts that merely approximate its true constituent concepts. For example, suppose that a concept C has constituents X1, X2, and X3 and we want to test whether
X1 is a constituent. We can test for false recall of C in the presence of X1 and X2 without invoking X3 or by approximating X3.

2. Some philosophers have argued that many philosophically interesting concepts are simple (e.g., Fodor, 1998; Williamson, 2002). Yet, these concepts may still have interesting necessary conditions. For example, Williamson (2002) holds that although KNOWLEDGE is simple, the concept still has philosophically important necessary conditions. Many philosophers hold that a necessary condition for S knowing P is that P be true. Semantic integration can be used to examine whether the necessary condition is something that leads people to false recall the presence of that the word denoting the target concept.

3. Some philosophers think that some concepts are sensitive to certain parameters and that this sensitivity is accessible to lay people. For example, Knobe (2010) holds that competent folk mental state attributions are sensitive to the moral valence of the content attributed, and some epistemologists have claimed that competent folk knowledge ascriptions are sensitive to practical interests (Pinillos, 2012; Stanley & Sripada, 2012) and moral properties (Beebe & Buckwalter, 2010). In these cases, parameters like moral valence and practical interests do not necessarily constitute interesting necessary conditions for the concept. Yet, semantic integration is still apt for testing these parameters. We can do this by developing vignettes that include a critical sentence whose truth, together with a parameter, is thought to yield the target concept. If people consider the parameter to be relevant to the target concept, then the presence of the parameter ought to lead to greater false recall for words that lexicalize the concept (see Henne & Pinillos, in preparation; Waskan et al., 2014 for further examples for how semantic integration can produce these results).
What these three points reveal is that the viability of the semantic integration method does not depend on any particular understanding of concepts. On the contrary, the method is applicable under a wide variety of assumptions about concepts. The versatility of the method is then especially useful for philosophers who themselves might disagree about the very nature of concepts.

Section 3: Two experiments using semantic integration

In the remainder of this section, we discuss two experiments we conducted that demonstrate how semantic integration can be used to investigate philosophically significant concepts. In this research we focus on the concept KNOWLEDGE, but recently other researchers have adopted our method in order to examine scientists’ concept of EXPLANATION (Waskan et al., 2014) and CAUSATION (Henne & Pinillos, in preparation).

There are three main components in a semantic integration study. The first component is the passage containing the contextual information to semantically activate the target concept. In order to construct passages that yield false recall of KNOWLEDGE, we altered contextual information in different versions of a main story, controlling for word count, sentence length, and overall structure. In a preliminary study, we constructed two versions of a story about a detective (Jack Dempsey) who forms the belief that a suspect (a teenager named Will) is guilty. In the experimental condition, the detective’s belief is justified by legitimate evidence and his belief is true (the suspect is in fact guilty). In the control condition, the detective cannot find any evidence and participants are not told whether the suspect is guilty, but the detective forms the belief anyway.
In each of these stories, we included a sentence containing a critical verb (shown below). Recall that when sufficient contextual information licenses using a more specific verb, people will falsely recall the more specific verb as having appeared in the passage. The critical verb must be consistent with the concept under investigation, but must not entail it. In our knowledge experiment, we chose “thought” as our critical verb since thinking that P is consistent with knowing that P, but does not entail knowing that P. We predicted that when people read a sentence containing thought in the right context, it would lead to false recall of the word “knew.”

Critical sentence: “Whatever the ultimate verdict would be, Dempsey thought Will was guilty.”

We predicted this will occur more frequently in the experimental condition where the appropriate context is supplied than in the control condition.²

The second component of a semantic integration study is a distractor task. In principle, this distractor task could consist of almost anything. The purpose of the distractor is simply to diminish the effect of episodic memory in the recall task. Importantly, however, distractors should not contain either the critical verb or the target word. After reading the distractor, participants advance to the third part of the experiment, the recall task. There they are shown several sentences from the passage that we manipulated, each with one word removed. Their task is to recall the word that appeared in the blank. In our experiment, we were interested in their

² We should note that an additional consideration when choosing a critical and target word is the frequency with which that word occurs in English communications. Generally, it has been found that recall performance is improved for high-frequency than for low-frequency words, and that the opposite is true for recognition performance (Kintsch, 1970). That said, there is some evidence that low-frequency words might benefit at recall when they presented together with high-frequency words (Duncan, 1974; Gregg, 1976), as will likely be the case in semantic integration experiments. A good practice is to ensure that critical and target words are matched for frequency of occurrence as closely as possible. “Thought” and “knew” are reasonably well matched as the 179th and 300th most common English words, respectively (Wolfram|Alpha 2013a, 2013b).
recall performance for the critical sentence. During the recall task, participants were shown this sentence with the word “thought” replaced with a blank, as shown below:

*Recall Task: “Whatever the ultimate verdict would be, Dempsey ______ Will was guilty.”*

Participants typed in the word that they recalled as having appeared in the original story. Consistent with our predictions, participants were more likely to recall “knew” as having appeared in the sentence when the detective’s belief was justified and true (Powell, Horne, Pinillos, & Holyoak, 2013). Clearly, this finding does not demonstrate anything particularly interesting about KNOWLEDGE, but it does demonstrate that the semantic integration paradigm can be extended to examine philosophical concepts. Subsequently, we have begun investigating people’s concept of knowledge in Gettier cases, a clearly more substantive issue in philosophy (Powell, Horne, Pinillos, & Holyoak, 2015).

Section 3.1: Pragmatics and demand characteristics

Semantic integration tasks offer two important advantages over more explicit survey methods. For one, semantic integration tasks avoid many of the concerns raised by Cullen (2010) over pragmatic cues. Researchers using survey methods need to account for pragmatic cues in the stimuli that they present to participants as well as in their instructions, questions, and response options. In a semantic integration experiment, participants are told they are performing a memory task and nothing in the instructions, response prompts, or options indicates otherwise. While these materials are not devoid of pragmatic cues, pragmatic factors in this context are less problematic and better understood. Psychologists have studied memory since Ebbinghaus (1885/1964), and have developed reliable methods for testing people’s recollection of presented material. While it is clear that stimuli may still contain pragmatic cues and conversational
implicatures, this fact is not in any way unique to semantic integration. For one, survey methods will also face these same concerns. Moreover, if one were skeptical about an experimental paradigm for this reason, one would also have to be skeptical about research on causal reasoning, decision-making, psycholinguistics, or nearly any line of research that involves presenting text to participants. The pressing concern is that pragmatic cues in instructions will lead participants to approach the experimental task incorrectly, or to interpret their response options in a manner inconsistent with the researcher’s intentions. Semantic integration tasks avoid these difficulties.

Second, semantic integration tasks largely preclude demand characteristics. Even if participants are apprehensive about being evaluated, their apprehension is unlikely to lead researchers to erroneous conclusions about the importance of their manipulation. Evaluation apprehension should motivate participants to perform the task well, and since there is no reliable way for participants to produce “desirable” answers except by probing their own memory, there is little risk of evaluation-apprehension leading to spurious findings. Rather, the method is quite conservative in that if participants perform the task optimally, then researchers ought to observe no differences between conditions since the correct answer in both conditions is exactly the same.

Finally, because the memory task is both intelligible and naturalistic, participants are less likely to take on the role of the faithful participant (Weber & Cook, 1972). Even if some participants do ignore experimenters’ conversational implicatures, this is unlikely to affect their performance, as the instructions of a memory task can be made comprehensible without many contextual cues.
Section 3.2: Caveats

The interpretation of findings from semantic integration tasks depends on resolving two questions:

(a) How are concepts structured?

(b) What mental process leads to integration of semantic information?

The structure of concepts

If semantic integration directly measures the semantic activation of people’s concepts, then one might wonder about the nature and structure of these concepts. As discussed, Gentner (1981) hypothesized that verb concepts are represented as structured collections of subpredicates. On the basis of this view, she made and confirmed very specific predictions about how representations would be combined during the processing of connected discourse, lending support for her theory. Still, psychologists have attempted to describe concepts using a number of representational formats (e.g., Posner & Keele, 1968; Medin & Schaffer, 1978). This may prompt some to doubt that Gentner’s model of concepts is accurate, or to worry that, even if it accurately describes the representations of certain concepts, different types of concepts may be represented in other ways (e.g., natural kind terms, prototype or exemplar models, distributed representations, etc.). Although these possibilities may complicate the interpretations of semantic integration experiments, researchers who use semantic integration can remain agnostic to the “true” psychological theory of concepts. In fact, the method rests on two basic assumptions: (1) semantic concepts are mentally represented in some fashion and (2) memory for the meaning of a passage is more robust than memory for its exact wording. The first claim is a fundamental assumption of modern psychology and one that we will not defend. The second is supported by a
large body of research on memory, some of which we discussed in Section 2 (e.g., Bransford & Franks, 1971; Brewer, 1977; Barclay, 1973; Cofer, 1973; Flagg, 1976; Sachs, 1967).

*Mental processes and semantic integration*

Thus far we have reasoned as if integration occurs during comprehension and encoding, but another possibility is that integration actually occurs at recall. That is, during encoding people store the meanings of individual propositions separately. Then, at recall, they integrate these meanings by a process of inference to form a reconstruction of the memory for an individual sentence or proposition. Supposing this is true, it is worth noting that semantic integration still overcomes concerns about demand characteristics and pragmatic cues. However, it can no longer be said to provide a more direct a measure of semantic activation. Rather, in this case, the responses that participants give to recall prompts are just as dependent on inferential processes as responses to surveys are. Fortunately, Gentner (1981) tested this possibility by inserting contextual information both before and after the critical sentence in a passage. She found that false recall for critical items was greater when the inserted material came before the critical sentence, supporting the interpretation that meanings are integrated online during discourse comprehension rather than after the fact during recall. This supports the claim that semantic integration isolates conceptual activation from downstream decision-making processes.

**Section 4: Alternate experimental designs and surveys**

*Similar experimental paradigms*

In this chapter we described an experimental method modeled on Gentner’s (1981) work on the semantic integration of verb meanings, and described its use for examining people’s concept of KNOWLEDGE. It bears noting that there are a number of other related experimental
paradigms that have been used to examine semantic integration in discourse comprehension (e.g., Bransford & Franks, 1971; Brewer, 1977; Barclay, 1973; Cofer, 1973; Flagg, 1976; Sulin & Dooling, 1974; Thorndyke, 1976; Owens et al. 1979). However, Gentner’s (1981) paradigm has several qualities that are desirable for experimental philosophers, even relative to other semantic integration tasks.

First, the use of a free recall task makes its results more compelling than tasks that rely on recognition judgments. Participants’ responses to recognition tasks can be influenced by both true recollection as well as mere feelings of familiarity (Tulving, 1985). In contrast, explicit recall of the word “knew” provides unambiguous evidence for the semantic activation of the concept KNOWLEDGE, especially when the demands of the experiment push participants to recall the actual word that appeared in the story (i.e., “thought”).

Second, this paradigm focuses responses onto a single specific word of interest, whereas other semantic integration paradigms often ask participants to evaluate larger semantic units, such as phrases or sentences (e.g., Bransford & Franks, 1971; Sulin & Dooling, 1974). Specifying a target verb can reduce ambiguity in investigations of individual concepts. Thus, where possible, the semantic integration tasks we’ve discussed may be a superior method for examining the parameters involved with instantiating people’s concepts. Of course, not all concepts of interest will necessarily have a verb form (“knew”), with nearby synonyms (“thought,” “believe”). Where this is not the case, other semantic integration tasks may be more appropriate. The disadvantages associated with semantic integration tasks measuring recognition for sentences or phrases (e.g., Bransford & Franks, 1971; Owens et al., 1979) are not insurmountable. In particular, memory researchers have developed procedures, like the Remember-Know procedure (Tulving, 1985), that can help distinguish between genuine
recollection and familiarity. With sufficient care, phrases or sentences can be crafted to unambiguously express whatever concept may be of interest to researchers (e.g., Waskan et al., 2014). In the next chapter, we will discuss one such set of experiments.

Surveys and semantic integration

The methodological advantages of semantic integration owe to the indirect nature of the task. However, this also marks semantic integration tasks as importantly different from the explicit measures collected during survey tasks. Different research questions might warrant the use of either surveys or semantic integration. Many experimental philosophers hope to assess philosophical arguments by examining the psychology of concepts they employ. We have argued that, in general, semantic integration tasks are well suited for accomplishing this goal. Semantic integration tasks provide an implicit measure of conceptual activation, making them ideal for capturing these sorts of intuitive reactions.

However, some philosophical concepts may also be applied to situations by more effortful cognitive processes. In these cases, explicit survey questions that are consciously considered may be better suited if these questions can be adequately constructed. Additionally, surveys may be more appropriate where experimental philosophers are interested in people’s judgments.

Section 5: Conclusion

In this chapter, we discussed the ways in which pragmatic cues and demand characteristics can affect the results of surveys. In light of these problems, we argued that experimental philosophers should adopt a new experimental paradigm that we call semantic integration. Our experimental investigations of KNOWLEDGE demonstrate how this method can
be used to examine philosophical concepts. Semantic integration can be applied to investigate complex concepts in a manner consistent with the aims of traditional conceptual analysis, and used to examine other parameters relevant to the instantiation of concepts. This method avoids concerns about pragmatic cues and demand characteristics because participants’ conceptual activation is measured implicitly through a memory task. For these reasons, semantic integration represents an important methodological advance in experimental philosophy.\(^3\)

---

\(^3\)The proceeding chapter was adapted from a longer paper on this topic written with Derek Powell and N. Angel Pinillos.
CHAPTER 3
EVIDENCE FOR ANTI-INTELLECTUALISM ABOUT KNOW-HOW FROM A SEMANTIC INTEGRATION TASK

1. Introduction

One of the central aims of epistemology has been to analyze the concept of knowledge. Several types of data have been brought to bear in pursuit of this project, including analyses of epistemic language (e.g., Stanley & Williamson, 2001), philosophers’ intuitive responses to thought experiments, and more recently, the use of psychological methods to investigate how non-philosophers’ conceive of knowledge (e.g., Buckwalter, 2010). This latter methodology is often employed with an eye towards determining if, and to what extent, philosophers’ theories of knowledge are similar to people’s conceptions of knowledge. Are philosophers’ theories of knowledge revisionary, or do they articulate people’s concept of knowledge more generally?

However, like epistemology in general, this methodology has primarily been employed to investigate propositional knowledge, that is, knowledge that a proposition is true (e.g., knowledge that the Earth is a sphere). While knowing-how has received increased attention in the last decade, much of this attention has been dedicated to establishing intellectualism, the thesis that knowing-how is a species of, or reducible to, propositional knowledge (e.g. Stanley & Williamson, 2001; Stanley, 2011; Bengson and Moffett, 2007; Bengson, Moffett, & Wright, 2009). This is quite surprising as prior to this shift in thinking about knowing-how, many philosophers and psychologists assumed that procedural knowledge (or knowing-how) and declarative knowledge (or knowledge-that) are distinct (e.g., Ryle, 1945; Ryle, 1949; Hawley, 2003; Noë, 2005; Wallis, 2008). This view is sometimes known as anti-intellectualism.

In mainstream epistemology, the claim that there is a distinction between knowing-how
and knowing-that has been rejected on both linguistic and psychological grounds in three high profile papers.\(^4\) First, Stanley and Williamson (2001) and Stanley (2011), draw on Karttunen’s (1977) linguistic analysis of embedded questions to argue that sentences containing know-how ascriptions should be analyzed as ascriptions of propositional knowledge. For instance, on Stanley and Williamson’s view, sentences of the form

(1) Hannah knows how to ride a bicycle.

should be analyzed as:

(2) Hannah knows that some way, \(w\), is a way for her to ride a bicycle.

Since (2) is a propositional knowledge ascription, and because, according to Stanley and Williamson, it is the correct treatment of (1), Stanley and Williamson conclude that knowing-how is a species of, or reducible to, knowing-that.\(^5\)

Second, Bengson, Moffett, and Wright (2009) reject the distinction between knowing-how and knowing-that on the basis of psychological data purporting to show that people have an intellectualist conception of know-how. They claimed to demonstrate that people’s concept of know-how is intellectualist with a simple observational study: Participants were presented with a short story about a ski instructor who could not perform the ski stunts he successfully taught others how to perform. After reading this short story, participants were asked whether the ski instructor knows how to perform the stunts. Participants were only asked this and two other reading comprehension questions about the short passage they read. Bengson and colleagues found that, consistent with intellectualism, the majority of their participants judged both that the ski instructor knows how to perform the stunts and that he does not have the ability to perform them himself.

\(^4\) This is not an exhaustive review of the arguments for intellectualism.
\(^5\) For Stanley and Williamson’s complete account see (Stanley & Williamson, 2001; Stanley, 2011).
While the work reviewed above has been quite influential, we believe it falls short of providing compelling evidence for intellectualism. In the next section, we discuss some challenges to this evidence, beginning with the linguistic evidence offered by Stanley and Williamson.

2. Challenges to Intellectualism

In response to Stanley & Williamson’s account of know-how, some philosophers have questioned whether linguistic analyses have any relevance to the question of whether knowing-how and knowing-that are distinct (Noë, 2005). For instance, Noe writes, “Why should linguistic analysis be regarded as dispositive in matters like this? Is it not a home truth of analytic philosophy that grammar can mislead? What does the grammar have to do with what we are talking about or thinking about or studying when we study practical knowledge?” (Noë, 2005 p. 286). Along these lines, Noë raises compelling worries that Stanley and Williamson’s linguistic analysis is not descriptively accurate.

If Stanley and Williamson’s linguistic analysis is descriptively accurate, then most English speakers should judge that, for instance, a ski instructor knows how to perform ski stunts in the absence of having the ability to do them. According to their analysis, knowing how to perform the stunts just requires having the right propositional knowledge. One can have the ability to do them as well, but this is not required. Of course, this is an empirical question and so Noë’s evaluation of the ski instructor case is, in and of itself, purely speculative.

As we discussed, a recent observational study seems to put Noë’s empirical prediction to the test. Recall that Bengson and colleagues found that a majority of participants made judgments consistent with intellectualism. Nevertheless, the study in question suffers from a straightforward methodological problem.
The problem with Bengson and colleague’s study is that, because the aim of the study was in no way concealed from its participants, and participants were only asked one substantive question about the passage they read, it is quite likely that their results can be explained by demand characteristics of the experimental paradigm (e.g., Cullen, 2008; Firestone & Scholl, 2014; Orne, 1962; Powell, Horne, & Pinillos, 2014). When the aim of the task is not concealed from participants, then this can remove the effects of random assignment. In the case of an observational study, concerns about the effects one observes being the product of demands are greatly increased since there is no control group. Thus, it is difficult to determine whether participants were clued into the aim of the experiment and made judgments consistent with the experimenters’ hypotheses or whether the results truly reflect people’s concept of know-how. Consequently, we think gathering additional empirical evidence is required to determine whether people’s concept of knowing-how is consistent with intellectualism or anti-intellectualism.

3. The Present Studies

Our interest in this chapter is primarily psychological. However, we also assume there is an important connection between the psychological and the semantic evidence for intellectualism though we will not argue for that claim here. Instead, we intend to investigate whether there is psychological evidence for intellectualism. In order to avoid the criticisms we raised in section 2, we used an implicit measure of conceptual activation that avoids concerns about demand-characteristics.

Rather than using a standard survey measure, we investigated people’s concept of knowledge-how using a semantic integration task (Gentner, 1981; Powell et al., 2013; Powell et al., 2014; Waskan et al., 2014). Semantic integration is the cognitive process by which semantic information is combined to form structured representations (Gentner, 1982). Prior psychological
research has shown that when people read sentences within a larger context (i.e., a passage), individual sentences are not encoded in isolation, but instead their meanings are concatenated to form a semantic whole (e.g., Bransford & Franks, 1974). Other related studies have shown that people's memory for semantic content tends to be more robust than their memory for the syntactic structure of individual sentences (e.g., Sachs, 1967). Thus, when participants are asked to remember a passage of text after a delay, what they remember reflects their interpretations of the passage they read rather than the actual sentences they read. For example, Bransford and Franks (1971) found that reading several interrelated sentences (e.g., “The frog was on the log” and “The fish swam under the log”) caused participants to falsely recognize sentences that expressed a description that could be inferred from a combination of those two sentences (e.g., “The fish swam under the frog”). In a task that was analogous to Bransford and Franks’ task, we investigated people’s concept of know-how by using a sentence recognition task as a measure of semantic integration.

Semantic integration imposes fewer demand characteristics on participants than traditional survey methods because participants are led to believe they are completing a reading comprehension or memory task. Moreover, semantic integration tasks do not raise any concerns that participants’ responses are the product of unthoughtful reactions to philosophical thought experiments; a concern that has been raised by several philosophers about experimental philosophy in general (e.g., Bengson, 2013; Kauppinen, 2007). Instead, the data from integration tasks appears to reflect the concepts activated when participants passively encode the story they read. In short, unlike traditional survey measures, data from semantic integration tasks is not subject to the worry that it is the product of responses that are unreflective of participants’ concepts.
In the present studies, we used a sentence recognition task as an implicit measure of semantic integration. Participants were asked to read a story about a ski instructor (Experiment 1) or a chess instructor (Experiment 2). In a 2x2 between-subjects design, participants read a story in which the instructor has both propositional knowledge about and an ability to perform the task in question (KNOW-HOW), only propositional knowledge (INT) but no ability, only the ability in question (ANTI) but false beliefs about how one performs the stunts, or neither (NO KNOW-HOW). Following prior semantic integration designs (e.g., Powell, Horne, Pinillos, & Holyoak, 2015; Waskan et al., 2013), after reading a distractor story and answering comprehension questions about it, participants were asked to indicate whether a series of sentences had appeared in the story about the instructor. Our hypothesis was that the story people read (i.e. the condition they were assigned to) would influence what they encoded about the instructor. Differences in what participants encoded about each story would be exhibited through their false recognition of a sentence that measures people’s concept of know-how. Henceforth, we refer to this sentence as the critical sentence.

Predictions

We predicted that participants would incorrectly report that a sentence had appeared in the story (false alarm) when the sentence fit with what they semantically encoded about the story. Specifically, we hypothesized that a sentence that said that an instructor knew how to perform a task would be falsely recognized by participants as present in the story they read when the sentence fit with what they semantically encoded about the story. For example, we predicted that if participants read a story in which an instructor was described as having the ability to perform ski stunts and propositional knowledge about how he performed these stunts, people would falsely recognize that a sentence saying the instructor knew-how to perform ski stunts...
appeared in the story (even though this sentence does not appear in the story). By comparing participants’ false alarm rates on the critical sentence across conditions, as well as their overall false alarms on other control sentences, we were able to measure the degree to which these stories led participants to encode that the instructor knows how. Thus, using this method we are able to determine which factor – ability or propositional knowledge – leads participants to more frequently encode that an instructor knows-how.

Our main predictions were that there would be a main effect of ability on the proportion of false alarms to the critical sentence, and no effect of propositional knowledge on false alarms to the critical sentence. In addition, we predicted that there would be a greater proportion of false alarms on the critical sentence for the ANTI condition (short for anti-intellectualism) than for the INT condition (short for intellectualism).

4. Experiment 1

Participants

Participants were 208 workers recruited on Amazon Mechanical Turk. Fifty percent of participants were female. The mean age of participants was 35.6 years old.

Procedure and Materials

Participants read one of four versions of a story about a ski instructor, Patrick, teaching a student how to perform ski stunts. Participants’ task was to read this story and remember it to the best of their ability. After reading a distractor story (a fake popular science article on gamma ray bursts) and answering comprehension questions about it, they were asked to judge whether a series of sentences (randomly ordered) appeared in the story about the ski instructor. During this sentence recognition phase, sentences were presented to participants in bold and they were forced to choose either, “Yes, this sentence appeared in the story,” or, “No, this sentence did not
appear in the story.”

In addition to presenting the critical sentence to the participants, we also included several control sentences. One of these sentences (the teacher sentence) read, “Patrick was definitely a great teacher.” This was included to ensure that participants did not assume the ski instructor was a better teacher in virtue of having the ability to perform the stunts in conditions where he had this ability. It was important to include this because a difference in how participants perceived the quality of Patrick as a teacher could spuriously drive a difference in false alarm rates to the critical sentence. The other control sentences were included to rule out the possibility that some versions of the story were more difficult to remember than others.

**Experiment 1 Results**

Of the original 208 participants, two participants were removed for failing to pay attention. These participants were removed before performing subsequent analyses. However, all of our effects (and non-effects) remain unchanged if these participants are included in the analyses.

![Figure 1. The proportion of false alarm rates for the critical sentence “It was clear that Patrick knew how to perform ski stunts”](image-url)
Recall that we predicted that we would observe a main effect of ability on false alarms for the critical sentence, but no effect of propositional knowledge and no interaction between these factors. Because our main dependent variable was dichotomous – whether you false alarmed to the critical sentence – we used logistic regression to examine the effect of ability and propositional knowledge on false alarms to the critical sentence. Consistent with our predictions, this analysis revealed a main effect of ability on false alarms to the critical sentence (Wald $\chi^2 (1) = 13.705, p < .001$). Contrary to the intellectualist account of know-how, we observed no main effect of propositional knowledge on false alarms for the critical sentence (Wald $\chi^2 (1) = .131, p = .718$), nor did we observe a significant interaction between these factors (Wald $\chi^2 (1) = .016, p = .90$).

Since it is possible this main effect could be driven by the KNOW-HOW condition, we wanted to confirm that the ANTI condition elicited a higher proportion of false alarms than the INT condition. A $\chi^2$ test revealed that the ANTI condition had significantly more false alarms to the critical sentence than the INT condition ($\chi^2 (1, N = 101) = 4.91, p = .026$).

*Table 1. The proportion of false alarms for the Teacher and the Control sentences.*

<table>
<thead>
<tr>
<th>Ability</th>
<th>Propositional Knowledge</th>
<th>Teacher Sentence</th>
<th>Control Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False Alarms</td>
<td>SE</td>
<td>False Alarms</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>.47</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.48</td>
<td>.067</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>.62</td>
<td>.071</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.21</td>
<td>.056</td>
</tr>
</tbody>
</table>
As shown in Table 1, these effects cannot be attributed to an overall difference in recognition performance since we observed no main effect of condition on control sentence performance, though this effect was marginally significant \(F(3, 204) = 2.49, p = .058\), nor can it be attributed to any difference in participants attributing more skill to Patrick as a teacher since the ANTI condition did not differ from the INT condition on false alarms for the teacher sentence \(\chi^2(1, N = 101) = 2.39, p > .1\).\(^6\)

Our results indicate that, contrary to prior empirical research on people’s concept of knowledge-how (e.g., Bengson et al, 2009), people encoded that an instructor knew how when the instructor had the ability to perform the stunts, but false beliefs about how he performed them, over having the correct beliefs about how to perform the stunts.

5. Experiment 2

Participants

Participants were 213 workers recruited on Amazon Mechanical Turk. Fifty-One percent of participants were female. The mean age of participants was 33.3.

Procedure and Materials

Experiment 2 used the same methods and analyses as experiment 1. However, rather than using the ski instructor story we used a story about a chess instructor. Chess is, intuitively, a more intellectual activity than skiing, and we wanted to test the possibility that people’s concept of know-how might be task-specific. In other words, one possibility is that people think that knowing how to do physical tasks, such as ski stunts, consists in having the appropriate ability,\(^6\)

\(^6\) In fact, participants were more likely to encode that Patrick was a great teacher in the intellectualism condition than the anti-intellectualism condition, though this difference was not reliable. Moreover, participants’ overall control sentence score was actually better (though non-significantly) in the ANTI condition than the INT condition, so it is not as if participants in the ANTI condition were simply worse at remembering what occurred in the story than in the INT condition.
whereas they may think that knowing how to do more intellectual activities, such as playing chess, consists in declarative knowledge of rules and strategies. Nevertheless, if an anti-intellectualist story still leads to a higher proportion of false alarms for the critical sentence in a paradigmatically intellectual activity, this would indicate that people still think knowledge-how is more tightly connected to having an ability than to having propositional knowledge. Thus, in experiment 2, Patrick was a chess instructor who trained a student how to play chess. Everything else about the task was the same.

Results

Of the original 213 participants, two participants were removed for failing to pay attention. These participants were removed before performing any analyses. However, all of our effects (and non-effects) remain unchanged even if we include these participants in our analyses.

![Figure 2: The proportion of false alarms for the critical sentence “It was clear that Patrick knew how to play chess”.

As in experiment 1, our dependent variable was dichotomous so we again used logistic regression to examine the effect of ability and propositional knowledge on false alarms to the
critical sentence. Consistent with anti-intellectualism about know-how, this analysis revealed a main effect of ability on false alarms for the critical sentence (Wald $\chi^2 (1) = 12.49, p < .001$).

Against the intellectualist account of know-how, we observed no effect of having propositional knowledge about chess on false alarms for the critical sentence (Wald $\chi^2 (1) = .13, p = .718$), nor any interaction between these factors (Wald $\chi^2 (1) = .708, p = .40$).

Following up on this main effect, a $\chi^2$ test revealed that the ANTI condition again led to significantly more false alarms to the critical sentence than the INT condition ($\chi^2 (1, N = 110) = 8.97, p = .002$), suggesting that the main effect of ability was not solely driven by the KNOW HOW condition.

Table 2. The proportion of false alarms for the Teacher and Control sentences.

<table>
<thead>
<tr>
<th>Ability</th>
<th>Propositional Knowledge</th>
<th>Teacher Sentence</th>
<th>Control Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>.56</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.58</td>
<td>.22</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>.58</td>
<td>.20</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>.14</td>
<td>.22</td>
</tr>
</tbody>
</table>

We also found that this effect cannot be attributed to an overall difference in the difficulty of remembering these stories since we observed no effect of condition on control sentences ($F (3, 207) = .227, ns$), nor can it be attributed in any difference in participants attributing more skill to Patrick as a teacher, since there was no significant difference in false alarm rates for the teacher sentence comparing the ANTI and INT conditions ($\chi^2 (1, N = 110) = .004, ns$).
Contrary to the conjecture that people will have a more intellectualist conception of know-how in a paradigmatically intellectual domain, reading the ANTI story led to more false alarms for the critical sentence than reading the INT story. Even in the case of chess, participants semantically encoded that the instructor knew how to play chess when he had the ability to play chess (and false beliefs about how he executed chess moves). On the other hand, participants false alarm rates in the INT condition were not significantly higher than they were in the NO KNOW-HOW condition.

6. Discussion

In the last several years the question of what types of mental states are constitutive of knowledge-how has received attention from epistemologists and philosophers of mind. Intellectualists have argued that knowing how to, for instance, ride a bicycle consists in knowing certain kinds of propositions about bicycle riding. Anti-intellectualists, on the other hand, have argued that knowing how to ride a bicycle consists in having the ability to ride a bicycle. Here we discussed two kinds of evidence for intellectualism – linguistic and psychological. The present empirical research, however, may call this evidence into question.

First, our results directly contradict those of Bengson et al. (2009). Whereas Bengson et al. (2009) found that the majority of their participants made judgments consistent with intellectualism, we found that people are far more likely to semantically encode that an instructor knows how in cases where an instructor has the ability to do a task, despite the instructor having false beliefs about the task, than in cases where they have propositional knowledge about the task but lack the ability to perform it. This, in and of itself, indicates that there is at least equal psychological support for intellectualist and anti-intellectualist concepts of know how. However,
our study has the additional virtue of being more methodologically sound. Bengson and colleague’s results can be explained by appealing to demand characteristics of their study.\textsuperscript{7}

Finally, our results may call into question the descriptive accuracy of Stanley and Williamson’s linguistic analysis. If this analysis were descriptively accurate, then participants should have had a higher proportion of false alarms to the critical sentence for stories in which a subject has propositional knowledge (but no ability) about a task than in cases where a subject has the ability (but lacks propositional knowledge) about a task. This, of course, is the opposite of what we found. Hence, our data may suggest that their analysis is descriptively inaccurate.\textsuperscript{8}

\textsuperscript{7} In contrast, our study employs a methodology that avoids these concerns, and also avoids the additional concern that participants are making judgments that reflect pragmatic factors downstream from their concepts (a concern Bengson himself has raised about experimental philosophy, Bengson, 2013). In other words, the very structure of our task yields results that more directly reflect people’s concepts (e.g., Powell et al., 2014).

\textsuperscript{8} The proceeding chapter was adapted from a longer paper on this topic written with Ian Harmon.
CHAPTER 4
SYNTACTIC CUES INDUCE DIVERGENT EPISTEMIC INFERENCE

1. Introduction

In the last decade, an increasing number of epistemologists working at the intersection of
cognitive science and philosophy have sought to understand people’s concept of knowledge by
empirically investigating people’s uses of epistemic terms (see Knobe, 2015 for a review). Prior
to this methodological shift in thinking about how to investigate attributions of knowledge,
philosophers often argued for epistemic theories by considering whether the theory correctly
classified thought experiments and linguistic utterances containing epistemic statements into
cases of knowledge or mere belief. Although there is a general consensus among philosophers
about some of the parameters that affect whether an agent know—for instance, justification and
truth—substantial disagreement remains (e.g., Buckwalter & Schaffer, 2013; Stich &
Buckwalter, 2011; Buckwalter, 2012; Pinillos, 2012). More pressing perhaps, several
experimental philosophers have taken aim at the traditional philosophical approach by arguing
that people’s intuitions about thought experiments cannot serve as reliable support of an
epistemic theory because intuitions cannot be calibrated (e.g., Cummins, 1998) or because they
are influenced by irrelevant features that are not truth-tracking (e.g., Alexander, Mallon, &
Weinberg, 2010). In line with these prior criticisms, this chapter raises a new set of concerns
against a fundamental assumption made by most, if not all, epistemologists. We refer to this
claim as the *know-based definition* of knowledge.

The know-based definition of knowledge is that knowledge can be defined in terms of
what an agent knows. For instance, Knowledge \( =_{\text{def}} \) An agent knows some proposition just in
case \([A, B, C]\). In this definition, A, B, and C, might stand in for concepts like JUSTIFIED, TRUE,
BELIEF. While some researchers have questioned whether, for instance, belief is required to know (e.g., Murray, Sytsma, Livengood, 2013), the claim that knowledge can be defined in terms of what an agent knows has gone unquestioned. While this assumption may seem uncontroversial, in this chapter we present empirical evidence from 8 experiments that suggests ‘know’ and ‘knowledge’ may pick out genuinely distinct concepts.

What motivates this bold prediction? Ongoing research in developmental and cognitive psychology suggest that subtle syntactic differences (similar to the syntactic differences between ‘know’ and ‘knowledge’) can have significant effects on people’s inferential processes and behaviors. In the next section, we summarize some of this literature. The upshot of this research is a completely novel hypothesis about the nature of knowledge and the nature of knowing, according to which knowledge seems to require more certainty, and appears to depend on less pragmatic and contextual features of an epistemic agent’s situation than does knowing.

2. Syntactic markers affect inferences and behavior

Psychological research in developmental and cognitive psychology provides compelling evidence for the prediction that syntactic differences can have significant effects on people’s inferences and behavior.

Inferences

Gelman and Heyman (1999) investigated how linguistic labels affect children’s inferences about novel categories. In their study, these researchers either described an agent using a novel category label, for instance as a carrot-eater, or they described the same agent as “someone who eats carrots all the time and loves to eat carrots.” After describing an agent in one of these two ways, they measured participants’ beliefs about the stability of the agent’s carrot-eating tendencies. They found that children were more likely to think that the agent in the story
had an enduring and stable trait when the agent was described as a carrot-eater (i.e., the noun form) than when the agent was described as loving to eat carrots. Gelman and Heyman concluded that the label denoting a category led children to think that the tendency to eat carrots was stronger and more stable.

This result was consistent with earlier research that suggested that nouns and verbs elicit different intuitions about the strength and stability of a trait. In an earlier study by Markman and Smith (cited in Markman, 1991) participants were asked to compare nouns (e.g., John is an intellectual) and semantically matched adjectives (e.g., John is intellectual). They found that participants judged that the noun conveyed stronger information than a matched adjective. Participants explained this difference by noting that the noun seemed more central to the identity of the category (e.g., Gelman and Heyman, 1999; Markman, 1991).

These results provide preliminary support for the hypothesis that the words knowledge and knows might likewise elicit subtly different epistemic intuitions, as words that denote categories (like nouns) seem “stronger” and seem to convey more stable information about an agent.

**Behavior**

Even a cue as seemingly insignificant as a word or a phrase can have a tremendous impact on children’s behavior. For example, Cimpian, Arce, & Markman (2007) have shown that praising children’s successes on a drawing task with statements as similar as “You are a good drawer” and “You did a good job drawing” led 4-year-old children to make different inferences about the source of their success and, as a result, to behave differently when they subsequently encountered a setback (Cimpian et al., 2007). When they heard “You are a good drawer,” children were more likely to infer that they had succeeded because of a talent they possess,
which then made subsequent mistakes more difficult to overcome (because such mistakes reflected negatively on this supposed talent) and made children less likely to take on challenges (because they feared they would make more mistakes).

In a similar line of work, Bryan and colleagues have shown that asking kids to be a _helper_ versus asking them to _help_ (Bryan et al., 2014) causes them to help more with household chores. In another study, Bryan and colleagues (2011) demonstrated that asking adults the question "How important is it to you to be a voter?" was more likely to cause them to vote than asking them “How important is it for you to vote?” (also see Bryan, Adams, & Monin, 2013).

In sum, these results speak to the pervasive influence of subtle syntactic cues on people’s inferences about themselves (Bryan et al., 2013) and others (Gelman & Heyman, 1999). Moreover, they show that people use these cues to inform their subsequent behavior (e.g., Bryan et al., 2011; Bryan et al., 2013). Finally, this effect not only influences adults, but also children’s inferences and behavior (e.g., Cimpian et al., 2007; Gelman & Heyman, 1999).

Predictions

Prior psychological research suggests that syntactic differences can lead adults and children to infer different information about agents, and these inferences can affect their behavior. On the basis of these findings, we predicted that subtle differences between the category form and the verb form of knowledge (i.e., “knowledge” and “knows”) would lead to different epistemic attributions. In particular, we predicted that because words denoting categories (like knowledge) tend to be viewed as more enduring and saying something stronger (Markman, 1991), that people would think it is better to have knowledge than to know (Experiments 1-5), and it is harder to attain knowledge than to know (Experiments 6 and 7). Finally, we predicted these syntactic differences would also affect philosophers’ defenses of
epistemic theories (experiment 8). If our predictions are borne out, it may have significant implications for debates about the nature of knowledge, philosophical methodology, and perhaps may even offer some explanation for the shift away from thinking that knowledge requires certainty in the history of epistemology.

3. Experiment 1 – Knowledge and knowing

Participants. Participants in Experiment 1 were 107 Amazon Mechanical Turk workers, who were paid $0.50 for participating in the study.

Procedure. In experiment 1 we investigated whether participants distinguished between “knowing” and “knowledge”. Participants were presented with two phrases, “Knowing that ______” and “Knowledge that ______”, in which the blank was always left empty. After reading each phrase, participants were asked four questions about the epistemic implications of these phrases. The items in the scale were (1) Which of these two seems to say something stronger? (2) Which of these two seems to be more dependent on one’s circumstances? (reversed) (3) Which of these two seems to say something less definite? (reversed) (4) Which of these two seems to be more certain? At the end of each of these phrases, participants chose the knowing or knowledge phrase. The epistemic implications questions and response options were shown in randomized order. After completing this task, participants briefly explained their responses.

Results

We computed a scale average based on participants responses to the epistemic implication questions. The reliability of the scale was satisfactory (α = .74). Knowledge responses were coded as 1 and knowing responses were coded as 0. The mean scale response
was \( .60 (SD = .37) \). We compared participants’ responses against chance responding with a one sample \( t \)-test, revealing a significant difference \( (t(106) = 2.88, p = .005, 95\% CI \) of the difference [.032 to .173], \( d = .28 \)).

Consistent with our predictions, participants appeared to distinguish between the epistemic implications of knowledge and knowing. Experiment 1 thus provides preliminary support for the hypothesis that knowledge may be misrepresented by defining it in terms of what an agent knows.

4. **Experiment 2 - Comparing complete sentences**

**Participants.** Participants in Experiment 2 were 104 Amazon Mechanical Turk workers, who were paid $.50 for participating in the study.

**Procedure.** One limitation of the first experiment is that we did not present participants with complete sentences, which may have lead them to artificially differentiate knowing and knowledge. Experiment 2 was conducted to address this concern, and demonstrate that we would observe a similar effect even when participants consider two very similar complete sentences.

In experiment 2, participants were presented with two sentences and then were asked to answer a series of questions that tapped into their epistemic assumptions about two agents. The two sentences were: “Michael has a lot of knowledge” and “Christopher knows a lot.” The names associated with each sentence were counterbalanced, as was the order of the sentences.

We created a set of questions to measure whether syntactic differences between “know” and “knowledge” would lead participants to make different epistemic inferences about two agents. Participants responded to our epistemic scale, which was composed of six items, three of which were reverse coded (\( \alpha = .81 \)). The items in the epistemic scale were (1) If you had to choose, which person would you trust less? (2) If you had to choose, which person is more likely
to be a college professor? (3) If you had to choose, which person is more likely to be misinformed? (4) If you had to choose, which person do you think is more likely to be competent? (5) If you had to choose, which person do you think is more likely to make bad decisions? (6) If you had to choose, which person would you rather have your kids learn from? At the end of each of these questions, participants had to choose “Michael” or “Christopher”. We also embedded a catch question in the scale that instructed participants to select a particular response. This item was included to ensure participants were paying attention. After making judgments about the epistemic questions, participants briefly explained their responses.

Results

We excluded two participants for having IP addresses outside of the United States, and one participant for missing a catch question. After these exclusions, 101 participants remained in our analyses. We calculated a scale average based on participants six responses. Knowledge responses were coded as 1 and Know responses were coded as 0. The mean scale response was .78 (SD = .29). We compared participants’ responses against chance responding with a one sample t-test. This test revealed a significant difference in the predicted direction (t(100) = 9.76, p < .001, 95% CI of the difference [.226 to .341], d = .97). Subsequent binomial tests revealed that all six items in the scale were themselves different than chance responding in the predicted direction (ps < .001).

Even after eliminating the concerns about the previous experiment, we still found that people appear to distinguish between an agent that knows and an agent that has knowledge. These results again support the hypothesis that knowledge and knows may pick out genuinely different epistemic concepts.
5. Experiment 3 - Between-subjects manipulation

In experiment 3, we sought to rule out alternative interpretations of our findings. One possible criticism of the prior experiments is that the within-subjects nature of the task might lead participants to see the items as more different than they really are. Consequently, one possibility is that we would not observe any difference in people’s inferences when an agent has “knowledge” and “knows” in a between-subjects design. A second concern with experiment 2 is that “knows a lot” is similar to “know it all”, which has a clear negative connotation. This might lead participants to associate the positive epistemic statements with knowledge and the negative epistemic statements with knowing. Finally, one might also worry that our results were due to superficial associations between the word ‘knowledge’ and the academic or book-smart questions in experiment 2 (e.g., Who would you want your kids to learn from?; Which one is a college professor?).

We addressed all of these concerns in experiment 3. First, we predicted that we would observe a difference between participants’ responses to knows and knowledge in a between-subjects design. Second, we eliminated the negative reading of “knows a lot” by changing the sentences to “Michael has a lot of knowledge about the world” and “Michael knows a lot about the world.” Finally, we replaced the items in the scale that may be associated with book-smarts from experiment 2 with two new scale items.

Participants. Participants were 233 Amazon Mechanical Turk workers and students recruited from the undergraduate psychology subject pool from a large Midwestern research university. To ensure adequate power for our between-subjects design, additional participants were recruited to participate. Participants were paid $.50 for completing the study, or given course credit.
Procedure. Participants were presented with one sentence in a between-subjects design. Participants were randomly assigned to either the knowledge condition “Michael has a lot of knowledge about the world” or the Knows condition “Michael knows a lot about the world”. After reading one of these two sentences, participants rated their agreement on a 100-point Likert scale. The endpoints of the scale were labeled “Very Unlikely” and “Very Likely”. Participants responded to six items, three of which were reversed (α = .83). The items in the scale were as follows: (1) How likely is it that Michael would provide good advice? (2) How likely is it that Michael will succeed in life? (3) How likely is it that Michael is misinformed? (reversed) (4) How likely is it that Michael is competent? (5) How likely is it that Michael would make bad decisions? (reversed) (6) How likely is it that Michael makes unwise financial investments? (reversed) After making judgments about every item, participants briefly explained their responses.

Results

We excluded 10 participants for having IP addresses outside of the United States, and 12 participants for missing a catch question. After these exclusions, 211 participants remained in our analyses. We calculated a scale average based on participants’ responses. The mean scale responses for the knowledge condition and the know condition were 66.39 (SD = 12.14) and 63.08 (SD = 11.71) respectively. We compared participants’ average scale responses in the knowledge and know conditions with an independent samples t-test. This t-test revealed a significant difference between the knowledge and know conditions (t(209) = 2.01, p = .045, 95% CI of the difference [.070 to 6.54], d = .28). Additionally, this effect was not driven by any single question on the scale, as participants responses to all six items in the scale were in the predicted direction (p < .001). However, the straightforwardly epistemic items on the scale (e.g., How
competent? How misinformed?) exhibited the largest differences between the knowing and knowledge conditions, while other questions exhibited smaller differences (e.g., bad decisions, financial investments).

Demonstrating the robustness of the distinction between knowledge and knows, experiment 3 revealed that even when participants were only told that an agent had knowledge or that an agent knew, participants tended make different epistemic inferences about the agent. Moreover, this effect emerged despite controlling for a potentially problematic reading of the knows sentence, and changing the two (potentially) problematic items in the scale.

6. Experiment 4 - Entailment

The current results suggest that syntactic cues lead people to make different epistemic inferences when they are told that an agent has knowledge or knows. As we have noted, this result may also suggest a further claim – ‘knowledge’ and ‘knows’ may pick out semantically distinct concepts. To test this claim, we examined the intuitive claim that there is an a priori connection between what an agent knows and the knowledge that they have. The assumption is that intuitively people should equally agree with both of the following conditional claims: “If you know that P, then you have knowledge that P” and “If you have knowledge that P, then you know that P.” While we think this is a reasonable conjecture, we predicted an asymmetry in people’s agreement with these conditionals since apparently, saying someone has knowledge is stronger than saying they know. Consequently, we predicted that people would agree with the conditional “if you have knowledge that P, then you know that P” more than the conditional “if you know that P, then you have knowledge that P.” As more evidence may be required to have knowledge than to know, having knowledge should be stronger than knowing. Thus, people should think that knowledge entails knowing more strongly than knowing entails knowledge.
If this result obtained, it would suggest that not only do syntactic cues induce divergent epistemic inferences, but also that ‘knowledge’ and ‘knows’ may pick out distinct concepts.

**Participants.** Participants were 100 Amazon Mechanical Turk workers and students enrolled in an introductory psychology course at a large Midwestern research university. Participants were paid $.50 for participating in the study or received course credit for their introductory psychology course.

**Procedure.** In Experiment 4, participants were first presented with directions explaining how to understand conditional claims. After reading these instructions, participants were then presented with two practice questions about conditional statements. The practice questions were “If Harry is a cat, then Harry is a mammal.” and “If Larry is a mammal, then Larry is a cat.” The order of presentation of these conditionals was randomized.

After answering these practice questions, participants advanced to the next page and rated their agreement with two more conditional statements in a within-subjects design. The know-to-knowledge conditional read “If you know a lot about x then you have a lot of knowledge about x” and the knowledge-to-know conditional read “If you have a lot of knowledge about x then you know a lot about x.” Again, our prediction was that participants would agree significantly more with the second conditional (i.e., knowledge-to-know) than the first conditional claim (i.e., know-to-knowledge).

**Results**

We excluded 13 participants for rating the conditional “If Larry is a mammal, then Larry is a cat” greater or equal to the conditional “If Harry is a cat, then Harry is a mammal”. The principle behind this exclusion was that, at a minimum, if a participant understood the task then they ought to think that the inference from cat-to-mammal is stronger than the inference from
mammal-to-cat. However, all the effects we report remain significant without excluding these participants, \((p < .01)\). After these exclusions, 87 participants remained in our analyses. Participants' mean responses on the 10 point Likert scale was 9.89 \((SD = .45)\) for the cat-to-mammal conditional and 1.18 \((SD = 1.67)\) for the mammal-to-cat conditional. A paired sample t-test revealed that this difference was significant \((t(86) = 47.87, p < .001)\). Thus, it appears participants understood the task.

We then examined participants’ responses to the conditional from know-to-knowledge (Mean = 8.49, \(SD = 2.2\)) and from knowledge-to-know (Mean = 9.23, \(SD = 1.4\)). A second paired sample t-test also revealed participants agreed significantly more with the conditional from knowledge-to-know than the conditional from know-to-knowledge \((t(86) = -3.13, p = .002; 95\% CI of the difference \([-1.20 to -.27]\), \(d = .33\))

These results are striking as they show that, despite the entailment claim appearing quite intuitive, people agreed more with the entailment from knowledge-to-know than the entailment from know-to-knowledge. These results provide compelling evidence against the claim that there is an a priori connection between what one knows and the knowledge one has.

7. Experiment 5 - Single Proposition

In experiment 5 we sought to rule out another possible alternative interpretation of the results of our experiments. Thus far we have examined the contrast “knows a lot” vs. “has a lot of knowledge” or “knows a lot about the world” vs “has a lot of knowledge about the world”. One potential problem with this comparison is that people might make very different assumptions about the kinds of things that an agent knows. This assumption would be largely independent of any important differences between knows and knowledge. For instance, perhaps when an agent knows a lot, he has a lot of know-how rather than propositional knowledge and
perhaps people regard know-how as less reliable than propositional knowledge. If this argument is right, then people might then see an agent who knows a lot as less competent that one who has a lot of knowledge.

Another potential limitation of the prior experiments is that it is possible to get a non-factive reading of know (Hazlett, 2010), whereas no such reading may be possible with knowledge. In this case, perhaps all of our prior results can be explained by the fact that people have assumed a non-factive reading of knows, but no such reading for knowledge. In experiment 5, we controlled for these possibilities by having participants compare two people who “fully know” or “have full knowledge” that a particular proposition is true. We added “fully” in order to block any non-factive reading of know. Moreover, we had participants focus on specific propositions about ordinary topics to rule out the possibility that participants were giving different judgments because knowledge and knows are associated with different propositional contents.

**Participants.** Participants were 131 Amazon Mechanical Turk workers who were paid $.50 for participating in this study.

**Procedure.** In experiment 5, participants were presented with four different pairs of sentences about everyday topics. The four pairs of sentences participants read were:

1. Michael has full knowledge that you can clean stained coffee cups with baking powder. Christopher fully knows that you can clean stained coffee cups with baking powder.
2. Robert fully knows that avocados ripen faster in a paper bag. William has full knowledge that avocados ripen faster in a paper bag.
3. Matthew has full knowledge that apples are more effective at waking you up in the morning than bananas. Joshua fully knows that apples are more effective at waking you up in the morning
than bananas.

(4) David fully knows that a wet rubber dish-washing glove is the quickest way to clean up pet hair. James has full knowledge that a wet rubber dish-washing glove is the quickest way to clean up pet hair.

After reading these two sentences, participants were asked four questions ($\alpha = .95$):

1. If you had to choose, who would you trust more on this matter? 2. If you had to choose, who is better informed on this matter? 3. If you had to choose, who is more likely to give you bad advice on this matter? (reverse coded) 4. If you had to choose, who is less confident on this matter? (reverse coded). Following each of these questions were the corresponding names for the case. The names associated with knowledge and knew were again counterbalanced.

Results

We excluded two participants for having IP addresses outside of the United States, and 8 participants for missing a catch question. After these exclusions, 121 participants remained in our analyses. We calculated a scale average based on participants four responses. Knowledge responses were coded as 1 and know responses were coded as 0. The mean scale response was .68 ($SD = .355$). We compared participants’ responses against chance responding with a one sample t-test. This test revealed a significant difference in the predicted direction ($t(120) = 5.45$, $p < .001$, 95% CI of the difference [.112 to .240], $d = .49$). Additionally, this effect was not driven by any single question on the scale, as every scale question was itself significantly different from chance responding ($p < .001$).

Thus, even when blocking non-factive readings of know and focusing only on a single proposition, participants assumed that an agent with knowledge (compared to an agent that knows) was more trustworthy, more informed, more confident, and would give better advice.
This result again suggests that not only do syntactic differences lead to different epistemic inferences, but also that having knowledge and knowing are viewed as distinct epistemic states.

8. Experiment 6 - Lotteries

The results thus far suggest that people distinguish between knowing and having knowledge. Philosophers may raise the worry that these prior findings do not bear on philosophical debates, since we only examined participants’ responses to simple sentences. Given that our results show that these differences can be elicited in ordinary sentences, we think there is no doubt that the distinction between knowing and having knowledge will affect people’s judgments about thought experiments, and other more traditional philosophical cases. Nonetheless, we sought to demonstrate the distinction between knowing and having knowledge in classic philosophical thought experiments since it is quite possible that the distinction plays a larger (or smaller) role in some epistemic debates.

Experiment 6 served two ends: First, it extended the scope of our results by showing the differences we observed between knowing and knowledge would emerge in response to thought experiments. Second, an important test of our theory is that we would observe no differences between know and knowledge when the evidence required for knowledge (and thus knowing) is met. Thus, we predicted that people would not distinguish between knowing and knowledge in some contexts, further suggesting that our results are not due to task demands (e.g., Firestone & Scholl, 2014). Our first test case was people’s judgments about Lottery cases.

Participants. Participants were 139 Amazon Mechanical Turk workers, who were paid $0.50 to complete the experiment.

Procedure. To test our predictions, we employed a 2 x 2 repeated measures design (Vignette: between-subjects factor x Knowledge question: within-subjects factor). Participants
were randomly assigned to read either a No Chance vignette or the Chance vignette. In the No Chance vignette, Emma has no chance of winning the company raffle (because she doesn’t enter the raffle) and forms the belief that she will not win the raffle. In the Chance condition, Emma enters the company raffle but has extremely low probability (1 in 100,000) of winning and forms the belief that they will not win. After reading one of these stories, participants rated their agreement with two statements that were randomly ordered and also displayed on the same page. The knows statement read, “Emma truly knows that she won't win the company raffle.” The knowledge statement read, “Emma has true knowledge that she won't win the company raffle.” We included “truly” in both sentences to eliminate any non-factive reading of these sentences. Participants rated their agreement with these statements on a Likert scale ranging from 0 to 10 with the endpoints labeled “Completely Disagree” and “Completely Agree”.

We predicted that after reading the No Chance vignette, participants would not distinguish between the knows and knowledge question, since in both cases the threshold for knowledge that Emma will not win is evidently met (Emma did not enter the raffle so she is certain she will not win). In contrast, we predicted that participants would distinguish between knowing and knowledge after reading the Chance vignette, rating the knowledge question significantly lower than the know question.

Results

Of the 139 participants in the study, three participants had IP addresses outside of the U.S., and two missed a catch question or said they were not paying attention during the study. These participants were excluded before performing any further analyses. However, all of our results remain significant even if these participants are included in our analyses.

First, we performed a 2 x 2 repeated measures ANOVA to test whether there was an
interaction between the between-subjects Vignette factor and the within-subjects Knowledge question factor. As we predicted, we observed an interaction between Vignette and Knowledge question \((F(1, 132) = 7.12, p = .009)\). In the No Chance condition, the mean response for the know question was 9.82 and the mean response for the knowledge question was 9.65. This difference was not statistically significant \((t(67) = 1.53, p = .12, 95\% CI \text{ of the difference }[-.052 \text{ to } .405], d = .18)\). In contrast, in the Chance condition the mean response for the know question was 4.20, whereas the mean for the knowledge question was significantly lower at 3.06 \((t(65) = 3.29, p = .002, 95\% CI \text{ of the difference }[.447 \text{ to } 1.82], d = .40)\).

These findings provide compelling evidence that knowing and having knowledge may be distinct epistemic concepts, and furthermore, they suggest that both concepts may actually require different strengths of evidence to elicit their application. Finally, these results provide a strong test of the claim that people are not merely distinguishing between knowing and having knowledge artificially, as we observed no difference between knowing and knowledge in the No Chance lottery condition.

9. Experiment 7 - Gettier

Our aim in Experiment 7 was to extend our findings to other widely discussed philosophical thought experiments. The methods and analyses in Experiment 7 were identical to those used in Experiment 6.

Participants. Participants were 271 mturk workers, who were paid $0.50 to complete the experiment. Additional participants were recruited in this experiment for fear that we were unpowered for reliably detecting the interaction in experiment 6.

Procedure. To test our predictions, we employed a 2 x 2 repeated measures design (Vignette: between-subjects factor x Knowledge question: within-subjects factor). Participants
were randomly assigned to read either a Knowledge vignette or the Gettier vignette. In the Knowledge vignette, Emma is shopping for a diamond necklace at a jewelry store with honest employees, and picks a necklace from a tray labeled “Diamond Necklaces”. In the Gettier condition, Emma is shopping for a diamond necklace but the store is run by a dishonest clerk, who replaces almost every diamond with a fake diamond. Fortunately for Emma, the one she selects from the tray is a real diamond. After reading one of these stories, participants rated their agreement with two statements. The knows statement read, “Emma fully knows that the stone she bought was a real diamond.” The knowledge statement read, “Emma has full knowledge that the stone she bought is a real diamond.” Participants rated their agreement with these statements on a 0 to 10 Likert scale with the endpoints labeled “Completely Disagree” and “Completely Agree”.

We predicted that after reading the Knowledge vignette, participants would not distinguish between knowing and knowledge, as the threshold for knowing or having knowledge is met. In contrast, we predicted that participants would distinguish between knowing and knowledge after reading the Gettier vignette, and rate the knowledge question significantly lower than the know question.

**Results**

Of the 271 participants in the study, 8 participants had IP addresses outside of the U.S., and 9 missed a catch question or said they were not paying attention during the study. These participants were excluded before performing any further analyses.

First, we performed a 2 x 2 repeated measures ANOVA to test whether there was an interaction between the between-subjects Vignette factor and the within-subjects Knowledge question factor. We observed a marginally significant interaction between Vignette and
Knowledge question \( F(1, 252) = 3.35, p = .068 \). In the Knowledge condition the mean response for the know question was 7.58 and the mean response for the knowledge question was 7.57. This difference was not statistically significant \( t(126) = .091, p = .93, 95\% CI \) of the difference [-.164 to .180], \( d = .008 \). Consistent with our predictions, in the Gettier condition the mean response for the know question was 4.54, whereas the mean for the knowledge question was significantly lower at 4.20 \( t(126) = 2.15, p = .03, 95\% CI \) of the difference [.027 to .634], \( d = .19 \).

Our findings in Experiment 7 again demonstrate that the distinction between knowing and knowledge is exhibited even in response to Gettier cases. However, the effects we observed in Experiment 7 were weaker than those observed in Experiment 6, suggesting that this distinction plays a more minor role in people’s knowledge ascription behaviors in Gettier cases.

10. Experiment 8 - Philosopher’s papers

The proceeding seven experiments, which controlled for several alternative explanations, suggest that ordinary people distinguish between knowing and having knowledge in more ordinary contexts (Experiments 1 – 5) and in the context of thought experiments (Experiments 6 and 7). One common response to empirical data on laypeople’s knowledge attributions is that because ordinary people are making judgments about topics that they are unfamiliar with, they will tend to give responses which are not actually representative of their concepts. Consequently, in experiment 8, we sought to test the prediction that the difference between knowing and having knowledge would affect how philosophers argue for claims about knowledge. Thus, experiment 8 served as a more direct test of the claim that these subtle syntactic differences can influence people’s behavior. As philosophers have traditionally been at forefront of epistemic debates, if they exhibit a sensitivity to the distinction between knowing and having knowledge in their own
writing then that would suggest that our prior results are not simply due to the fact that laypeople are not experts on understanding epistemic sentences. Instead, the differences we observed would more likely owe to the effect of subtle syntactic differences on people’s inferences and behavior.

We tested this prediction by looking at philosophers’ own papers in which they defend one of two epistemic theories. First, we recruited philosophers to participate in our study by posting on a widely read philosophy blog (Leiter Reports: A Philosophy Blog). The post read that we were seeking a list of the best papers defending fallibilism about knowledge (roughly, the view that having knowledge does not require certainty) and the best papers defending infallibilism about knowledge (roughly, the view that having knowledge requires certainty). We predicted that papers in which philosophers argued for fallibilism or infallibilism would exhibit different proportions of use of the verb form or the noun form of knowledge when making their arguments. This result would suggest that philosophers are implicitly exhibiting a sensitivity to the distinction between knowing and having knowledge in their own writing. We computed this proportion by simply counting the number of verb forms of knowledge (i.e., know, knows, knew, known, knowing) the author used in their paper and divided that by the number of noun forms they used in their paper (i.e., knowledge). This provided a verb form of knowledge to noun form of knowledge ratio for each paper (henceforth, verb-to-noun ratio).

As fallibilist argue that knowledge does not require certainty, and our results suggest that people think less is required to know than to have knowledge, we predicted that fallibilists’ verb-to-noun ratio would be larger than infallibilists’ verb-to-noun ratio.

Participants. Participants in our study were 18 professional philosophers. 72% of these philosophers held tenure track jobs at major research universities.
**Procedures.** Participants were asked to list the top five papers defending fallibilism about knowledge and the top five papers that defended infallibilism about knowledge (this ordering was counterbalanced). The listing task yielded 32 distinct papers, as many philosophers recommended the same papers. Participants that simply listed philosophers (e.g., Kant), but not specific papers, were excluded before performing additional analyses.

**Results**

After calculating each paper's verb-to-noun ratio, we compared fallibilists’ median ratio (Mean = 2.90, Median = 1.51) to infallibilists’ median ratio (Mean = .71, Median = .63) using a Mann-Whitney-U test. Consistent with our prediction, this test revealed that fallibilists’ ratio was significantly larger than infallibilists’ ratio ($U = 52, p = .008, CL = 78.4$). We followed up on this difference with a Spearman’s rho correlation that showed that the author’s philosophical position (fallibilism or infallibilism) was correlated with their verb-to-noun ratio $r_s = -.48, p < .01$.

These findings are quite striking, as philosophers arguing for either fallibilism or infallibilism were presumably unaware that they used a greater proportion of the verb form of knowledge than the noun form (and vice versa) in advancing their arguments. The results show that the effects we observed among ordinary people can likewise be found in philosopher’s own writing, suggesting that this effect has little to do with a lack of expertise among ordinary speakers but instead is more likely explained by the effect subtle syntactic differences can have on people’s epistemic inferences and behavior. Further, these results lend support to the hypothesis that knowing and knowledge pick out genuinely distinct epistemic concepts.

**11. General Discussion**

The know-based definition of knowledge has not been sufficiently vetted. In 8
experiments, we raised the concern that the know-based definition of knowledge is incorrect. In our experiments, both laypeople and philosophers treated knowing and knowledge as though they have different evidential requirements. We observed these findings using explicit questions in within and between-subjects designs, as well as looking at the word choice of philosophers’ papers defending competing epistemic theories.

Recall, the know-based definition of knowledge defines knowledge in terms of what an agent knows. However, because knowledge appears to have a distinct epistemic nature from knowing (e.g., a higher degree of certainty is required, it entails more reliability and so forth), this definition seems to misrepresent knowledge. Below we discuss how our findings may offer some explanation of several heretofore-unresolved puzzles in epistemology.

12. The shift from infallibilism to fallibilism

The view that knowledge requires certainty is often associated with Descartes, who in the Meditations, argues that all doubt needs to be removed from what we believe if we are to provide a solid foundation for science and knowledge writ large (Descartes, 2013; 1641). Subsequently, as philosophy has turned towards the study of epistemic language (e.g., MacFarlane, 2005; Stanley, 2003) this view has largely gone out of fashion. Indeed, much of the motivation for this shift was that there are several well-known linguistic counter-examples to the thesis that knowledge requires certainty. That is, there are utterances where it seems perfectly apt to say that one knows even if one is uncertain. For example, it seems perfectly apt for me to say that “I know my car was not stolen from the parking lot” even though I cannot be certain that it wasn’t stolen.

We think it is possible that the fallibilist shift in thinking about the nature of knowledge may be explained, at least in part, by our findings about the distinction between knowing and
knowledge. Our findings suggest that “knowledge” is viewed as being more reliable, more certain, and consequently, more difficult to attain than knowing. As Descartes famously attempted to provide a foundation for *scientia*, meaning the object *knowledge*, it seems plausible that the shift in popularity of infallibilism to fallibilism in the history of epistemology was not simply due to counter-examples about the nature of knowledge, but instead about a shift in philosophers’ focus from *knowledge*, the category, to *knowing* the action.

Of course, we do not assume our hypothesis about this shift in thinking in the history of epistemology entirely explains why most epistemologists do not think certainty is required for knowledge. The rise of naturalized epistemology, which builds epistemic theories on the basis of psychological research, likely plays an important role in the changing tides. As many naturalized epistemologists know, psychological research has long shown that we do not have unmediated access to the world. Thus, if certainty were required for knowledge, it would appear very difficult for us to have any knowledge at all. Since that claim seems false, many conclude that knowledge must not require certainty.

Nonetheless, we conjecture that the shift in focus from the object *knowledge*, in Descartes writings, to the action of *knowing* may offer some explanation for the paradigm shift in the history of epistemology. We conclude that more psychological and philosophical research is necessary to determine how plausible this conjecture is. However, we think that the data we have provided here may offer some insight into both first order questions about the nature of knowledge, and second order questions about shifts in the thinking in the history of epistemology.

13. Additional Theoretical Implications

We think our data also speak to several issues in traditional and experimental philosophy.
that have been hotly debated in the last decade. First, we note that philosophy has undergone a revolution in the last decade in which experimental philosophers have attempted to impugn the traditional philosophical practice of relying on “intuitions” as the primary arbitrators of truth (e.g., Alexander, Mallon, & Weinberg, 2010). It is clear that our data offers some support for the concern that intuitions cannot always serve this function in debates about epistemology, or philosophy in general. We found that subtle syntactic differences, differences that went completely unnoticed in the history of epistemology, can elicit different epistemic intuitions. We assume that these differences are not, on the face of it at least, relevant to “the facts of the matter” about the nature of knowledge. Indeed, some may worry that our data shows that the debate between fallibilists and infallibilists bottoms out in a linguistic distinction that makes it appear that these camps of thought have simply been talking past each other – one has focused on the nature of knowledge, while the other has focused on the nature of knowing. Further research is necessary to draw firmer conclusions about the role these syntactic differences play in epistemology, and philosophy at large.

Second, although further research is necessary, our data may also have implications for analogous debates in other areas of philosophy. By analogous debates we mean that there are debates about the threshold required for, say, something to be an explanation or say, an agent having the intention to act. In light of our findings, we would not be surprised to find that some of these ongoing debates may bottom out in different proportions of the word explanation to explains, or a difference in the an ethicists focus on acting intentionally vs having the intention to act. Given that children are also sensitive to syntactic cues as early as 4 years old (Gelman & Heyman, 1999), we think it’s quite likely that the syntactic effect is not about knowledge per se. Rather, we think it is more likely that this finding represents a single example of a far-reaching
bias that has simply gone unnoticed in philosophical theorizing.\textsuperscript{9}

\textsuperscript{9} This research was conducted with Professor Andrei Cimpian.
CHAPTER 5
HOW LARGE IS THE ROLE OF EMOTION IN JUDGMENTS OF MORAL DILEMMAS?

1. Introduction

How do we decide whether an act is morally right or wrong? Though this question has a long history, recent insights into the psychological processes involved in moral judgment have reinvigorated normative ethical debates about our moral obligations to ourselves and others (e.g. Greene, 2007; 2013; Prinz, 2007; Singer, 2005). Despite heightened interest among researchers, the nature of the controversies surrounding moral decision-making has not fundamentally changed. Historically, there has been debate between philosophers who stressed the role of reason and deliberation in moral judgment (e.g., Kant, 1797/2002), and those who argued that moral judgments are driven by emotional processes (e.g., Hume, 1739/2012; 1748/2011). These contrasting emphases are also evident in the course of psychological research on moral judgment. Early investigations were chiefly concerned with how morality was shaped through cognitive development (e.g., Kohlberg, 1976). More recently however, a great deal of research has focused on the role of emotion in moral decision-making (e.g., Eskine, Kacinik, Prinz, 2010; Greene, Nystrom, Engell, Darley & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001; Inbar, Pizarro, Knobe, & Bloom, 2009; Prinz, 2007; Rottman & Kelemen, 2013; Rottman, Kelemen & Young, 2014; Seidel & Prinz, 2013).

The roles of reason and emotion are integrated in the dual-process theory of moral judgment (Greene et al., 2001; Greene, 2014). According to the dual-process theory, “cold” reasoning processes are recruited when making utilitarian moral judgments, but these judgments can be preempted by “hot” affective processes that lead people to make deontological moral
judgments. The signatures of these two processes are thought to be evident in the so-called personal-impersonal distinction: researchers have found that people are less likely to approve of sacrificing one person to save others if a dilemma requires an “up-close-and-personal” action, such as physically pushing someone to their death, than if a dilemma requires an action that operates at greater distance, such as flipping a switch that leads to someone’s death. The dual-process theory has proved very influential within the field of moral psychology. It is now widely accepted that people are less likely to approve of personal violations because they evoke strong emotional reactions compared to impersonal actions (e.g., Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Greene et al., 2001; Greene et al., 2004; Koenigs, Young, Adolphs, Tranel, Cushman et al., 2007; Moretto, Làdavas, Mattioli, & Di Pellegrino, 2010; however also see Kahane, Everett, Earp & Farias, 2014).

2. Reexamining the relationship between emotion and judgment in moral dilemmas

While there has been widespread acceptance of the dual-process theory of moral judgment, there is a troubling lack of research directly linking emotion and people’s judgments about moral dilemmas using traditional self-report measures. Proponents of the dual-process theory have discussed the role of “emotional processes,” but their methods often do not directly measure which emotions, if any, are involved in judgments of moral dilemmas (e.g., Costa et al., 2014; Cummins & Cummins, 2012; Greene et al., 2008; Paxton, Ungar, & Greene, 2012; Suter & Hertwig, 2011). To be sure, there is compelling evidence that emotions are linked to judgments about norm violations such as committing incest or suicide (Eskine et al., 2010; Haidt, 2001; Inbar et al., 2009; Prinz, 2007; Rottman & Kelemen, 2013; Rottman, Kelemen & Young, 2014; Seidel & Prinz, 2013). Yet, it is unclear whether emotion plays a similar role in judgments about moral dilemmas. Moral dilemmas pit competing moral norms against one another (e.g.,
save lives vs. don’t kill), and each of these norms may elicit an emotional response (see Prinz, 2007). Therefore, people’s emotional responses to moral dilemmas may be complex (e.g., a person might at once feel positively about saving lives and negatively about the possibility of failing to save lives and about killing). Consequently, these emotional reactions may not straightforwardly inform people’s judgments about dilemmas as they might for judgments of a single norm violation. Consistent with this, a number of researchers have shown how induced emotions affect judgments of simple norm violations (e.g., Eskine et al., 2010; Seidel & Prinz, 2013), but attempts to demonstrate these effects on judgments about moral dilemmas have produced inconsistent results (e.g., Johnson, Cheung, & Donnellan, 2014; Schnall, Benton, & Harvey, 2008; Strohminger, Lewis, & Meyer, & 2011; Valdesolo & Desteno, 2006).

The most widely cited evidence for the role of emotion in judgments of moral dilemmas, and for the dual-process theory, has come from research examining people’s judgments about a particular battery of moral dilemmas (henceforth, the standard battery; e.g., Ciaramelli et al., 2007; Greene et al., 2001; Greene et al., 2004; Koenigs et al., 2007; Moretto et al., 2010). Most prominently, a number of neuroimaging studies have examined people’s judgments about dilemmas taken from the standard battery. For instance, in two studies, Greene et al. (2001; 2004) found increased activation in brain areas associated with emotion when participants made judgments about personal dilemmas, and increased activation in areas associated with reasoning processes when they considered impersonal dilemmas. In other studies, researchers showed similar effects using psychophysiological measures of affect (Moretto et al., 2010), and when examining clinical populations with ventromedial prefrontal cortex lesions (an area of the brain thought to be critically involved in emotion and emotion regulation; e.g., Ciaramelli et al., 2007; Koenigs et al., 2007).
However, there appear to be serious problems with the standard battery (Cushman, Young, & Greene, 2010; Greene et al., 2008; Kahane & Shackel, 2010). For instance, researchers have argued that some items in the standard battery do not actually constitute moral dilemmas, that impersonal dilemmas often involve abstract or probabilistic reasoning but personal dilemmas do not, and that the dilemmas are not matched in certain basic respects such as perspective (second- or third-person) and word length (e.g., McGuire, Langdon, Coltheart, & Mackenzie, 2009; Moore, Clark, & Kane, 2008; Nakamura, 2013).

In addition to these concerns, we observed that personal dilemmas in the standard battery more often involve physically harming a moral patient than do impersonal dilemmas. In fact, it appears that all of the personal dilemmas in the standard battery involve physical harm, whereas only half of the impersonal moral dilemmas do. This presents a potentially serious confound, as the harmfulness of an action ought to be orthogonal to its up-close-and-personal nature. What’s more, it appears that personal dilemmas in the standard battery tend to involve more graphic and grisly descriptions of harm than do impersonal dilemmas, even when focusing only on impersonal dilemmas involving harm. For example, personal dilemmas ask participants to consider cutting off a man’s head, smothering a baby, or subjecting children to painful medical experiments. In contrast, impersonal dilemmas ask participants to consider venting invisible but deadly fumes into a room, or voting for a new environmental policy that will harm people.

Researchers using the standard battery have often argued that the “closeness” of personal moral actions elicits a strong negative emotional reaction that in turn leads participants to make deontological moral judgments. However, one obvious possibility is that researchers have observed stronger emotional reactions to personal dilemmas because the personal dilemmas in the standard battery more often involved grisly and harmful actions than did the impersonal
dilemmas, and not because of the closeness of personal actions. If so, then prior studies have made the unremarkable observation that graphic descriptions of harmful acts are emotionally salient—an altogether trivial finding. Thus existing studies do not provide a meaningful test of the central claim of the dual-process theory, as it remains unclear whether emotional responses explain the difference in people’s judgments about personal and impersonal dilemmas.

Although a number of researchers have raised concerns about the standard battery, no research to date has determined whether emotional responses explain people’s moral judgments when the confounds in the standard battery are eliminated. Confirming this connection would lend support to the dual-process theory. However, if this claim is not borne out, it would suggest that people’s judgments about personal and impersonal moral dilemmas are driven by factors other than emotion.

The notion that other factors might play a role in judgments of is consistent with recent findings that suggest moral decision-making is a complex cognitive process that takes many different inputs (Gong, Iliev, & Sachdeva, 2012; Iliev & Medin, 2012; Miller & McFarland, 1986; Rai & Fiske, 2011; Rai & Holyoak, 2012; Uhlmann, Zhu, & Tannenbaum, 2013; Waldmann & Dieterich, 2007). Similarly, it is consistent with a number of domain-general explanations of moral decision-making (e.g., Bennis, Bartels, Medin, 2010), including dual-process views that do not make specific commitments about emotional processes (e.g., Tremoliere & Bonnefon, 2014).

3. The Current Experiments

We have argued that existing research has failed to establish the connection between specific emotions and people’s judgments about moral dilemmas. We sought to establish this
connection in two ways: First, we used a self-report measure to compare participants’ emotions before and after considering a moral dilemma. This allowed us to establish the connection between specific emotions and moral judgments, and to demonstrate the strength of these relationships. Second, we tested a battery of matched personal and impersonal dilemmas – drawn from an existing battery created by Moore et al., (2008) – to assess whether emotions play a substantial role in moral decision-making.

4. General Methods for the Current Experiments

   **Participants.** Participants were recruited online from Amazon’s Mechanical Turk work distribution website (mTurk). After recruitment, participants were redirected to a Qualtrics website where the experiment was administered. Participants were paid $0.60 to participate in Experiments 1 and 3, and $0.75 to participate in experiment 2.

   **Materials: The Standard Battery.** The standard battery is composed of 44 moral vignettes describing situations in which a moral decision must be made. Each vignette describes an action which must be taken to avoid an undesirable outcome, but which comes at the cost of another morally undesirable outcome. These vignettes are divided into two groups: personal moral dilemmas (25 dilemmas) and impersonal moral dilemmas (19 dilemmas). Personal moral dilemmas involve more intimate and direct moral actions than impersonal dilemmas. Within and between these conditions, the vignettes range over a wide variety of situations and actions, from donating to a charitable organization, to pushing a man in front of a train.

   **Materials: Revised Battery.** We compiled a new set of moral dilemma vignettes to address the concerns we raised about the standard battery. Most of these scenarios were taken directly from Moore et al. (2008), and others were developed by modifying their materials to
further improve the match between personal and impersonal versions of the dilemma. We used eight different scenarios presenting a dilemma in which a moral patient must be harmed in order to maximize utility. For each scenario, we created a pair of personal and impersonal vignettes that were matched as closely as possible, save for the intimacy and directness of the action considered in the vignette. For instance, in the “Space Station” scenario, a fire threatens to break out in the international space station unless a module is vented of oxygen. Unfortunately, an astronaut is trying to exit the module, and his presence in the doorway will prevent the fire safety system from activating. In the personal version of this scenario, the astronaut is stuck in the doorway, and participants must consider whether or not to push him back into the module so that the fire system will be activated. This action will kill him but will save the others on the station. In the impersonal version of the scenario, participants must consider whether to press a switch that will seal the doorway before he reaches it, with the same consequences as the personal scenario (Moore et al., 2008).

**Emotion Measure: Positive and Negative Affect Schedule - Expanded Form.** In Experiments 1 and 3 we measured participants’ emotional responses to moral dilemmas using scales from the Positive and Negative Affect Schedule expanded form (PANAS-X), a comprehensive emotional state, trait, and mood self-report measure (Watson & Clark, 1994; Watson, Clark, & Tellegen, 1988). The PANAS-X is among the most commonly used self-report measures of emotion (according to Google Scholar, Wallace et al.’s 1988 paper has been cited over 15,000 times at the time of this writing). A number of papers in moral psychology have also used this measure to examine the link between emotion and moral judgments or moral behavior (e.g., Gray, Ward, Norton, 2014; Paxton, Unger, Greene, 2012). These measures ask participants to rate the extent to which they are experiencing different emotions, on a scale from 1 (very
slightly or not at all) to 5 (extremely). The measure is composed of subscales, each of which is composed of several emotion words. We presented participants with the positive and negative affect scales, as well as the guilt, hostility, and joviality scales. Table 3 provides more detail about the variety of emotion words that participants rated.

Table 3. Items included in each of the PANAS-X scales.

<table>
<thead>
<tr>
<th>Negative Affect (10)</th>
<th>afraid, scared, nervous, jittery, irritable, hostile, guilty, ashamed, upset, distressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Affect (10)</td>
<td>active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong</td>
</tr>
<tr>
<td>Hostility (6)</td>
<td>angry, hostile, irritable, scornful, disgusted, loathing</td>
</tr>
<tr>
<td>Guilt (6)</td>
<td>guilty, ashamed, blameworthy, angry at self, disgusted with self, dissatisfied with self</td>
</tr>
<tr>
<td>Joviality (8)</td>
<td>happy, joyful, delighted, cheerful, excited, enthusiastic, lively, energetic</td>
</tr>
</tbody>
</table>

*Note.* The number of terms comprising each scale is shown in parentheses.

We chose the hostility scale to measure anger and disgust, which many researchers have identified as potentially key emotions involved in moral judgment (e.g., Eskine, Kacinik, Prinz, 2010; Haidt, 2001; Inbar, Pizarro, Knobe, Bloom, 2009; Prinz, 2007; Rottman & Kelemen, 2013; Rottman, Kelemen & Young, 2014; Seidel & Prinz, 2013). The guilt scale was chosen because the experience of guilt may serve adaptive functions in deterring moral norm violations and in regulating relationships affected by norm violations (e.g., Baumeister, Stillwell, & Heatherton, 1994; Hareli & Eiskovits, 2006; Giner-Sorolla et al., 2008; Zeelenberg & Breugelmans, 2008). Finally, the positive affect and joviality scales were used along with the negative affect scale as general measures of affect, respectively.
Filler Task. We used a filler task to reduce any memory effects caused by repeated administrations of the PANAS-X. Participants were presented with three images taken from the *Where’s Waldo* book series. Participants were asked to search for Waldo, and to click on him when they found him.

Catch Questions. All experiments included catch questions to identify participants who were not paying attention, or were clicking through the study. These questions were embedded in the response scales, and instructed participants to choose a specific response option. For instance, a catch question embedded in the emotion measure was, “For this item please respond ‘a little’.”

5. Experiment 1

In Experiment 1 we sought to validate our measure by reproducing Greene and colleagues (2001) original effect.

Participants. We recruited 266 participants to participate in Experiment 1. Of these, 141 participants were female and 125 were male, with mean age of 32.8 years old (SD = 11.51).

Materials. In Experiment 1, each participant read one of the 44 moral dilemma vignettes from the standard battery, reproduced verbatim from Greene et al. (2001). Each item was followed by an additional sentence directing participants’ attention to the moral action they should consider before rating their emotions. For instance, for the “Standard trolley” dilemma this sentence read, “You are thinking about flipping the switch in order to save the five workmen.”

Procedure. After collecting demographic information, participants were directed to complete an emotion pre-test. Participants were asked to rate how they were feeling at the present moment using the selected subscales from the PANAS-X. This established a baseline for
each participant’s emotional state upon entering the study. A catch item was included within the emotion scale. After completing the pre-test, participants completed the filler task.

Participants were then randomly assigned to read a personal or impersonal moral dilemma. Participants assigned to the personal condition read one of the 25 personal moral dilemmas, and participants in the impersonal condition read one of the 19 impersonal moral dilemmas (in a between-subjects design). After reading a moral dilemma, participants completed the emotion post-test, rating their emotions using the PANAS-X scales a second time. Participants were given specific instructions to rate their emotions as they were currently experiencing them, and not to respond based on how they imagine they might feel if they were actually in the situation described. After rating their emotions, participants were asked to make a moral decision. They responded using a six-point labeled scale that ranged from “completely inappropriate” to “completely appropriate.”

Results and Discussion

Five participants were excluded for missing at least one catch question, leaving 261 participants in the final analyses.

First, we examined participants’ moral judgments, and replicated prior findings. Participants made more deontological moral judgments for personal dilemmas (Mean = 2.82, SD = 1.915) than for impersonal dilemmas (mean = 3.47, SD = 2.00), \( t(259) = -2.67, p = .009, 95\% CI [.165 to 1.121], d = .33 \). Next, we examined the effect of reading moral dilemmas on participants’ emotional states by computing an emotional reaction score. We calculated an emotion reaction score for each subscale by subtracting participants’ pre-test emotion ratings from their post-test emotion ratings on each scale. Mean emotional reaction scores for each
condition are shown in Figure 3. Reading both personal and impersonal moral dilemmas led to increased negative emotions (negative affect, guilt, hostility), and decreased positive emotions (positive affect, joviality), when compared to pre-test emotional states.

Figure 3. Mean emotion difference scores for PANAS-X subscales across personal and impersonal conditions in Experiment 1. Error bars represent ±1 standard error.

We conducted a series of 2 x 2 ANOVAs, one for each emotion subscale (summarized in Table 4). We examined two factors with these ANOVAs: a within-subjects Emotional Reaction factor comparing pre-test scores and post-test scores, and a between-subjects Condition factor comparing personal and impersonal dilemmas. Within each ANOVA, the main effect of the Emotional Reaction factor tested whether reading a moral dilemma affected participants’ emotional state, and the interaction between Emotional Reaction and Condition factors tested whether the change in participants’ emotional state was greater for personal dilemmas than for
impersonal dilemmas. We observed a significant Emotional Reaction effect for every subscale, indicating that both personal and impersonal dilemmas elicited emotional reactions. We also observed significant interactions between the Emotional Reaction factor and Condition for the negative affect, guilt, hostility, and joviality subscales.

Table 4. Summary of ANOVAs conducted on emotion ratings of participants in Experiment 1. The Emotional Reaction factor has been abbreviated as “Reaction.”

<table>
<thead>
<tr>
<th>Effect</th>
<th>Positive Affect</th>
<th>Negative Affect</th>
<th>Hostility</th>
<th>Guilt</th>
<th>Joviality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df  F   p   η²</td>
<td>F   p   η²</td>
<td>F   p   η²</td>
<td>F   p   η²</td>
<td>F   p   η²</td>
</tr>
<tr>
<td>Reaction</td>
<td>1    79.9 &lt;.001 .058</td>
<td>125 &lt;.001 .134</td>
<td>114 &lt;.001 .109</td>
<td>71.5 &lt;.001 .088</td>
<td>148 &lt;.001 .113</td>
</tr>
<tr>
<td>Condition</td>
<td>1    8.63 .004 .024</td>
<td>8.67 .004 .019</td>
<td>18.3 &lt;.001 .040</td>
<td>2.66 .104 .006</td>
<td>7.38 .007 .019</td>
</tr>
<tr>
<td>Interaction</td>
<td>1    3.04 .082 .002</td>
<td>14.5 &lt;.001 .015</td>
<td>32.6 &lt;.001 .031</td>
<td>5.50 .02 .007</td>
<td>7.73 .006 .006</td>
</tr>
</tbody>
</table>

We found that considering a personal dilemma led to significantly greater emotional changes than considering an impersonal dilemma, reproducing the widely reported personal-impersonal effect using a self-report emotion measure. However, we also observed that reading a moral dilemma led to increased negative emotions and decreased positive emotions across both personal and impersonal dilemmas, and this effect accounted for the greatest proportion of variance among the effects we examined.

6. Experiment 2

How should we interpret the results of Experiment 1 and the findings of other researchers who have examined participants’ responses to the standard battery? Many have argued that people’s stronger emotional responses to personal dilemmas explain the differences in their judgments about personal and impersonal moral dilemmas. However, this interpretation may be
unwarranted if these emotional differences can be explained by uncontrolled factors in the standard battery.

The most notable difference we observed between personal and impersonal moral dilemmas is the amount of physical harm and graphicness of the descriptions in personal dilemmas. To determine whether this issue was as problematic as we suspected, we performed a norming study on the standard battery and on our battery of revised dilemmas.

Participants. We collected responses from 256 participants for this experiment. Of these, 100 were female and 156 were male. Their mean age was 31.90 years old ($SD = 10.90$).

Materials and Procedure. After responding to demographic questions, participants were randomly assigned to read four personal and four impersonal vignettes from either the standard battery or from our revised battery of moral dilemmas.

Participants were instructed to first read each vignette and then to rate their agreement with each of the statements that composed the Harm ($\alpha = .80$) and Graphicness ($\alpha = .81$) scales (both scales can be found in the Appendix). A catch item was included in a random place in both scales. After participants made these ratings, the vignette remained onscreen while they were asked to make a moral judgment about the vignette. Vignettes were presented in a random order and the ordering of the harm and graphicness scales and their items were also randomized.

For both the harm and graphicness scales, participants were asked to rate their agreement with five statements, two of which were reverse coded. The harm scale included statements such as, “The situation is violent.” The graphicness scale included statements such as, “The language used to describe the situation evokes disturbing images.”
Results and Discussion

Participants who missed any catch questions were excluded from analyses, leaving 215 participants. Participants’ ratings were averaged to compute harm and graphicness scores for each of the 44 vignettes in the standard battery (mean number of ratings = 29.4) and for the 16 vignettes in our revised battery (mean number of ratings = 26.5).

The results of Experiment 2 are shown in Figure 4. First, we conducted a pair of one-way ANOVAs to determine if the personal dilemmas in the standard battery were viewed as more harmful and more graphic than the impersonal dilemmas. Confirming our predictions, we found that personal dilemmas were rated significantly higher on the harm scale than impersonal dilemmas, \( F(1,43) = 25.80, p < .001, \eta^2 = .38 \). Personal dilemmas also had higher graphicness ratings than impersonal dilemmas, \( F(1,43) = 33.66, p < .001, \eta^2 = .45 \). In contrast, no significant differences were observed between harm ratings for personal and impersonal dilemmas in our revised battery, \( F(1,15) = .307, p = .588, \eta^2 = .021 \). Personal dilemmas tended to be rated as more graphic than impersonal dilemmas, but this trend was not statistically significant, \( F(1,15) = 3.98, p = .066, \eta^2 = .22 \).

We also compared the normed ratings for our items with ratings of items from the materials used in prior studies using a 2 x 2 (condition x experiment) ANOVA for each scale (see Figure 4). The harm ratings of both our personal and impersonal items were comparable to the personal items in the standard battery. We observed significant differences between conditions \( (F(1,59) = 6.97, p = .011, \eta^2 = .111) \) and between experiments \( (F(1,59) = 8.35, p = .005, \eta^2 = .130) \). Most importantly, we observed a significant interaction \( (F(1,59) = 10.19, p = .002, \eta^2 = .154) \), indicating that the differences between personal and impersonal dilemmas were
stronger among the moral dilemmas from the standard battery than among those in our revised battery.

A comparison of graphicness ratings revealed a significant main effect of experiment \(F(1,59) = 93.00, \ p < .001, \ \eta^2 = .624\), indicating that our items were more graphic than those in the standard battery. An ANOVA also revealed a significant effect of condition \(F(1,59) = 14.38, \ p < .001, \ \eta^2 = .204\), and a significant interaction \(F(1,59) = 6.48, \ p = .014, \ \eta^2 = .10\). The significant interactions in the ANOVAs conducted on harm and graphicness ratings demonstrate that the differences between personal and impersonal items from the standard battery are not driven by inherent differences in the interpretation of personal and impersonal moral dilemmas, but rather by confounds in their items.

**Figure 4.** Mean graphicness and harm ratings for items from the Standard battery and our new revised battery of moral dilemmas. Error bars represent ±1 standard error.

**Correlational Analyses.** As we hypothesized, personal and impersonal dilemmas in the standard battery were not matched for two potentially confounding factors—how harmful the
actions were and how graphically the actions were described. In fact, we suspect that these confounding factors are at least partially responsible for participants’ differing emotional reactions to personal and impersonal dilemmas. Supporting this, we found small but significant correlations between a dilemma’s harm score and the emotion difference scores observed for participants in Experiment 1 (negative affect ($r = .155, p = .011, 95\% CI [.04 to .27]$), hostility ($r = .143, p = .020, 95\% CI [.02 to .26]$), and guilt ($r = .124, p = .043, 95\% CI [.003 to .24]$). We also observed stronger correlations between an item’s graphicness score and participants’ emotional difference scores on negative affect ($r = .219, 95\% CI [.10 to .33], p < .001$), hostility ($r = .272, 95\% CI [.38 to .16], p < .001$), guilt ($r = .170, 95\% CI [.05 to .29], p = .005$), and joviality ($r = -.165, 95\% CI [-.28 to -.05], p = .007$). These confounds may call into question the conclusions prior studies have drawn about the role of emotion in people’s judgments about moral dilemmas.

7. Experiment 3

**Participants.** *A priori* power analyses were conducted to determine desired sample sizes for ANOVA and correlational analyses. We wanted to ensure adequate power to detect effects similar to those observed in Experiment 1. Among the ANOVAs conducted in Experiment 1, the smallest significant effect was observed in participants’ emotion ratings for the guilt subscale. Power analysis conducted using G*power (Faul & Erdfelder, 1998) revealed that a total of 284 participants would be required to achieve 99% power to detect an effect of this size. Meanwhile, hostility was the only emotion subscale to correlate significantly with participants’ moral judgments in Experiment 1 ($r = -.178$). To detect similar correlations with 99% power, 568 participants would be required.
In Experiment 3, we recruited 654 participants anticipating that we might need to remove some participants from our analyses for missing catch questions. Of these participants, 359 were female and 295 were male. Their mean age was 36.02 years old ($SD = 13.01$).

**Materials and Procedure.** The norming study confirmed that the revised battery does not suffer from significant extraneous differences in the degree of harm nor in the graphicness of our vignettes. By examining participants’ reactions to these new vignettes, we are able to test whether personal dilemmas are actually more emotional than impersonal dilemmas, and the extent to which a difference in emotional responses predicts participants’ judgments about moral dilemmas.

This experiment used our revised battery of eight moral scenarios, for a total of 16 vignettes (personal vs impersonal x 8 scenarios). All other materials and procedures were identical to those used in Experiment 1. After completing these procedures, some participants were also asked to answer an additional set of questions: the Trait Meta-Mood Measure (Salovey, Mayer, Goldman, Turvey, 1995). We had hoped that this measure might identify those participants for whom emotion and moral judgments would be most related. However, we did not find any stronger connection between emotion and moral judgment for participants who scored higher on this measure than those who scored lower. Therefore, we have omitted discussion of these analyses.

**Results and Discussion**

Thirty-eight participants were excluded for missing at least one catch question, or for indicating that they had not paid attention when participating, leaving 616 participants in the final analysis.
First, we examined participants’ emotional reactions to our revised battery of moral dilemmas. We again calculated an emotional reaction score for participants by subtracting their pre-test emotion ratings from their post-test emotion ratings. Mean differences for participants in each condition are shown in Figure 5. Both personal and impersonal moral dilemmas produced increased negative emotions (negative affect, hostility, and guilt) and decreased positive emotions (positive affect and joviality).

![Figure 5](image)

*Figure 5. Mean emotion difference scores for PANAS-X subscales across personal and impersonal conditions in Experiment 3. Error bars represent ±1 standard error.*

As in Experiment 1, we performed a series of 2 x 2 (Emotional Reaction x Condition) ANOVAs for each emotion subscale (summarized in Table 5). Comparing participants’ emotions at pre-test and post-test, we observed a significant change on every emotion subscale. Then, we tested whether personal dilemmas elicited stronger negative emotional reactions than impersonal
dilemmas after matching the dilemmas on graphicness and harm dimensions. To test this, we examined the interaction between the Emotional Reaction and Condition (personal vs. impersonal) variables. We observed significant interactions between the degree of change in participants’ emotional state and their assigned condition for all subscales except the positive affect and joviality scales, indicating that personal dilemmas had a significantly larger effect on participants’ emotional state.

Table 5. Summary of ANOVAs conducted on emotion rating data from Experiment 3.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Positive Affect</th>
<th>Negative Affect</th>
<th>Hostility</th>
<th>Guilt</th>
<th>Joviality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>F</td>
<td>p</td>
<td>η²</td>
<td>F</td>
</tr>
<tr>
<td>Reaction</td>
<td>1</td>
<td>274</td>
<td>&lt;.001</td>
<td>.064</td>
<td>422</td>
</tr>
<tr>
<td>Condition</td>
<td>1</td>
<td>1.72</td>
<td>.220</td>
<td>.002</td>
<td>2.40</td>
</tr>
<tr>
<td>Interaction</td>
<td>1</td>
<td>2.49</td>
<td>.115</td>
<td>.006</td>
<td>6.89</td>
</tr>
<tr>
<td>Error</td>
<td>614</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The Emotional Reaction factor has been abbreviated as “Reaction” and interaction terms are labeled using the first letter of each crossed factor.

Although we were able to observe significant differences between the emotions elicited by personal and impersonal dilemmas in these experiments, these effects were all vanishingly small relative to the size of the emotional reaction effect for each subscale and to the size of the effects we observed in Experiment 1. In Experiment 1, the strongest interaction effect was observed for the hostility subscale, which accounted for approximately 29% as much variance as the main effect of hostility, or a ratio of approximately 3:1. In experiment 3, the variance
accounted for by the hostility interaction term was less than 3% of that accounted for by the main effect of hostility, or a ratio of approximately 35:1. Comparing participants’ hostility reaction scores across Experiments 1 and 3 using a 2 x 2 ANOVA (Condition x Experiment), we found a significant interaction between these factors ($F(1, 873) = 10.609, p = .001, \eta^2 = .011$), indicating that differences in the amount of anger and disgust elicited by personal versus impersonal dilemmas were greater in Experiment 1 than 3. Thus, when examining matched personal and impersonal dilemmas, a participant’s emotional state depended more on whether they had read a moral dilemma than on whether that dilemma was personal or impersonal in nature.

When matched in terms of graphicness and harm, personal and impersonal dilemmas elicited very similar emotional reactions. How would this affect participant’s moral judgments about these dilemmas? We found that participants’ moral judgments were still significantly lower for personal dilemmas (mean = 2.98, $SD = 1.755$) than impersonal dilemmas (mean = 3.72, $SD = 1.865$), $t(614) = 5.294, p < .001$. Comparing participants’ moral judgments across Experiments 1 and 3 in a 2 x 2 ANOVA (Condition x Experiment) revealed no significant main effect of Experiment ($F(1, 873) = 2.467, p = .117, \eta^2 = .002$) nor any interaction ($F(1, 873) = .221, p = .639, \eta^2 = .0002$). Eliminating confounds in the standard battery did not significantly affect the pattern of moral judgments we observed among personal and impersonal dilemmas.

Altogether, it appears confounds in the standard battery were responsible for the differences in emotions elicited by personal and impersonal dilemmas, but were not responsible for the differences in participants’ moral judgments.

**Correlational Analyses.** We examined the relationship between each participant’s emotional states and their moral judgments. First, we correlated participants’ difference scores for each emotion scale and their moral judgments (see Table 6), finding significant correlations
between participants’ moral judgments and the change in their positive affect and hostility scores. However, the size of these correlations would conventionally be considered quite small, suggesting that the dual-process theory has placed too great an emphasis on the role of emotion in judgments of moral dilemmas.

Table 6. Correlations between emotion scales and moral judgments after collapsing across conditions in Experiment 3 (above the diagonal). For correlations between moral judgments and the emotion scales, 95% confidence intervals are enclosed in brackets (below the diagonal).

<table>
<thead>
<tr>
<th></th>
<th>Negative Affect</th>
<th>Positive Affect</th>
<th>Hostility</th>
<th>Guilt</th>
<th>Joviality</th>
<th>Gender</th>
<th>Moral Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg. Affect</td>
<td>-</td>
<td>-.265***</td>
<td>.823***</td>
<td>.853***</td>
<td>-.473***</td>
<td>.030</td>
<td>-.033</td>
</tr>
<tr>
<td>Pos. Affect</td>
<td>-</td>
<td></td>
<td>-.246***</td>
<td>-.289***</td>
<td>.771***</td>
<td>-.064</td>
<td>.117**</td>
</tr>
<tr>
<td>Hostility</td>
<td>-</td>
<td></td>
<td>-</td>
<td>.737***</td>
<td>-.399***</td>
<td>.015</td>
<td>-.108**</td>
</tr>
<tr>
<td>Guilt</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-.441***</td>
<td>-.036</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Joviality</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-.066</td>
<td>.023</td>
</tr>
<tr>
<td>Gender</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-.180**</td>
</tr>
<tr>
<td>Moral</td>
<td>[-.11 .05]</td>
<td>[.039 .194]</td>
<td>[-.19 -.03]</td>
<td>[-.08 .08]</td>
<td>[-.06 .10]</td>
<td>[-.25 -.10]</td>
<td>-</td>
</tr>
</tbody>
</table>

Note. *p < .05 **p < .01 ***p < .001.

Of course, these effect sizes would be better compared against the predictions of the theory and against prior findings rather than against conventions. Unfortunately, proponents of the dual-process theory have not made specific predictions about the strength of correlations between emotion and moral judgments, nor about how reliably emotional reactions are expected to elicit deontological moral judgments. Nonetheless, one point of comparison that is available and that could be used to establish the size of a “small” effect is a participant’s gender. We found that participants’ gender was more predictive of their moral judgments ($r = -.18$, $p < .001$), where women were less likely to approve of taking action than men. To our knowledge, there are no
major contemporary theories of moral judgment that argue for the importance of gender in moral judgment. Assuming these theories are not all at once neglecting an important factor in moral judgment, it would seem fair to consider whatever effect gender has on moral judgments to be small. Therefore, the effect of emotion on moral judgment, which was smaller than the effect of gender, must also be small.¹⁰

Still, it is possible these observed correlations underestimate the true size of the correlation between emotion and moral judgment due to measurement error. This attenuation is likely minor: prior research has established that the PANAS-X is highly reliable (Moment instructions hostility α = .82; hostility test-retest reliability r = .65; Watson & Clark, 1994), and although its reliability has not been explicitly tested, the moral judgment scale we used has been used by a number of other researchers (e.g., Horne, Powell & Spino, 2013; Horne, Powell & Hummel; in press; Paxton, Ungar, & Greene, 2010). However, as gender was presumably measured without error, any attenuation may have affected the comparison between emotion and gender correlations. To test whether the observed correlation between hostility and moral judgment was smaller than the correlation between gender and moral judgment due to measurement error, we examined the relationship between these correlations after correction for attenuation (computing ρ; Spearman, 1904). Correction for attenuation attempts to estimate the “true” correlation between two variables by accounting for attenuation due to measurement error.

The corrected correlation is computed as: $\rho_{a,b} = \frac{corr(a,b)}{\sqrt{R_a R_b}}$

Where $R$ is the reliability of each measure.

¹⁰ Note that participants’ gender did not correlate with their emotional responses, indicating that the effect of gender is not mediated through different emotional responding.
Although the reliability of the moral judgment scale is not known, we can compare the ratios of the disattenuated correlations for gender and moral judgment $\rho_{g,m}$ and for emotion and moral judgment $\rho_{e,m}$ (using the lower test-retest reliability coefficient and assuming sex is measured with perfect reliability) to determine their rank order:

\[
\frac{\rho_{g,m}}{\rho_{e,m}} = \frac{\text{corr}(g,m)}{\text{corr}(e,m)} \frac{\sqrt{R_g R_m}}{\sqrt{R_e R_m}} = \frac{-0.180}{-0.108} \frac{\sqrt{R_g R_m}}{\sqrt{R_e R_m}} = \frac{-0.180 \sqrt{R_g R_m}}{-0.108 \sqrt{R_e R_m}} = \frac{-0.180 \sqrt{R_e}}{-0.108 \sqrt{R_g}} = \frac{-0.180 \sqrt{0.65}}{-0.108 (1)} = \frac{0.145}{108} > 1 \therefore \rho_{m,s} > \rho_{m,e}
\]

After some simple algebra, we show that the ratio’s magnitude is greater than one and therefore that gender remains a better predictor of moral judgment than emotion, even taking into account the measurement error of the PANAS-X.

**Mediation analysis.** We observed significant changes in participants’ feelings of hostility between personal and impersonal conditions, and we found that participants’ feelings on this subscale were reliably correlated with their moral judgments. Accordingly, we performed a mediation analysis (Preacher & Hayes, 2006) to test whether differences in moral judgments for personal and impersonal dilemmas can be attributed to increased anger and disgust in response to personal dilemmas. We found a significant indirect effect of condition on moral judgments (10,000 Bootstrapped Samples; Effect: -.0295; 95% CI: lower bound = -.0846, upper bound = -.0017). However, the direct effect of condition remained significant (Effect: -.7150; 95% CI: lower bound = -1.004, upper bound = -.4258) indicating that emotional reactions only partially mediated the effect of condition on moral judgments. Indeed, the relative sizes of the observed effects and confidence interval bounds for the direct and indirect effects clearly demonstrate a stronger direct effect of condition. Therefore, although differences in feelings of anger and disgust did partially account for differences in participants’ moral judgments about personal and
impersonal dilemmas (as the correlation suggests), much of this difference must be attributable to other factors.

8. General Discussion

The present experiments constitute a systematic and direct investigation of the role of emotion in people’s judgments about moral dilemmas. There has never been any question as to whether emotions play some role in moral decision-making—even Kant (1798/2006; 1797/2002) recognized that “sympathies” and “sentiments” influence people’s moral behavior. Rather, the substantive issue concerns the relative contribution of emotion to people’s decisions and whether or not emotion plays an important role in people’s judgments about moral dilemmas. Our findings clearly indicate that emotions play some role in judgments about moral dilemmas: we found a statistically significant difference in the emotions elicited by personal versus impersonal dilemmas as well as statistically significant correlations between emotions and moral judgments. However, these effects were very small.

Thus, our investigation has serious implications for the dual-process theory of moral judgment. This theory predicts that emotional responses to moral dilemmas will account for a non-trivial portion of the variance in people’s moral judgments (e.g., Greene et al., 2001; Greene et al., 2004; Greene et al., 2008). We found that emotion accounted for only a very small proportion of the variance (i.e., 1%) in people’s moral judgments. Participants’ emotional responses were less predictive of their moral judgments than was their gender, a factor that the dual-process theory has not identified as being particularly relevant to moral judgment.

Similarly, once confounds in the standard battery were eliminated, we observed only very small differences between participants’ emotional reactions to personal versus impersonal
dilemmas. In contrast, participant’s moral judgments about these dilemmas continued to differ, demonstrating the dissociation between people’s emotional responses and their moral judgments (also see Cima, Tannaer, & Hauser, 2010). Although the personal-impersonal distinction appears to affect people’ moral judgments, we conclude that the dual-process theory fails to explain this effect.

9. Criticisms and Responses

Do our data represent a failure of the dual-process theory, or a failure of our method? Specifically, one might worry that the PANAS-X is not sufficiently sensitive to detect the emotional states that drive moral judgments, or that the scales we used did not measure the right emotional states. Several points speak against these concerns. First, we not only tested differences between the emotions elicited by personal and impersonal dilemmas, but also tested for differences in participants’ emotions before and after reading a moral dilemma. We observed that reading a moral dilemma had a strong effect on participants’ emotional states, leading to increased negative emotions along with decreased positive emotions. These main effects demonstrate that the emotions we chose to measure are induced during moral decision-making and that the PANAS-X scales are sufficiently sensitive to detect emotions induced by moral dilemmas (consider that the smallest $F$ value we observed among these main effects was 212!). Moreover, we provided further validation of our measure by replicating the original effects found by Greene et al. (2001) using the items from the standard battery.

Second, we attempted to employ analyses and effect size comparisons that were independent of the PANAS-X scales sensitivity. Rather than consider effects of the personal-impersonal distinction in isolation, we compared the differences between emotions elicited by personal and impersonal dilemmas to the overall effect of reading a moral dilemma. Compared
with the variance accounted for by the main effects examining emotions measured before and after reading a dilemma (e.g., hostility $\eta^2 = .113$), the personal-impersonal distinction accounted for only a small fraction of the variance in participants’ emotional states (hostility $\eta^2 = .003$; a ratio of approximately 35:1). Similarly, we also did not rely on conventional effect size when evaluating correlations between emotion and moral judgment. Instead we compared these correlations to the correlation between gender and moral judgment, finding that gender was a more reliable predictor of moral judgment than was emotion—even after adjusting correlations between emotion and moral judgment for attenuation due to measurement error.

Would other measures of emotion such as GSR, FACS, facial electromyography, or fMRI have better captured people’s emotional responses to moral dilemmas? These measures could offer an advantage if the affective states that drive moral judgments are not available to conscious introspection and therefore not measurable by self-report. To our knowledge, no existing research has demonstrated a link between subconscious emotional states and moral judgment, and other researchers have already used the PANAS to successfully demonstrate links between emotions and moral judgments about single norm violations (e.g., Gray, Ward, & Norton, 2014). In addition, leading proponents of the role of emotion in moral judgment are quite explicit in asserting that the emotions experienced during moral judgments are experienced consciously (e.g., Greene & Haidt, 2002), so we see no reason to suspect the relevant emotions cannot be measured using self-report. Although it is possible that future investigations could uncover a stronger connection between emotional responses and people’s judgments about moral dilemmas, at this point these concerns are purely speculative.
10. Theoretical Implications

According to the dual-process theory, people make deontological judgments about personal dilemmas because these dilemmas evoke strong emotions. However, the theory fails to explain the personal-impersonal distinction adequately: our findings suggest that the role of emotion in people’s judgments of moral dilemmas is significantly smaller than predicted by the theory. However, this result is consistent with recent evidence that moral decision-making is a complex cognitive process that is affected by many factors. First, calculations of moral utility depend on a variety of considerations that may or may not be seen as normative. For instance, Goodwin & Landy (2014) found that the utility of a person’s life is judged differently depending on the person’s age, and these judgments appear to operate independently of affect. Moral judgments are also affected by domain-general biases present in many other decision-making contexts (e.g., Gong, Iliev, & Sachdeva, 2012; Rai & Holyoak, 2012). In addition, making judgments about moral dilemmas may require the formation of complex representations of causal relationships, and even subtle changes in these representations may influence how moral actions are evaluated (e.g., Iliev & Medin, 2012; Waldmann & Dieterich, 2007). Finally, people’s moral judgments are influenced by factors that extend beyond the act in question, including consideration of counterfactual alternatives to the outcomes being evaluated (e.g., Miller & McFarland, 1986) and evaluation of what an act indicates about an individual’s character (e.g., Rai & Fiske, 2011; Uhlmann, Zhu, & Tannenbaum, 2013). All of these findings speak to the complexity of moral decision-making, suggesting that it is unlikely that any one factor explains people’s judgments about moral dilemmas.\(^1\)

\(^1\) The proceeding chapter was adapted from a longer paper with Derek Powell.
1. Introduction

Conventional wisdom advises one to steer clear of religion and politics in polite conversation. One reason for avoiding these topics is that they can quickly turn to disagreements about what is moral. Such disagreements are often fruitless: people’s moral beliefs are highly resistant to revision (e.g., Lord, Ross, & Lepper, 1979; also see Skitka, 2010) as they are thought to be rooted in emotion, motivation, and socialization (e.g., Greene, 2007; Haidt, 2001; Prinz, 2007; Rai & Fiske, 2012). Some of these roots go deep: there is evidence that infants have moral beliefs as early as 19 months of age (e.g., Sloane, Baillargeon, & Premack, 2012; also see Hamlin, Wynn, & Bloom, 2007), and have developing moral frameworks by age five (e.g., Kohlberg, 1976; Rochat et al., 2009; Turiel, 2006). Moral beliefs may be particularly steadfast because, as Skitka (2010) observes, “to support alternatives to what is right, moral and good is to be absolutely wrong, immoral, if not evil”. Yet, people can and often do change their moral beliefs, so it seems that research on moral conviction and moral development bring a puzzling psychological picture into focus. Moral beliefs are at once stubbornly intransigent and resistant to revision, and yet their change and development throughout people’s lifespan is almost inevitable.

Yet, assimilating new evidence into one’s moral framework is not a trivial task. The vast literature on confirmation bias has shown that people have an overwhelming tendency to seek out evidence that confirms their beliefs, and to interpret any new evidence (even evidence objectively contrary to their beliefs) as confirmatory (e.g., Klayman, 1995; Nickerson, 1998). This resistance to change has also been observed in the moral domain: for instance, people do
not temper their credences about the death penalty in the face of compelling contradictory evidence. Rather, they often view evidence that contradicts their belief as confirmatory, leading to even more polarized attitudes (e.g., Lord et al., 1979).

There are, of course, instances where people do revise even their deeply held beliefs. For a particularly notable example, consider former United States Federal Reserve chair Alan Greenspan, who changed many of his views on U.S. government regulation in light of the economic crises of 2008 (Time, October 23, 2008). The overarching point of confirmation bias is not that people never change their minds -- just that it takes more evidential support than seems (optimally) warranted, such as a global economic collapse. Even here, Greenspan altered his beliefs about economic policy, but there is reason to think his attitudes about the moral or normative aspects of economic policy did not change. In describing his new position on increasing market regulation, Greenspan prefaced by saying, “As much as I would prefer it otherwise, in this financial environment I see no choice …” (emphasis ours). All of this research raises the question, what kind of evidence leads people to revise their moral beliefs?

We propose that the method of cases, a common practice in philosophy (e.g., Nagel, 2012), provides an avenue for approaching this question. In ethical theory, moral dilemmas (i.e., a kind of case) are constructed with the aim of advancing or countering moral theories. Consider the Transplant dilemma, a case in which one must decide whether to harvest a single person’s organs in order to save the lives of five dying patients. Many philosophers think it is intuitively wrong to kill the single person, even to save the many (e.g., Kamm, 2007; Thomson, 1976), and therefore see their judgments about this dilemma as inconsistent with utilitarianism.\(^\text{12}\) Likewise,

\(^{12}\) Crudely put, utilitarianism is the view that one should maximize utility. However, see Parfit (2011) for an argument that philosopher’s intuitions about this dilemma do not actually run contrary to utilitarianism.
we hypothesized that people would view their judgments about the Transplant dilemma as being inconsistent with some of their moral beliefs (for instance, the belief that you should always save the most lives). If so, then exposure to this dilemma may lead to moral belief revision.

2. Coherence, the method of cases, and belief revision

An important epistemic standard for evaluating a belief is how it coheres with one’s other beliefs. We think that coherence may be particularly important for ethics, where there is little hope of independent and objective verification of beliefs through, say, perception (e.g., Campbell & Kumar, 2012). Is there any evidence that people value coherence when evaluating their own or other’s beliefs? First, research on cognitive dissonance suggests that incoherence between beliefs and actions can lead to psychological discomfort—motivating people to revise their beliefs (see Cooper, 2007 for a review). Second, people’s distaste for hypocrisy also suggests a certain level of concern for coherence. For example, consider the U.S. congressman from Florida who was recently charged with cocaine possession. The public seemed less concerned with his crime (a misdemeanor) than with his prior vote in favor of a drug-testing requirement for welfare recipients (New York Times, January 27, 2014). More than his drug use, it was his inconsistency that was repugnant.

Still, maintaining coherence in one’s system of moral beliefs is not always easy. People hold a multitude of moral beliefs, some portion of which are recruited when making judgments in any particular moral situation (e.g., Nichols & Mallon, 2006; Nichols, Kumar, & Lopez, Draft). It’s possible that incompatibilities between stored moral beliefs could go largely unnoticed, as only a handful of beliefs may ever be actively represented together. In contrast,
considering moral dilemmas pits opposing attitudes against one another, making salient the tensions between the beliefs one recruits when considering these dilemmas (e.g., Nichols, 2004; Prinz, 2007). This leads us to predict that making judgments about moral dilemmas will lead to moral belief revision by encouraging people to be coherent.

This account suggests that considering a moral dilemma is quite different from considering the sorts of evidence typically used in research on confirmation bias. In these studies, participants are often presented with external evidence (e.g., relevant statistics about the efficacy of the death penalty at deterring murder) that contradicts a belief they hold. Faced with external conflict, participants tend to either reject the presented evidence, or distort it until it conforms to their beliefs. However, when considering a case like the Transplant dilemma, tension arises due to an internal conflict among participants’ own beliefs. This difference, between internal and external conflict, motivates our prediction that considering a moral dilemma will lead to moral belief revision, even though external evidence typically does not.

Consistent with this prediction, one recent study suggests that people revise their beliefs about the impermissibility of killing after considering moral dilemmas. Horne, Powell, and Spino (2013) asked participants to make a judgment about a moral dilemma and then immediately after to rate their agreement with different moral beliefs.13 These researchers found that considering the footbridge dilemma -- a moral dilemma that tends to elicit deontological judgments -- led people to lower their credence in the utilitarian belief, “In the context of life or death situations,

---

13 This study primarily examined the connection between belief revision and transfer or “ordering effects” currently being discussed in the literature (see Schwitzgebel & Cushman, 2012; Wiegman & Waldmann, 2014 for discussions of these effects).
always take whatever means necessary to save the most lives.”

They concluded that people revise their moral beliefs after exposure to a single dilemma.

It is striking that these researchers did not have to coax a belief change from their participants. Rather, participants seemed to recognize that their deontological judgments were inconsistent with a utilitarian belief they otherwise strongly agreed with, and updated their agreement with the belief accordingly.

3. The present research: Does the method of cases actually lead to moral belief revision?

Although suggestive, Horne et al.’s (2013) findings leave open two important questions concerning (1) the nature of the changes in people’s moral beliefs and (2) the mechanisms behind these changes.

First, it is unclear whether the changes that Horne and colleagues (2013) observed in their participants’ belief ratings are truly reflective of changes in their participants’ underlying beliefs. For example, their participants may have changed their belief reports to appear consistent (e.g., in response to demand characteristics), without actually revising their beliefs. Another possibility is that Horne and colleagues’ findings reflect simple emotional priming, or some other transient contextual factor caused by considering an emotionally evocative moral dilemma. If the changes in participants’ belief reports reflect genuine changes in their moral beliefs, then they should remain relatively stable over time. In contrast, if demand characteristics or emotional priming can explain participants’ reports, then any shift in belief ratings should be fleeting.

---

14 In the footbridge dilemma, a runaway trolley threatens to kill five workmen standing on the tracks. A large man is standing on a footbridge over the tracks. Participants are told that the only way to save the lives of the five workmen is to push the man off the bridge and onto the tracks below. His body will stop the trolley but he will die.
A second question concerns the mechanisms underlying the hypothesized belief revision process. Does moral belief revision occur spontaneously, as a side-effect of simultaneously representing two conflicting beliefs, or does it require effortful deliberation, for example in response to explicit questions about one’s beliefs? This latter possibility suggests that people will not revise their moral beliefs unless they are immediately prompted to do so. On the other hand, if participants revise their beliefs without immediate prompting then this would suggest the process occurs more spontaneously, further cutting against the possibility that people only change their belief reports to appear consistent. Regardless of whether belief revision in response to moral dilemmas requires prompting or is spontaneous, demonstrating that people change their beliefs in response to such dilemmas would provide an avenue for overcoming confirmation bias. Although one’s beliefs may be comparatively impervious to external, contradictory evidence, perhaps they are less impervious to one another.

We conducted two experiments to investigate moral belief revision after exposure to a single dilemma. In Experiment 1, we examined how considering the Transplant dilemma affected participants’ agreement with a utilitarian belief both immediately following exposure to the dilemma and after a six-hour delay. We found that exposure to a dilemma induced changes in people’s moral beliefs that remained stable over a delay, suggesting that the method of cases provides an avenue for approaching moral belief change. Then, in Experiment 2, we examined whether participants would spontaneously revise their beliefs in response to the Transplant dilemma by asking participants to rate their agreement with moral beliefs only after a delay.

4. General Methods for Experiments 1 and 2

**Belief Selection.** Much of the research in moral psychology has focused on utilitarian moral judgment (e.g., Bartels & Pizarro, 2011; Greene, Nystrom, Engell, Darley & Cohen, 2004;
Greene Sommerville, Nystrom, Darley, & Cohen, 2001). We examined one operationalization of utilitarianism previously used in the literature: “In the context of life or death situations, always take whatever means necessary to save the most lives.” Existing research has found that the majority of people agree with this moral principle, and that their agreement ratings tend to be quite strong (Horne et al., 2013).

Although we assume that people’s reports of their beliefs are reliable indicators of those beliefs, recent data also suggests that utilitarian principles guide people’s judgments about moral dilemmas. Indeed, researchers have argued that people’s judgments about these moral dilemmas do not owe purely to the idiosyncrasies of each case, but are instead influenced by general moral rules, some of which are utilitarian in nature (e.g., Horne et al., 2013; Lombrzo, 2009; Nichols & Mallon, 2006; Nichols et al., Draft; Royzman, Goodwin, & Leeman, 2011). For instance, Lombrzo (2009) found that participants’ agreement with utilitarian principles predicted their judgments about moral dilemmas. In line with this finding, several studies have shown that people tend to make utilitarian moral judgments--choosing to sacrifice the life of one person to save many--across a number of different moral dilemmas (e.g., Côté, Piff, & Willer, 2013; Greene et al., 2001; 2004; Cushman, Young, and Hauser, 2006; Hauser, 2006; Moore, Clark, and Kane, 2008). So, it does not appear that participants naively assent to utilitarian principles without appreciating at least some of their negative implications. Of course, this is not to say that

---

15 There may be more than one reading of the utilitarian statement. We intended for participants to get a strong utilitarian reading, recognizing that “whatever means necessary” includes acts like killing an innocent person. However, it is also possible that participants get a weaker utilitarian reading on which these sorts of acts were not salient. Given the context of the utilitarian statement (i.e., “In life or death situations ...”), we think it is quite plausible that participants imagine “whatever means necessary” would include acts like killing.
people wholeheartedly endorse the doctrine of utilitarianism, as some philosophers do; it is quite likely that they also hold deontological beliefs (e.g., Nichols et al., Draft).

**Dilemma Selection.** Prior research suggests that people revise their moral beliefs after considering the footbridge dilemma (Horne et al., 2013) -- a moral dilemma that tends to elicit deontological moral judgments (e.g., Greene et al., 2001; Greene et al., 2004). However, as many people in our target population (Amazon Mechanical Turk workers) have been exposed to this dilemma in prior studies (e.g., Bauman, McGraw, Bartels, and Warren, 2014), we chose to use the Transplant dilemma, which also elicits high rates of deontological judgments (e.g., Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008), but has been less widely examined.

**Power Analysis.** An *a priori* power analysis was conducted based on an expected effect size estimated from the analyses reported by Horne et al. (2013). A good point of comparison for the current experiments is provided by the contrast between their participants’ utilitarian agreement ratings after reading the footbridge and control dilemmas, in which they observed a large effect ($d = 1.10$; Cohen, 1988). A power analysis conducted in G*power (Faul & Erdfelder, 1998) revealed that at least 23 participants would be required in each experimental group in order to achieve 95% power to detect effects of this size ($\alpha = .05$). Because it was difficult to predict dropout rates in our studies, we often exceeded the sample size required for 95% power.

5. **Experiment 1**

In Experiment 1 we sought to examine whether changes in participants’ moral beliefs would remain stable over time. We hypothesized that people would revise their moral beliefs after considering the Transplant dilemma. As we have discussed, people almost universally agree that it is inappropriate to harvest a single patient’s organs to save others, despite the fact that this action “maximizes the good” (Greene et al., 2008). Likewise, people also strongly endorse the utilitarian
moral principle "In the context of life or death situations, always take whatever means necessary to save the most lives" (Horne et al., 2013). We predicted that participants in our study would revise their agreement with this principle after making a judgment about the Transplant dilemma, because they would view their judgments about the Transplant dilemma as inconsistent with this principle. If our participants genuinely change their agreement with this utilitarian belief, then their revised agreement ratings should remain stable even after a delay. In contrast, if participants’ revisions are no longer present after a delay, then this would suggest that their reports changed due to demand characteristics, emotional priming, or other transient contextual factors. The first outcome would provide substantial evidence that moral dilemmas, and the method of cases more generally, can lead to belief revision. The second outcome may undermine this proposal.

Method

Participants. This experiment was conducted online with 195 participants (49% female; mean age of 34.9 years) recruited via the Mechanical Turk Work Distribution website between the times of 10:00 am and 11:30 am Central Standard Time. These participants were then encouraged to return later that same day between 5:00pm and 6:30pm to complete part two of the experiment. Of the 195 participants who initially completed part one, 67 returned and completed part two. Total compensation for completing both parts of the experiment was $1.50.
Design, Materials and Procedure.

Figure 6. Diagram summarizing procedures used in Experiment 1. Participants progressed through stages of the study in order from top to bottom.

This study utilized a 2 x 2 (vignette x delay) factorial design. We manipulated the moral vignette that participants read immediately before reporting on their moral beliefs: participants were assigned to either the Control or the Transplant condition (vignette; between-subjects). Participants were also asked to rate their agreement with the utilitarian moral belief both immediately following their moral judgments about the Transplant dilemma and after a delay of approximately six hours (delay; within-subjects). Figure 6 provides a diagram summarizing the experimental procedures.

Part one of the study began with a judgment task. Participants were presented with a series of three moral and three non-moral vignettes and were asked to make judgments about these situations. In the Transplant condition, the judgment task began with five distractor vignettes and ended with a judgment about the Transplant case. In the Control condition, the judgment task ended with a non-moral control vignette, and the transplant case was replaced.
with a different moral vignette (a vignette about incest) in order to maintain the same number of moral and non-moral distractors. Participants made their moral judgments on a 6-point Likert scale with endpoints labeled “Completely Inappropriate” and “Completely Appropriate.”

After the judgment task, participants completed a belief task: Participants rated their agreement with six statements expressing moral and non-moral beliefs. They made these ratings using a sliding scale, the endpoints of which were labeled “Completely Disagree” and “Completely Agree.” In order to make it difficult for participants to remember their responses to these belief statements in the second part of the experiment, there were no hashmarks or numerical indicators of their response. The only visual feedback provided by the slider was the position of the marker relative to the endpoints. The position of the marker was coded by the computer into a response between 0 and 100. Our primary dependent measure was participants’ ratings of the utilitarian belief statement “In the context of life or death situations, you should always take whatever means necessary to save the most lives.” This statement was presented at the beginning of the belief task, immediately following the moral judgment task. Five other belief statements, two-moral and three non-moral, were presented after the utilitarian belief statement. An example of a non-moral belief statement was, “I laugh out loud when someone tells me a joke that I think is funny,” and a moral belief statement was, “Incest is always morally wrong.” Additionally, half of these belief statements corresponded to the stories that participants made judgments about earlier in the experiment. The moral and non-moral distractors and belief statements were intended to conceal the aim of the experiment, and further reduce the likelihood of demand characteristics. After rating the six belief statements, participants completed a series of comprehension questions to ensure they paid attention to the stories that they read and then were asked to participate in part two of the experiment.
Part two was identical for participants assigned to either condition. Participants first made judgments about four unrelated distractor vignettes (two novel, two old; two moral, two non-moral), none of which were the Transplant dilemma. After making judgments about these vignettes, participants again rated their agreement with five distractor belief statements and the utilitarian belief (our primary dependent measure). Three of the belief statements were originally presented in part one of the experiment and three were novel.

**Results and Discussion**

- **Figure 7.** Average agreement ratings for the utilitarian belief statement across vignette and delay conditions in Experiment 1. Error bars represent ± 1 SE.

Participants’ agreement ratings for the utilitarian belief statement are shown in Figure 7. Of the 195 participants who initially completed part one, 67 returned and completed part two. To rule out the possibility of self-selection effects, we compared the immediate utilitarian belief ratings of those participants who returned for part two and those who dropped out. Results of two
independent samples *t*-tests revealed no reliable differences between these two groups in either the Control condition (*t*(99) < 1, *p* = .40) or the Transplant condition (*t*(94) = 1.48, *p* = .14).

We computed a two-way repeated measures ANOVA to determine how participants’ endorsements of the utilitarian belief were affected by reading the Transplant dilemma and by the delay. Participants in the Transplant condition gave lower ratings to the utilitarian belief than those in the Control condition, as indicated by a reliable main effect of Vignette, (*F*(1, 134) = 11.32, *p* < .001). Moreover, and consistent with our prediction that people’s beliefs would remain stable over time, participants’ ratings were not affected by an approximately six-hour delay (*F*(1, 134) < 1, ns. Moreover, there was no reliable interaction between Vignette and Delay, *F*(1, 134) < 1, ns. Subsequent independent *t*-tests revealed that participants in the Transplant condition gave reliably lower agreement ratings for the utilitarian belief statement both immediately following the Transplant dilemma in part one, (*t*(67) = 3.26, *p* < .01, *d* = .79) and after a delay in part two (*t*(67) = 2.43, *p* < .05, *d* = .59). Both of these effects are conventionally considered medium to large effects (Cohen, 1988). Finally, a paired sample *t*-test found no reliable difference in belief ratings in the Transplant condition between parts one and two of the experiment, *t*(38) = .92, *p* = .36.

**Discussion**

Consistent with our predictions, the results of Experiment 1 demonstrate that people revise their beliefs after exposure to a moral dilemma and that this revision persists for at least six hours after the dilemma is considered. The stability of participants’ revised beliefs suggests that this effect is not the product of emotional priming or other transient contextual factors. Rather, the results suggest that people consider their judgments about the Transplant dilemma as
inconsistent with a utilitarian belief about sacrificing others to save lives, and revise their agreement with this utilitarian belief accordingly.

6. Experiment 2

Experiment 1 suggests that the method of cases may provide a means for inducing moral belief revision. Still, one possibility is that people will only revise their beliefs when they are asked about the utilitarian belief immediately after considering the Transplant dilemma. That is, the mechanisms underlying moral belief revision remain unclear. Experiment 2 examined whether participants would spontaneously revise their beliefs without prompting. This experiment was identical to Experiment 1, but with two additional conditions. In these new conditions, participants only rated their agreement with the utilitarian belief after the delay. If participants spontaneously revise their agreement with the utilitarian belief after considering the Transplant dilemma, then we should be able to observe the effects of this revision even when participants are not immediately prompted to rate their agreement with this belief.

Methods

Participants. This experiment was conducted online with 373 participants (52% female; mean age of 36 years) recruited via the Mechanical Turk Work Distribution website. These participants were then encouraged to return later that same day to complete part two of the experiment. Of the 373 participants who initially completed part one, 120 returned and completed part two. Three participants were removed for completing the study more than once,
leaving a sample of 117 participants. Total compensation for completing both parts of the experiment was $1.50.

**Materials and Procedure.** The stimuli and procedure of Experiment 2 were identical to those of Experiment 1, except for the inclusion of two additional conditions. In the Transplant Delay-Only and Control Delay-Only conditions, participants rated their agreement with the utilitarian belief only after the delay. In all other respects these conditions were identical to the Transplant and Control conditions in Experiment 1. Figure 3 provides a summary diagram of the procedures in Experiment 2.

![Diagram summarizing procedures used in Experiment 2. Participants progressed through stages of the study in order from top to bottom.](image)

*Figure 8.* Diagram summarizing procedures used in Experiment 2. Participants progressed through stages of the study in order from top to bottom.
The results of Experiment 2 are shown in Figure 9. First, to rule out the possibility of self-selection effects, we again compared the immediate utilitarian belief ratings of those participants who returned for part two and those who dropped out. The results of two independent samples t-tests revealed no reliable differences in participants’ immediate utilitarian belief ratings between these groups. This held for both the Control condition \( t(96) < 1.66, p = .10 \) and the Transplant condition \( t(94) = 1.69, p = .09 \).

The distribution of data in Experiment 2 was non-normal, rendering parametric statistics inappropriate for any subsequent hypothesis testing. Since parametric effect sizes were also inappropriate, we calculated a non-parametric common language effect size (McGraw & Wong, 1992) \( CL \), as recommended by Grissom & Kim (2012). This statistic represents the probability
that a score randomly drawn from one population A will be greater than a score randomly drawn from another population B.

First, we computed a series of Mann-Whitney U tests to determine whether Experiment 2 replicated the results of Experiment 1. Our findings replicated all previously observed effects, demonstrating again that participants revised their beliefs after reading the Transplant dilemma, and that those revisions remained stable even after a significant delay. Agreement ratings for the utilitarian belief were higher in the Control condition than in the Transplant condition when tested immediately ($U(27, 36) = 792, p < .001, CL = .815$), and after a delay ($U(27, 36) = 773.5, p < .001, CL = .796$). Participants’ immediate and delayed belief ratings were examined with a related-samples Wilcoxon Signed Rank test. The ratings did not reliably differ over time in the Transplant condition ($W(26) = 430.5, p = .058$), although this effect approached significance. However, a similar trend was observed in the Control condition ($W(27) = 200, p = .059$), suggesting that this trend was not associated with our experimental manipulation.

Of particular interest in Experiment 2 was whether people would spontaneously recruit and revise their beliefs after considering the Transplant dilemma—that is, whether participants’ agreement ratings with the utilitarian belief would be lower in the Delay-Only Transplant condition than in the Delay-Only Control condition. This test revealed a significant effect, where participants in the Delay-Only Transplant condition gave lower agreement ratings for the utilitarian belief than participants in the Delay-Only Control condition, $U(27, 26) = 471.5, p = .032, CL = .673$. We also compared participants’ delayed utilitarian agreement ratings from the Immediate-Delay Transplant condition and the Delay-Only Transplant condition. There were no reliable differences in their agreement ratings ($U(36, 26) = 360, p = .12$), even though participants in the Immediate-Delay Transplant condition also rated their agreement with the
utilitarian belief immediately after considering the Transplant dilemma earlier that day. These results are particularly striking as they suggest that people spontaneously recruit and revise their beliefs when considering moral dilemmas.

7. General Discussion

How is it that moral convictions are at once steadfast, and yet people can and do change their moral beliefs? And what factors can lead to authentic moral belief change? The present research demonstrates that considering a moral dilemma can produce authentic change in people’s moral beliefs. We found that making a judgment about the Transplant dilemma changed people’s beliefs in the utilitarian principle, “In the context of life or death situations, always take whatever means necessary to save the most lives.” This was the case both immediately following exposure to the dilemma and six hours later. Experiment 2 demonstrated that this change occurs even when participants are not immediately queried about their agreement with the utilitarian principle. This latter result suggests that being exposed to moral dilemmas may lead people to spontaneously revise their beliefs, even when they are not prompted to do so.

It is unlikely that participants’ belief reports owe simply to a desire to appear consistent, as our experiments included distractor vignettes and distractor belief statements that concealed our aims. Additionally, research on analogical problem solving has shown that, in the absence of featural support for memory retrieval, it is rare for participants to recall a relevant problem when attempting to solve a new problem, even after a short delay (e.g., Gick & Holyoak, 1983; Wharton & Holyoak, 1996). Accordingly, it is unlikely that participants in Experiment 2 recalled the Transplant dilemma when they rated their agreement with the utilitarian belief six hours later. Instead, a more plausible interpretation of the results of Experiment 2 is that exposure to the
Transplant dilemma changed participants’ utilitarian beliefs when they first encountered the dilemma. These changes then persisted, affecting agreement ratings six hours later.

8. What are the mechanisms underlying moral belief revision?

Our results suggest that moral dilemmas can lead people to revise their beliefs by pitting inconsistent moral beliefs against one another. This marks a significant advance for understanding the psychological processes underlying moral conviction and confirmation bias. It seems that people possess many moral beliefs that can be mapped onto a specific moral situation (e.g., Nichols & Mallon, 2006; Prinz, 2007). Once such a mapping is computed, a belief may suggest a certain response to the moral situation under consideration. When thinking about a moral dilemma, people might successfully map multiple beliefs to the same situation. If different beliefs entail different responses to the same situation, then this suggests that the elicited beliefs are inconsistent. In this case, one or more beliefs must be revised in order to restore coherence. For example, our participants may have reduced their agreement with the utilitarian belief after considering the Transplant dilemma because they mapped a utilitarian belief to the dilemma but ultimately recognized that their judgment about the dilemma was inconsistent this belief.

Of course, there are likely a number of factors that could interfere with this type of mapping, such as cognitive load or increased complexity. Additionally, if convictions are held strongly enough, it is possible that people might respond to perceived inconsistency by rejecting the mapping between a case and their belief. This outcome seems especially likely for convictions that have been publicly expressed, such as when people express positions on political issues, which have often been the focus of research on moral conviction (e.g., Skitka, 2010). Further research is necessary to more completely understand the connection between people’s judgments about dilemmas and moral belief revision.
9. Conclusion

The present studies reveal a means for inducing moral belief revision. Rather than presenting people with data that directly contradict their existing moral beliefs—and which could therefore be ignored or discounted due to confirmation bias—moral dilemmas, and the method of cases more generally, may circumvent confirmation bias by highlighting existing (but perhaps heretofore unnoticed) inconsistencies in people’s beliefs. While moral belief revision is a necessary first step toward moral progress, contemporary moral psychologists have often seen this step as insurmountable. This study gives reason for optimism, as we have provided a method for inducing moral belief revision.16

16 The proceeding chapter was coauthored with Derek Powell and John E. Hummel.
CHAPTER 7
ORDERING EFFECTS AND UPDATING EFFECTS

Some experimental philosophers have argued that intuitive judgments in ethics, epistemology, metaphysics, philosophy of mind, and philosophy of language are influenced by demographic factors, like personality (Schulz et al., 2011, Feltz & Cokely, 2012), socioeconomic status (e.g., Haidt et al., 1993), and culture (e.g., Machery et al., 2004). On the basis of these results, some experimental philosophers argue that because demographic factors do not track the truth about philosophical claims, intuitive judgments are unreliable. Consequently, they argue that the fact that a judgment is intuitive is not a good epistemic reason to believe that the judgment is correct. Likewise, some experimental philosophers argue that intuitive judgments in philosophy are strongly influenced by various contextual or situational factors, like whether one is reading a concrete vignette or a more abstract one (e.g., Nichols & Knobe, 2007), whether one is eating something sweet or something bitter (e.g., Eskine et al., 2011), and whether we take the perspective of an actor or an observer (e.g., Nadelhoffer & Feltz, 2008). Some experimental philosophers argue that because these contextual factors do not track the truth about philosophical claims, intuitive judgments are unreliable. Again, they conclude that the fact that a judgment is intuitive is not a good epistemic reason to believe that the judgment is correct. The upshot of both lines of argument is that the contents of intuitive judgments should not be treated as providing significant evidence for or against a philosophical theory in virtue of their apparent agreeableness to reason.

According to one widely-endorsed contextual argument, intuitive judgments are unreliable because they are influenced by the order in which thought experiments prompting those judgments are presented. For example, Tobia, Buckwalter, and Stich (2012) point out that:
Research in social psychology and experimental philosophy ... has shown that ordinary people’s moral intuitions are influenced by a variety of factors including order effects, framing effects, and environmental variables. Since it is widely agreed that those factors are irrelevant to the truth or falsity of the intuition, these empirical results cast doubt on the use of intuition as evidence for moral claims (3).17

The same basic argument has been deployed against the reliability of intuitive judgments in ethics (e.g., Schwitzgebel & Cushman, 2012; Petronovich et al., 1996; Sinnott-Armstrong et al., 2008; Tobia et al., 2012; Laio et al., 2011), in epistemology (e.g., Swain et al. 2008), and in philosophy of mind (e.g., Feltz & Cokely, 2012). Each of these arguments against the reliability of intuitive judgments has at its core what we will call The Argument from Ordering Effects, which we think may be fairly stated as follows:

[ Normative Principle ] If judgments about what some thought experiments show are systematically affected by the order in which those thought experiments are considered, then the judgments about what those thought experiments show are unreliable.

[ Empirical Observation ] Intuitive judgments about what some thought experiments show are systematically affected by the order in which the thought experiments are considered.

[ Conclusion ] Intuitive judgments about what some thought experiments show are unreliable.

Although there is considerable controversy about the empirical observation (e.g., Bengson, 2013; Cullen, 2010; Sosa, 2007), the normative principle has gone unchallenged. All parties involved

17 Throughout this chapter, we will talk about intuitive judgments and the (propositional) contents of intuitive judgments. In our experience, the term “intuition” too easily slips back and forth between a mental state and the (propositional) content of a mental state. Readers who think that intuitions are non-propositional intellectual seemings, inclinations to make judgments, or something similar may make suitable substitutions in our arguments and in the arguments that we attribute to various experimental philosophers.
seem to agree that *if* intuitive judgments about what some thought experiments show are sensitive to the order in which those thought experiments are presented, *then* intuitive judgments about those thought experiments are unreliable.

If the argument from ordering effects is successful, then it may plausibly be extended via a generalization step or steps. If intuitive judgments about *these* thought experiments are unreliable, then intuitive judgments about all (or most all) relevantly similar thought experiments are unreliable. All of the thought experiments in some wider class are relevantly similar to these thought experiments. Therefore, we should be skeptical of the reliability of intuitive judgments in that wider class. How to pick exactly *which* wider class of judgments is thrown into doubt—or more generally, how to evaluate the degree of support given to various generalizations by the conclusion of the argument from ordering effects—is an interesting problem of inductive logic. Experimental philosophers often endorse a very wide generalization. As we saw above, Tobia et al. (2012) generalize to all moral claims; Swain et al. (2008) offer a more circumspect objection to intuitive judgments but without identifying any particular class of thought experiments to which skepticism should apply.\(^{18}\) Sytsma et al. (forthcoming) explicitly consider a range of possible generalizations from new cross-cultural results in experimental philosophy of language that they report.

---

\(^{18}\) In their conclusion, Swain et al. write: “Specifically, we found that intuitions about the Trutemp Case vary depending on whether, and which, other cases are presented before it. Such variability calls into question the legitimacy of using the intuitions generated by the Trutemp Case as evidence against reliabilism. But it is unclear what about this case makes it susceptible to these effects, *which raises questions about the reliability on intuitions about thought-experiments more generally*, especially given that this is not the only case called into question by empirical research” (153, emphasis added). They go on to write: “We certainly do not take ourselves to have offered anything like a general proof of the unreliability of all intuitions (nor do we think that any such proof would be either possible or desirable). But we do take ourselves to have raised a serious empirical worry that philosophers need to begin deciding how to address.”
In the present chapter, we show that the argument from ordering effects sketched above is defective. The normative principle is ambiguous with respect to what it means for judgments about thought experiments to be systematically affected by the order in which the thought experiments are considered. We claim that owing to the ambiguity in the normative principle the argument from ordering effects is either invalid or unsound. Invalid because the kind of sensitivity to ordering mentioned in the empirical observation is not the same kind of sensitivity mentioned in the normative principle or unsound because the normative principle, interpreted in line with the empirical observation, is false. Having established that the argument from ordering effects is defective, we go on to consider one way that it might be rehabilitated: endorse a revised normative principle, and conduct experiments showing that intuitive judgments run afoul of that principle. However, we point out that psychological research shows that many cognitive faculties typically considered sources of justification run afoul of the revised normative principle we consider. Consequently, the revised normative principle risks leading to a highly skeptical thesis about the reliability of all or almost all of our cognitive faculties, not just the faculties responsible for our intuitive judgments. Since we are suspicious of arguments that lead to global or near-global skepticism, we think something has gone wrong either in the argument from ordering effects itself or in the generalization step. We consider whether the skeptical conclusion might be resisted by rejecting the revised normative principle that drives the rehabilitated argument from ordering effects.

The plan of the paper is as follows. In Section 1, we distinguish two kinds of effect that one might observe in experiments where more than one thought experiment is presented to participants: (genuine) ordering effects and updating effects. We point out that the normative principle in the argument from ordering effects may be interpreted either as saying that observing
an updating effect provides prima facie reason to think that the relevant cognitive process is unreliable or as saying that observing a genuine ordering effect provides prima facie reason to think that the relevant cognitive process is unreliable. We then show that if the normative principle were understood as referring to updating effects, then according to the normative principle, the act of updating one’s beliefs on the basis of evidence (i.e. learning) would be prima facie reason to think that the relevant cognitive process was unreliable. To learn would be to indict one’s ability to know things. Therefore, if the normative principle is to be plausible, it must be understood as referring to genuine ordering effects, not updating effects.

In Section 2, we argue that the empirical premise actually supported by the work of experimental philosophers is the unthreatening claim that intuitive judgments exhibit updating effects. In Section 3, we consider the possibility of rehabilitating the argument from ordering effects. We assume for the sake of argument that intuitive judgments exhibit genuine ordering effects, and we restate the normative principle in terms of genuine ordering effects. We then argue that to endorse the revised normative principle would potentially call perception, memory, testimony, and reasoning into question, since all of these human cognitive processes have been shown to exhibit genuine ordering effects. Finally, in Section 4, we consider two ways in which one might resist the argument from ordering effects and its generalization: (1) reject the generalization step on the grounds that experiments that have been conducted so far form an inadequate inductive basis; or (2) reject the revised normative principle on the grounds that the fact that a cognitive process exhibits a genuine ordering effect is not prima facie reason for thinking that the process is unreliable.

1. Ordering Effects and Updating Effects
In this section, we draw a distinction between genuine ordering effects and updating effects. We argue that whereas one might plausibly think that the fact that a cognitive process exhibits a genuine ordering effect is prima facie reason to think it is unreliable, the fact that a cognitive process exhibits an updating effect is not (by itself) even prima facie reason to think that it is unreliable.

1.1 Genuine Ordering Effects

Let us introduce some terminology. A genuine ordering effect occurs when each participant in a study is presented the same collection of stimuli, some participants receive the stimuli in one order, some participants receive the stimuli in a different order, and participants who receive stimuli in different orders perform differently on some task that they complete after they have received all of the stimuli in the collection. We want to emphasize that experimentalists must present all of the stimuli before measuring participants’ performance on a task in order to observe a genuine ordering effect. Our use of the term genuine ordering effect thus agrees with the account given by Hogarth and Einhorn in their seminal 1992 paper on belief adjustment:

There are two pieces of evidence, A and B. Some subjects express an opinion after seeing the information in the order A-B; others receive the information in the order B-A. An order effect occurs when opinions after A-B differ from those after B-A.

To illustrate, consider the following experiment. Participants are randomly assigned to one of two conditions and asked to determine the probability that a defendant is guilty. All participants are presented the same two pieces of evidence related to the defendant’s guilt: some DNA evidence and some eye-witness evidence. Participants in Condition 1 are first presented the DNA evidence and then presented the eye-witness evidence. After being presented all of the evidence, participants in Condition 1 assign some credence to the claim that the defendant is guilty. For
example, we might imagine that the average credence reported by the participants in Condition 1 is 0.4 that the defendant is guilty. By contrast, participants in Condition 2 are presented *exactly the same two pieces of evidence* in the reverse order. Participants in Condition 2 are first presented the eye-witness evidence and then presented the DNA evidence. Again, after being presented all of the evidence, participants in Condition 2 assign some credence to the claim that the defendant is guilty. For example, we might imagine that the average credence reported by the participants in Condition 2 is 0.8 that the defendant is guilty. A difference in the credences reported by the two groups would indicate the presence of a genuine ordering effect. Our imagined experiment, in which we observe a genuine ordering effect, is pictured in Table 7:

### Condition 1

<table>
<thead>
<tr>
<th>Given DNA evidence</th>
<th>Given eye-witness evidence</th>
<th>Is the defendant guilty?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>Time 2</td>
<td>Time 3</td>
</tr>
</tbody>
</table>

### Condition 2

<table>
<thead>
<tr>
<th>Given eye-witness evidence</th>
<th>Given DNA evidence</th>
<th>Is the defendant guilty?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>Time 2</td>
<td>Time 3</td>
</tr>
</tbody>
</table>

Table 7: Experiment Eliciting a Genuine Ordering Effect

If the relevant task in an experiment is answering a question posed by the experimenter, then in order to observe a genuine ordering effect with respect to that question, participants in each experimental condition must be asked the question after being presented with a complete
sequence of stimuli in an order unique to their condition. A genuine ordering effect occurs if participants in different conditions answer the question in (significantly) different ways.

1.2 Updating Effects

An *updating effect* occurs when two groups of participants are presented different sets of stimuli before completing some task, the set of stimuli presented to one group is a proper subset of the set of stimuli presented to the other group, and the two groups perform differently on the task. To illustrate, consider the following variation on the experiment described in Section 1.1. As before, participants are randomly assigned to one of two conditions and asked to determine whether a defendant is guilty. Again, the order in which evidence is presented is different in the two conditions. However, unlike the previous case, participants are asked for a new judgment after seeing each piece of evidence. Participants in Condition 1 are first presented with DNA evidence that makes the defendant appear innocent, after which they are asked to make a judgment. Then they are presented with eye-witness evidence that makes the defendant appear guilty, after which they are again asked to make a judgment. Thus, Condition 1 looks like this:
Condition 1:  (A) DNA evidence.
Question: Is the defendant guilty?

(B) Eye-witness evidence.
Question: Is the defendant guilty?

Participants in Condition 2 are first presented with eye-witness evidence that makes the defendant appear guilty, after which they are asked to make a judgment. Then they are presented with DNA evidence that makes the defendant appear innocent, after which they are again asked to make a judgment. Thus, Condition 2 looks like this:

Condition 2:  (A) Eye-witness evidence.
Question: Is the defendant guilty?

(B) DNA evidence.
Question: Is the defendant guilty?

An updating effect occurs if the two groups of participants give different judgments about whether the defendant is guilty when prompted after seeing the same piece of evidence. For example, if participants in Condition 1 tend to say that the defendant is unlikely to be guilty after seeing the eye-witness evidence, and participants in Condition 2 tend to say that the defendant is likely to be guilty after seeing the eye-witness evidence, then the participants are exhibiting an updating effect. The imagined experiment is pictured in Table 8:
**Condition 1**

<table>
<thead>
<tr>
<th>Given DNA evidence: Is the defendant guilty?</th>
<th>Given eye-witness testimony: Is the defendant guilty?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>Time 2</td>
</tr>
</tbody>
</table>

**Condition 2**

<table>
<thead>
<tr>
<th>Given eye-witness testimony: Is the defendant guilty?</th>
<th>Given DNA evidence: Is the defendant guilty?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>Time 2</td>
</tr>
</tbody>
</table>

Table 8: Experiment Eliciting an Updating Effect

In this case, there is an updating effect because people’s judgments in the two cases are based on different total information. In Condition 2, the judgment that the defendant is guilty is based only on the eye-witness testimony. But in Condition 1, the judgment that the defendant is guilty is based on both the eye-witness testimony and the DNA evidence.

1.3 Updating Effects are Not Prima Facie Vicious

One might be tempted to think that the fact that some cognitive process exhibits a genuine ordering effect is sufficient reason, in the absence of defeaters, to think that the process is unreliable. However, one should have no temptation to think that cognitive processes are impugned when they exhibit updating effects, since updating effects occur whenever there are perfectly ordinary and normatively correct instances of learning.
One might already have been convinced by the example in Section 1.2 that updating effects are at least sometimes virtuous. But we want to drive the point home, so let us consider some further examples. Begin with an obvious—albeit a bit silly—case. Suppose Billy and Suzy are trying to guess the identity of a card chosen at random from a standard deck of playing cards. The card happens to be the two of hearts, and the possible clues include the following: (1) the card is the two of hearts; and (2) the card is red. In the first round, Billy receives Clue #2. He thinks that all the red cards are equally likely given his evidence, so he just takes the first thing that comes to mind and guesses that the card is the ace of diamonds. Suzy is luckier than Billy, and she receives Clue #1. She then “guesses” that the card is the two of hearts. In the second round, Billy receives Clue #1 and changes his guess to the two of hearts, while Suzy gets Clue #2 and keeps her initial guess. Billy and Suzy say different things after seeing Clue #2, but that is no threat to their epistemic reliability, since they have different total evidence at the times of their guesses.

On another day, Billy and Suzy go on a tour around Fake Barn Land. While on their trip, Billy learns that most of the things that look like barns from the road are fakes, and Suzy learns that everything that looks like a barn as seen from the Realview Bridge really is a barn. Subsequently, they each observe what appears to be a barn as seen from the Realview Bridge, and they make a judgment about whether or not it is a real barn or a fake. Then they share their evidence with each other and make a second judgment. Before sharing their evidence, Billy judges that he is not looking at a real barn, but Suzy judges that she is looking at a real barn. After sharing their evidence, Billy changes his judgment, but Suzy sticks with hers. Again, Billy and Suzy are exhibiting an updating effect, but again, their reasoning is not thereby discredited.
Both of the previous examples are straightforward illustrations of how an updating effect might be observed even when every participant in a study reasons correctly. Updating effects can arise in *subtle* ways as well. To illustrate, suppose that Billy and Suzy are Bayesian agents with a credence function $Cr(\cdot)$, and assume that they also endorse the following simple account of incremental confirmation: $d(h, e) = Cr(h | e) - Cr(h)$. That is, the degree to which some hypothesis $h$ is confirmed by some evidence $e$ is just the difference between the credence in $h$ given $e$ and the credence in $h$. Now, assume that Billy and Suzy both adopt the following convention for rating the strength of a piece of evidence: a piece of evidence $e$ is classified as *weak* if it confirms the hypothesis $h$ by less than $c = 0.085$, as *strong* if it confirms $h$ by more than $c = 0.15$, and as *modest* otherwise.

Now, suppose that Billy and Suzy are considering evidence at a trial, as in the example from Section 1.2, but instead of making a judgment about whether the defendant is guilty, they are asked to make a judgment about how strong each piece of evidence is. Suppose that before seeing any evidence, Billy and Suzy both think that the defendant is just as likely as not to be guilty, and hence, they both have initial credences $Cr(\text{guilty}) = Cr(\text{not-guilty}) = 0.5$. Further suppose that Billy and Suzy have the following conditional credences: $Cr(\text{DNA} | \text{guilty}) = 0.1$; $Cr(\text{DNA} | \text{not-guilty}) = 0.9$; $Cr(\text{eye} | \text{guilty}) = 0.8$, and $Cr(\text{eye} | \text{not-guilty}) = 0.4$. Given their credences, Billy and Suzy should say that $Cr(\text{guilty} | \text{eye}) = 0.67$, $Cr(\text{guilty} | \text{DNA}) = 0.1$, and $Cr(\text{guilty} | \text{eye, DNA}) = 0.182$. Hence, the degree of confirmation given to the hypothesis that the defendant is guilty by the eye-witness evidence alone is $c = 0.17$, which Billy and Suzy classify as strong evidence. But the degree of confirmation given to the same hypothesis by the eye-witness evidence *in light of the DNA evidence* is only $c = 0.082$, which Billy and Suzy classify as weak evidence. Suppose, then, that Billy receives the eye-witness evidence first,
while Suzy receives the DNA evidence first. Then Billy will say that the eye-witness evidence is strong, while Suzy will say that the eye-witness evidence is weak.

The moral of these examples is that the fact that some people exhibit updating effects is no indication that any of their cognitive processes are unreliable. In fact, whenever one learns something, some cognitive process exhibits an updating effect. A cognitive process might be unreliable because it “learns” in an inappropriate way. For example, if ordinary conditionalization is the normatively correct way to update on evidence, then a cognitive process might be unreliable because it updates in a way that is not equivalent to ordinary conditionalization. But the mere fact that some cognitive process exhibits an updating effect does not provide any reason to think that it is unreliable.

According to the normative principle in the argument from ordering effects, if one’s judgments about some thought experiments are reliable, then they should not depend on the order in which the thought experiments are considered. On its face, the normative principle might be asserting that if a cognitive process exhibits an updating effect, then it is unreliable. Alternatively, the normative principle might be asserting that if a cognitive process exhibits a genuine ordering effect, then it is unreliable. However, we have argued that the fact that a cognitive process exhibits an updating effect does not provide prima facie reason to think that the process is unreliable. Hence, we think that if the principle is to be plausible at all, it must be interpreted as asserting that reliable cognitive processes do not exhibit genuine ordering effects.

However, as we will discuss in the next section, almost all of the empirical evidence presented by experimental philosophers goes to establish that intuitive judgments exhibit updating effects, not that they exhibit genuine ordering effects. If we are right in thinking that the empirical premise actually supported by work in experimental philosophy is that some cognitive
processes are sensitive to updating effects, then friends of the argument from ordering effects face a dilemma. Either the argument is invalid because it equivocates on what it means for a judgment to be sensitive to ordering or the argument is unsound because it relies on a faulty normative principle.

2. Updating Effects and Experimental Philosophy

In Section 1, we observed that genuine ordering effects are different from updating effects, and we argued that while genuine ordering effects might be prima facie vicious, updating effects are not. In this section, we argue that with only a couple of exceptions, the empirical premise actually supported by the work of experimental philosophers is the unthreatening claim that intuitive judgments exhibit updating effects.

Our argument proceeds by first observing in Section 2.1 that almost all of the experiments alleged to establish the second premise of the argument from ordering effects have the same structure as our examples of updating effects. We consider a case study to illustrate this point. In Section 2.2, we sketch a model according to which people update their beliefs in the light of thought experiments that they consider—a proposal that coheres nicely with research suggesting that considering hypothetical scenarios can lead people to update their preferences and beliefs (e.g., Holyoak & Simon, 1999; Horne, Powell, & Spino, 2013; Horne, Powell, & Hummel, 2015). We then argue that thought experiments should be treated as pieces of evidence, and we consider some empirical evidence that people do, in fact, treat thought experiments as pieces of evidence.
2.1 Mistaking Updating Effects for Ordering Effects

The so-called ordering effects typically observed in the experimental philosophy literature are not genuine ordering effects. They are updating effects. To make our point we will first consider the quintessential experiment that allegedly elicits an ordering effect: the Footbridge-Trolley Experiment. In this experiment, participants are randomly assigned to one of two conditions. In each condition, participants are presented with two moral dilemmas and asked to evaluate the degree to which some action is permissible. The only difference between the two conditions is the order in which the dilemmas are presented.

One dilemma, which we will call the Trolley Case, describes the well-worn scenario in which an out-of-control trolley will kill five workmen on the track unless the trolley is diverted onto a side track by flipping a switch. If the trolley is diverted, the five workmen will be saved. However, if the trolley is diverted, an innocent person standing on the side track will be killed. Participants are asked to decide whether it is permissible to flip the switch killing the single person and saving the five workmen.

The other dilemma, which we will call the Footbridge Case, describes a variation on the Trolley Case in which the workmen may only be saved by pushing a large man from a footbridge onto the tracks, killing the man but stopping the runaway trolley. The question put to participants in the Footbridge Case is whether it is permissible to push the fat man onto the tracks. Suppose that participants are asked to rate permissibility on a 6-point Likert scale ranging from (1) definitely impermissible to (6) definitely permissible. In the interests of working with a concrete example, suppose that the mean rating given by participants in Condition 1 for the Footbridge Case is 2.2, and the mean rating given by participants in Condition 1 for the Trolley Case is 3.5. By contrast, suppose that the mean judgment of participants in Condition 2 for the Trolley Case
is 4.3, and suppose that the difference between the ratings for the Trolley Case in the two conditions is statistically significant. Finally, suppose that the mean rating for the Footbridge Case in Condition 2 is 2.1, which we may suppose is statistically indistinguishable from the mean rating for the Footbridge Case in Condition 1 (for real examples, see Horne, Powell, & Spino, 2013; Wiegmann et al., 2012; and Wiegmann & Waldmann, 2014). The experiment is pictured schematically in Table 9.

<table>
<thead>
<tr>
<th>Condition 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Footbridge Case</td>
<td>Trolley Case</td>
</tr>
<tr>
<td>Mean = 2.2</td>
<td>Mean = 3.5</td>
</tr>
<tr>
<td>Time 1</td>
<td>Time 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Trolley Case</td>
<td>Footbridge Case</td>
</tr>
<tr>
<td>Mean = 4.3</td>
<td>Mean = 2.1</td>
</tr>
<tr>
<td>Time 1</td>
<td>Time 2</td>
</tr>
</tbody>
</table>

Table 9: The Footbridge-Trolley Experiment

Participants make different judgments about the Trolley Case in the two conditions. However, participants do not make their judgments about the Trolley Case after seeing both dilemmas.
presented in different orders. Thus, the effect elicited in the Footbridge-Trolley Experiment is not an ordering effect. Rather, the judgments that participants make about the Trolley Case are elicited at different points depending on the experimental condition. In Condition 1, participants see both dilemmas before making a judgment about the Trolley Case, but in Condition 2, participants do not see the Footbridge Case before being asked to make a judgment about the Trolley Case.

Many philosophers have construed the effect in the Footbridge-Trolley Experiment—and specifically the shift in people’s judgments about the appropriate action in the Trolley Case (or cases like it)—as evidence that people fail to reason as they should (e.g., Liao et al., 2011). But the conclusion has been drawn too hastily. Structurally, the Footbridge-Trolley Experiment is the same as Billy and Suzy’s guessing game and the DNA/Eye-Witness example. The Footbridge-Trolley Experiment looks like it is eliciting an updating effect. Hence, we have a defeasible reason to think that the Footbridge-Trolley Experiment really is eliciting an updating effect. And we have no reason to think that the Footbridge-Trolley Experiment is eliciting a genuine ordering effect. However, one might be suspicious of the claim that the effect elicited in the Footbridge-Trolley Experiment is really an updating effect. Ordering may not unduly influence participants, but perhaps they are nonetheless biased in some way that makes their judgments suspect.

The burden of proof is clearly on skeptics of our claim that the effect in the Footbridge-Trolley Experiment is an updating effect. The experimental design suggests that it is an updating effect, and we have no positive reason to doubt it. Moreover, we have no reason to think that the updating effect exhibited in the Footbridge-Trolley Experiment is vicious, as it should be no surprise that people presented different information may form different judgments. But we can
do better than note a structural similarity and shift the burden of proof. We can offer a cognitive model that explains the evidence on the assumption that what we are seeing is an updating effect.

2.2 An Explanation of the Footbridge-Trolley Effect

We claim that the effect in the Footbridge-Trolley Experiment (and experiments like it) is an updating effect. Below we briefly summarize some data on the Footbridge-Trolley Effect suggesting that it is an updating effect rather than an ordering effect.

In a recent study, Horne and colleagues found that people’s judgments about the Trolley dilemma are not affected when participants consider emotionally neutral, non-moral dilemmas beforehand (Horne et al., 2013). Unsurprisingly, the stimulus presented prior to the Trolley dilemma must be relevant in some fashion to the Trolley Case in order to produce the effect. But what features of the Footbridge and Trolley Cases are relevant? Subsequent research has revealed that negatively-valenced moral vignettes that do not involve a trade-off of lives (e.g., just pushing a man to his death) or that describe completely different but emotionally salient moral transgressions (e.g., stories about consensual incest) do not affect people’s judgments about the Trolley dilemma (Wiegmann & Waldmann, 2014). And in any event, effects like the Footbridge-Trolley Effect occur for cases in epistemology and philosophy of mind where emotion is unlikely to be salient. Hence, emotion-involving features of the Footbridge and Trolley Cases are probably not the relevant ones (for further discussion of this, see Horne, Powell & Spino, 2013; and chapter 4 and 5 of this dissertation).

We think that the relevant features are the relational structures of the cases involved. Say that two stories are relationally similar when the agents and objects stand in many of the same abstract relations in each story. For example, consider a story in which a magical hawk gives
feathers to a hunter so the hunter can make arrows. The hawk does this in exchange for the
hunter agreeing not to shoot the hawk. A story about Hungary giving missile guidance systems to
Russia in exchange for Russia agreeing not to attack Hungary is relationally similar to the story
of the magical hawk. By contrast, say that two stories are *superficially similar* when the agents
and objects in each story have similar non-relational features (e.g., Hummel & Holyoak, 2002).
For example, a story about an eagle that flies near an archery tournament is superficially similar
to the magical hawk story (Gentner et al., 1993). We claim that in order for the Footbridge-
Trolley Effect to occur, the thought experiments need to be relationally similar, not just
superficially similar.

Psychological research suggests that considering moral dilemmas relationally similar to
the Footbridge Case also alters people’s judgments about the Trolley Case (Petronovich et al.,
1996; Wiegmann et al., 2012). For example, participants who first see the *Transplant Dilemma*,
in which a physician (without obtaining consent) kills a healthy person in order to save the lives
of five other people who need organ transplants, tend to give lower ratings to the claim that
flipping the switch in the Trolley Case is permissible than do participants who do not see the
Transplant Dilemma beforehand (Petronovich et al., 1996). The finding that people change their
judgments about the Trolley Case after considering moral dilemmas that are relationally similar
to the Footbridge Case is reminiscent of results in research on analogical reasoning (e.g., Gick &

Why does relational similarity (but not superficial similarity) between the thought
experiments cause participants to alter their judgments about the Trolley Case after considering
the Footbridge Case? Prior research on analogical reasoning and relational matching offers some
guidance for answering this question. Psychologists that have investigated analogical inference
have shown that people solve “target” problems (in this case, the Trolley Case) on the basis of similar “source” problems (in this case, the Footbridge Case) because both problems lead people to recruit the same schema (e.g., Gick & Holyoak, 1983).\(^1\) The term *schema* as it is used in psychological research on relational reasoning is roughly a sentence-like formula involving one or more relations, where the relata are free variables. For example, “If \(x\) loves \(y\), and \(y\) loves \(z\), then \(x\) will be jealous of \(z\),” is a simple schema.\(^2\) We conjecture that updating effects only occur when people recruit the same schema (or similar schemas) across a series of relationally similar thought experiments.\(^3\) Recruiting the same schema leads people to update their credences with respect to some propositions related to that schema. The basic idea behind the recruitment model is that thought experiments prompt people to consider and revise some background beliefs that they hold.\(^4\) For example, people might recruit the belief that it is never permissible to kill or the belief that it is obligatory to save the most lives in life-or-death situations.\(^5\)

If the effect in the Footbridge-Trolley Experiment is an updating effect, then thought experiments sometimes lead people to update their beliefs. In this respect (at least), we think that thought experiments function as *evidence* in more or less the same way that everyday objects function as evidence. Suzy’s finger prints, found at the scene of a murder, are strong evidence

---

\(^1\) One point of interest is that if researchers have participants’ retrieve a source story from long term memory, and both the target and the source stories are particularly long, stories with superficial similarity are more likely to influence people’s judgments about the target story than stories that are structurally similar to each other. However, in the empirical studies of philosophical ordering effects, the thought experiments presented to people are both short and presented without delay between each thought experiment (see Hogarth and Einhorn’s meta-analysis for why length of vignette hinders ordering effects).

\(^2\) Hummel (personal communication) tells us: “There is no precise definition [of a schema]. Generally, it is a structured (i.e., explicitly relational; formally more powerful than a simple feature list) representation that summarizes the core/common properties of some domain of knowledge, especially the most important relations (including higher-order relations) governing that domain.” See Table 3 in Hummel and Holyoak (2003) for a short list of functional properties that schemas have.

\(^3\) See Horne et al. (2013) for a similar account.

\(^4\) However, at this point it is unknown whether people update their schemas explicitly or implicitly.

\(^5\) In typical cases, what we are calling the background belief will actually be a network of related beliefs.
that Suzy was there and weak evidence that she committed the crime. Similarly, Rachels’ (1975) bathtub thought experiment is evidence that allowing-to-die is sometimes just as morally wrong as actively-killing. Gettier’s (1963) thought experiments are evidence that knowledge is not equivalent to justified true belief. McDermott’s (1995) electrocuted professor thought experiment is evidence that causation is not simple counterfactual dependence. And so on.

In some cases, the evidence of thought experiments might lead one to form true beliefs, and in some cases, the evidence of thought experiments might lead one to form false beliefs. In some cases, the evidence of thought experiments is positively misleading. But then, all of those things are true of ordinary pieces of evidence. DNA evidence is typically very good evidence. If it was planted, though, it might be misleading. Thought experiment vignettes get used evidentially in the same ways as everyday objects, too. We present them. We put them “on the table.” We argue about their evidential significance. We use them to rule out some hypotheses and to confirm others.

Moreover, there is growing experimental evidence that ordinary people respond to thought experiments as pieces of evidence. Horne and colleagues (2013) assigned participants to read one of three thought experiments: a Control Case, the Trolley Case, or the Footbridge Case. After making a judgment about one of these thought experiments, participants were asked to rate the extent to which they agreed with the following statement on a 7-point Likert scale, anchored at (1) strongly disagree and (7) strongly agree:

(A) In the context of life or death situations, one should take whatever means necessary to save the most lives.

---

24 The Control Case was the Coupon Case: the 9th non-moral dilemma from the supplementary materials of Greene et al. (2001).
Their experimental design is pictured schematically in Table 10 below:

<table>
<thead>
<tr>
<th>Control Condition</th>
<th>Rate Agreement with (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Case</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Footbridge Condition</th>
<th>Rate Agreement with (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Footbridge Case</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trolley Condition</th>
<th>Rate Agreement with (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trolley Case</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Experiment 1 from Horne and colleagues (forthcoming)

Horne and colleagues found that participants tended to strongly agree with (A) when they were first asked to evaluate the Control Case. Participants who were asked to make a judgment about the Trolley Case prior to considering (A) tended to agree with (A) but to do so less strongly than participants who made a judgment about the Control Case. And participants who were asked to make a judgment about the right action in the Footbridge Case prior to considering (A) tended to neither agree nor disagree with (A).\textsuperscript{25} These results suggest that the Footbridge Case functions as

\textsuperscript{25} In Condition 2, the mean response was M=5.1, SD= 1.56. In Condition 3, the mean response was M=4.1, SD= 1.52.
disconfirming evidence with respect to (A). Moreover, the results suggest that the Footbridge Case is treated as *stronger* disconfirming evidence.

Horne and colleagues’ results suggest that people revise their beliefs after considering a thought experiment, and their data support construing the effect in the Footbridge-Trolley Experiment as an updating effect. Nonetheless, one might worry that their findings are the result of a demand characteristic created by their experimental design. According to this hypothesis, people only changed their rating of agreement with (A) because it was made salient by the experimental set-up. Participants would not have changed their real agreement with (A) in less artificial conditions. Horne and colleagues sought to rule out this hypothesis by showing that two predictions that it entails are not empirically borne out. First, the demand-characteristic hypothesis suggests that people’s updated ratings in (A) should not be stable over time because they are changed only in the context of the experiment. Second, the demand-characteristic hypothesis suggests that people will not recruit and update (A) without explicit prompting. If either of these predictions were borne out, then the claim that considering moral dilemmas causes people to update their moral beliefs would be called into question.

Horne, Powell, and Hummel (2015) conducted two experiments in order to provide evidence that their initial results were not the product of demand characteristics. First, they tested whether people’s credence in (A) would decay after updating. They predicted that if participants are in fact updating their moral beliefs when they make judgments about moral dilemmas, then their ratings should remain stable over non-trivial intervals of time. They found that participants’ updated credences in (A) were stable over time, even when participants were unaware what beliefs the experimenters were interested in measuring (Horne, Powell, & Hummel, 2015).
Second, they also tested whether people would recruit and update (A) without immediate prompting. They found that people still had an updated credence in (A) six hours after making a judgment about a moral dilemma. The fact that the effect of thinking about a moral dilemma has lasting impact on many people’s credences suggests that people recruit moral beliefs when making judgments about moral dilemmas, even when they are not immediately queried to do so. Taken together, these two experiments make it unlikely that their original result was due to a demand characteristic, instead suggesting that people are updating their beliefs after considering a thought experiment (see Holyoak & Simon, 1999 for more evidence that people update their beliefs and preferences after considering analogous hypothetical scenarios).

The effect in the Footbridge-Trolley Experiment looks like it produces an updating effect, we have an explanatory account consistent with a wide range of psychological data on the assumption that the Footbridge-Trolley Experiment produces an updating effect, and there is no alternative account that explains the available experimental evidence. Hence, we conclude that the Footbridge-Trolley Effect is an updating effect. We conjecture that experiments designed analogously to the Footbridge-Trolley Experiment (e.g., Feltz & Cokely, 2011; Liao et al., 2011; Petronovich et al., 1996; Sinnott-Armstrong et al., 2008; Swain et al., 2007; Tobia et al., 2012) likewise exhibit updating effects rather than genuine ordering effects. Indeed, the design of these experiments is quite suggestive. Consider Feltz & Cokely’s (2011) reported “ordering effect” on intuitions about a CEO intentionally helping or harming the environment.
<table>
<thead>
<tr>
<th>Condition 1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Harm CEO Case</td>
<td>Help CEO Case</td>
</tr>
<tr>
<td>Mean = 6.01</td>
<td></td>
<td>Mean = 3.59</td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
<td>Time 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Help CEO Case</td>
<td>Harm CEO Case</td>
</tr>
<tr>
<td>Mean = 2.54</td>
<td></td>
<td>Mean = 3.90</td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
<td>Time 2</td>
</tr>
</tbody>
</table>

Table 11: Feltz & Cokely (2011)

The effect elicited here does not have the necessary structure to be an ordering effect because participants have not been presented with all the same stimuli when they make their judgments about the Harm or Help version of the vignette. In this case, participants make different judgments about the Harm case because they have different information – they have considered the Help case. Consequently, it is likely that participants have learned something about the connection between a side-effect and intentional action and tempered their judgments about the CEO intentionally harming the environment.

To conclude, it seems that the empirical observation made by many experimental philosophers is that intuitive judgments exhibit updating effects, not that they exhibit genuine
ordering effects. Since updating effects are exhibited in ordinary instances of learning, it is unclear why they should pose any significant threat to the reliability of intuitive judgments.

3. Ordering Effects and the Specter of Global Skepticism

So far we have argued that the “ordering effects” that most experimental philosophers have reported are in fact updating effects, which are not prima facie vicious. And therefore, we have charged that the argument from ordering effects is either invalid or unsound. However, one might try to rehabilitate the argument from ordering effects by accepting a more plausible version of the normative principle—on which exhibiting a genuine ordering effect is prima facie evidence of unreliability—and then running experiments to show that intuitive judgments are affected by genuine ordering effects. The following Rehabilitated Argument from Ordering Effects is one way that such a project might go:

[Revised Normative Principle] If judgments about what some thought experiments show exhibit genuine ordering effects, then the judgments about what those thought experiments show are unreliable.

[Empirical Conjecture] Intuitive judgments about what some thought experiments show exhibit genuine ordering effects.

[Revised Conclusion] Intuitive judgments about what those thought experiments show are unreliable.

We are happy to grant the empirical conjecture. We would be very surprised (for reasons that we hope will soon be obvious) to find out that it is false, and there are some experiments reported in Schwitzgebel and Cushman (2012) suggesting that intuitive judgments about thought experiments do exhibit genuine ordering effects. However, fans of the old argument from
ordering effects should take no comfort in the lodgings of the rehabilitated argument, for the specter of global skepticism haunts the premises. As we will point out in the rest of this section, psychological research shows that perception, memory, deductive reasoning, and testimony are all liable to exhibit genuine ordering effects in some circumstances. Hence, one may offer the following expanded normative principle along with a corresponding empirical claim:

[Expanded Normative Principle] If the judgments delivered by some cognitive process exhibit genuine ordering effects in some cases, then the judgments delivered in those cases are unreliable.

[Expanded Empirical Claim] Perceptual, memorial, logical, and testimonial judgments exhibit genuine ordering effects in some cases.

[Expanded Conclusion] Perceptual, memorial, logical, and testimonial judgments are unreliable in those cases.

We see no reason to state the normative principle in terms of thought experiments, rather than in terms of instances or cases of the deliverance of a judgment by some cognitive process. But now, if one wants to criticize intuitive judgments (or the cognitive processes that produce intuitive judgments) by generalizing the conclusion of the rehabilitated argument from ordering effects, one must be prepared to generalize the expanded argument to the same degree. And we fear that the result of generalizing in the expanded case is global or nearly global skepticism. Our argument is similar in spirit to Williamson’s (2004, 2008, forthcoming) arguments that critiques of intuitive judgment that have been launched by experimental philosophers lead to skepticism about judgment simpliciter, which quickly spirals out of control. The main differences between Williamson’s arguments and ours, as we see things, are (1) that although we couch our
arguments in terms of judgments, they need not be framed that way but could be stated more
generically in terms of the products or deliverances of cognitive processes, and (2) that since
there is compelling psychological evidence that genuine ordering effects appear throughout
human cognition (as we show below), if one endorses a normative principle according to which
genuine ordering effects are prima facie vicious, the threat of skepticism is not idle.

3.1 Ordering Effects and Perception

Weiss and Anderson (1969) investigated the Averaging Model of visual estimation by having
participants look at a series of lines of different length and then estimate the average length of
the lines they looked at after considering the entire series. Participants were presented with 6
lines (either 16cm or 24cm in length) in random order in a single block and performed the
average estimation task for 11 blocks. Weiss and Anderson compared participants’ performance
between blocks with the same average line length but differing presentation order.

Participants’ estimations were significantly influenced by the order in which they
considered either 16cm or 24cm lines. Specifically, the line presented at the end of the series had
a larger effect on participants’ judgments than lines at the beginning of the series. This effect is
consistent with research showing that average loudness estimations are affected by the order in
which participants consider audio samples (e.g., Parducci, 1965). Likewise, in a now classic
paper on psychophysics, Stevens and Galanter (1957) demonstrate ordering effects in a dozen
perceptual phenomena including line length, visual numerosness, visual area, weight, loudness,
brightness, visual inclination, proportion, and pitch. Ordering effects have also been shown for
time-duration perception (Allan, 1979; Jamieson & Petrusic, 1975), product taste tests (Dean,
1980; Scarpi, 2004), and even aesthetic judgments (Englund & Hellstrom, 2012).
3.2 Ordering Effects and Memory

A very large body of research shows that people are better at recognizing and recalling the first and last items in a list (e.g., Anderson et al., 1998; Feigenbaum & Simon, 1962; Haarmann & Usher, 2001; Howard & Kahana, 1999; Murdock, 1962; Roediger & Crowder, 1976). For example, researchers ask participants to read a list of words that are presented sequentially. After a distractor, they ask participants to indicate which words appeared in the list. Psychologists have found that participants remember the first and last items better than any other items in the list. This is an ordering effect since participants in every condition are presented with all the same words, and in different presentations, the same words are recalled at different rates.

3.3 Ordering Effects and Reason

Ashton and Ashton (1988) investigated how professional auditors’ beliefs change when they consider evidence that a computer program will detect a payroll error. In this experiment, participants are asked to assume they are investigating the payroll records of a client. Participants are presented with four pieces of evidence in one of two orders. In condition 1, participants are first presented with two pieces of evidence that suggest that internal checks will detect errors in the payroll and then two pieces of evidence that suggest that internal checks will not detect errors in the payroll. In condition 2, participants are first presented with two pieces of evidence that suggest that internal checks will not detect errors in the payroll and then two pieces of evidence that suggest that internal checks will detect errors in the payroll. In both conditions, the total evidence participants assess is identical. After considering each piece of evidence, participants are asked to rate the likelihood that internal checks will detect errors in the payroll.
The results show an ordering effect wherein participants’ judgments about how likely internal checks are to detect errors in the payroll are primarily determined by the first pieces of evidence participants consider. This result suggests that even when people are familiar with the kind of evidence under consideration (e.g., professional auditors assessing payroll data), their assessments can still be altered by the order in which they consider evidence. More recent psychological research suggests that ordering affects even paradigmatic instances of deductive reasoning. Oberauer and colleagues (2005) report that altering the order in which participants see premises in syllogistic arguments significantly affects their judgments of the validity of these arguments, replicating an earlier effect of premise-ordering that Girotto and colleagues (1997) report.

3.4 Ordering Effects and Testimony

Cromwell (1950) investigated how listening members of an audience judged the persuasiveness of speeches presented when they hear the speeches in different orders. Participants sat in an audience and listened to argumentative speeches about labor and medicine in one of two orders: Affirmative – Negative or Negative – Affirmative. After hearing these two speeches, participants had to rate the extent to which they agreed with the proposition at issue in both speeches. This experiment revealed that participants’ attitudes changed depending on the order in which they heard the speeches. Specifically, Cromwell found that the second speech in the sequence had a larger effect on participants’ judgments than the first speech they evaluated. Cromwell then investigated whether presenting two speeches in different orders, both arguing for the same position but differing in strength, would affect participants’ agreement with the proposition at issue. Consistent with his original finding, Cromwell found that presenting the stronger argument
second led participants to agree more with the proposition at issue than presenting the weaker argument second.

Subsequent research has confirmed and expanded on Cromwell’s findings, for example by showing that jurors’ verdicts in criminal trials will change when they are presented with testimonial evidence in different orders (e.g., Pennington, 1982; Walker et al., 1972).

4. Reflecting on the Argument from Ordering Effects

Perception, reason, memory, and testimony all exhibit genuine ordering effects. Moreover, in the cases we surveyed in Section 3 above, perception, reason, memory, and testimony appear to be unreliable. Should we therefore be skeptics about perception, reason, memory, and testimony? We think not. But how should we resist the skeptical conclusion? In this section, we consider two possibilities: (1) reject the generalization step on the grounds that experiments that have been conducted so far form an inadequate inductive basis; or (2) reject the revised normative principle on the grounds that the fact that a cognitive process exhibits a genuine ordering effect is not prima facie reason for thinking that the process is unreliable. But before taking up either strategy, we want to reject an alternative empirical hypothesis.

One might hope to be able to accept the rehabilitated argument from ordering effects while avoiding skepticism by arguing that intuitive judgments are importantly different from perception, reason, memory, and the like because intuitive judgments exhibit ordering effects to a greater degree or with greater frequency than other basic sources of justification. Whether this differential exhibition response succeeds or fails depends crucially on whether or not intuitive judgments really do exhibit stronger ordering effects or whether they are in fact exhibited more frequently, which is an open empirical question. To give real teeth to the differential exhibition
response, one would need to know (at minimum) how often each of our cognitive faculties is sensitive to ordering. What is the relative frequency with which intuitive judgments are sensitive to the ordering of information? And how does that relative frequency compare to the same relative frequency for perception? At best, these are enormously fraught issues that require large-scale meta-analyses. Clearly, a few isolated studies are not adequate for answering them.

But in any event, the current evidence does not look good for the differential exhibition response. Counterbalancing is *standard practice* in psychology because it is widely believed that ordering effects are the norm, rather than the exception. In other words, ordering effects are widespread in human cognition as such, not just in (intuitive) cognition about thought experiments. In fact, in a large meta-analysis of psychological studies on belief updating, Hogarth and Einhorn (1992) found that ordering effects occurred in 69 out of 74 belief-updating experiments (i.e. nearly 95% of the time). Consequently, some psychologists have gone so far as to claim that because belief updating occurs sequentially and the psychological mechanisms at play treat different sequences of the same elements differently, nearly all processes that involve belief updating should be affected by ordering (e.g., Anderson, 1971; Hogarth & Einhorn, 1992; Newell and Simon, 1972). Hence, we think it is unlikely that intuitive judgments will turn out to exhibit ordering effects in a way that sets them apart from other cognitive processes.

However the arguments from ordering effects are to be resisted, they have to be resisted in the same basic way for intuitive judgments as they are resisted for perception, reason, memory, and testimony. Although one might maintain that the empirical premise of the argument has not been established beyond reasonable doubt for intuitive judgments, we are confident that it is true. Hence, if a skeptical conclusion is to be avoided, we need to reject either the expanded normative principle or the generalization step. Moreover, we think that rejecting
the expanded normative principle is plausible if and only if rejecting the revised normative principle is plausible. In Section 4.1, we consider the possibility that the normative principle is faulty. In Section 4.2, we consider the generalization step.

4.1 Are Ordering Effects Prima Facie Vicious?

Given how much hangs on it, we find it surprising that no one, as far as we know, has given an argument that anything like the revised normative principle is correct. Rather, most philosophers seem to think that some principle in the neighborhood of the normative principle or the revised normative principle is obviously true on its face. In what follows, we want to explore the possibility that exhibiting a genuine ordering effect is not necessarily a mark of unreliability. In the remainder of section 4.1, we accomplish two things. First, we defend the claim that genuine ordering effects are not always vicious. Our goal here is minimal: to make space for the thought that sensitivity to ordering may sometimes be due to the proper functioning of a normatively justified belief updating procedure. Second, we suggest that there is a deeper epistemological principle at stake in debates about ordering effects, and we argue that once one notices the deeper principle, experimentally driven arguments from ordering effects face a difficult dilemma.

A brief tour of a debate in formal epistemology about Jeffrey Conditionalization will be instructive. Jeffrey (1983) offers his brand of conditionalization in order to give an account of how an epistemic agent ought to update her beliefs on the basis of uncertain evidence—something beyond the scope of ordinary Bayesian conditionalization.26 Jeffrey observes in

26 Whereas ordinary Bayesian conditionalization is given by the equation $C_{\text{new}}(p) = C_{\text{old}}(p \mid e)$, the simplest case of Jeffrey conditionalization is given by the equation $C_{\text{new}}(p) = C_{\text{old}}(p \mid e) \cdot C_{\text{new}}(e) + C_{\text{old}}(p \mid \sim e) \cdot C_{\text{new}}(\sim e)$ to account for the fact that we might not be certain about our evidence $e$ for the proposition $p$. 

145
Section 11-11 of *The Logic of Decision* (182-183) that his brand of conditionalization is not commutative. Here is an illustrative example. Sarah is trying to figure out who robbed a local jewelry store. There are two possible suspects: Gertrude and Ichabod. Before consulting the videos, Sarah thinks that Gertrude and Ichabod are equally likely to be the culprit. Sarah has two grainy surveillance videos of the robbery. Call the first video $A$ and the second video $B$. Sarah thinks that if the culprit on the videos is a woman, it must be Gertrude, and she thinks that if the culprit on the videos is a man, it must be Ichabod. Let $a$ stand for the credence Sarah has after watching video $A$ that the culprit is a woman, and let $b$ stand for the credence she has after watching video $B$ that the culprit is a woman. After watching video $A$, Sarah revises her belief that the culprit is a woman up to $a = 0.53$. But after watching video $B$, she revises her belief that the culprit is a woman down to $b = 0.3$. After both revisions, Sarah has credence of 0.3 that Gertrude is the culprit. The second revision totally screens off the first revision. Hence, Jeffrey Conditionalization is not generally commutative in the following sense: If the numbers were reversed so that after the first revision, Sarah were to have credence of 0.3 that the culprit is a woman, and after the second revision, she were to have credence of 0.53 that the culprit is a woman, then her final credence would be 0.53 that Gertrude is the culprit. Since 0.3 is not equal to 0.53, Jeffrey Conditionalization is not commutative.

Several philosophers have argued that Jeffrey Conditionalization is defective in virtue of its failure to be commutative (e.g., Field, 1978; Domotor, 1980; van Fraassen, 1989; and Döring, 1999). The complaint raised by these philosophers looks very similar to the worry motivating the argument from ordering effects. For example, Döring (1999) writes:

Jeffrey conditionalization is sensitive to the order in which the evidence comes in. The probability function that results from first assigning some probability value $r_1$ to $e_1$ (and thus $1 - r_1$ to $\neg e_1$) and then probability $r_2$ to $e_2$ is not in general the same as the function
that results from first assigning \( r_2 \) to \( e_2 \) and subsequently assigning \( r_1 \) to \( e_1 \). … The difference will consist, presumably, in some reference to the order in which the evidence arrived. But intuitively, I think, this is a difference that should not make a difference. (S382-S384)

However, other philosophers, like Jeffrey himself, have argued that Jeffrey Conditionalization’s failure of commutativity is unproblematic.

Rosenkrantz (1981, 3.6-2) offers the following simple example to illustrate how the order in which events occur might make an evidential difference. Suppose a child clumsily spills a jar of paint. The child knows that her parents will be unhappy about the spill, but she doesn’t know whether or not she will be spanked. We might imagine two cases. In the first case, the child looks up to her father and observes an angry frown followed by a warm smile. In the second case, the child looks up to her father and observes a warm smile followed by an angry frown. We contend that the final opinion of the child with respect to the proposition, “I will be spanked,” should be different in the two cases. Many further cases could be given. To take just one simple example, suppose that you live in a country where traffic lights go from green to yellow to red and then back from red to yellow and then to green. A yellow light preceded by a green light indicates that one should prepare to stop, but a yellow light preceded by a red light indicates that one should prepare to go. Now suppose that while driving one day, you approach an intersection and see all three colors—red, yellow, and green—displayed in some order. If you want to know whether it is permissible to drive through the intersection, you need to know the order in which the lights were displayed. The order in which you see the colors should lead you to different beliefs about whether it is permissible to continue into the intersection. If you observe red-yellow-green, then you may proceed. But if you observe green-yellow-red, then you may not. Rosenkrantz remarks that the order in which one receives evidence clearly matters “in the special
case where two successive inputs both bear on the same proposition,” which is exactly what we think happens in cases where we observe genuine ordering effects. Hence, one should at least be open to the possibility that being sensitive to ordering with respect to thought experiments is actually appropriate.

We think that at this point, we have made space for the possibility that even genuine ordering effects are not necessarily counter-normative. But there is a further wrinkle worth considering. Lange (2000) argues that in Rosenkrantz’s example, the child would not be receiving the same pieces of evidence in different orders in the two cases. Rather, the child would actually have different evidence in the two cases. He suggests that in one ordering in Rosenkrantz’s case, the second piece of evidence is not simply an angry frown but that a warm smile turned into an angry frown; whereas, in the other ordering, the second piece of evidence is that an angry frown turned into a warm smile.27 “Intuitively,” Lange writes (396), “there is no reason why these two sequences should be treated in the same way, since they do not involve learning the same data merely in reverse order.” In this way, Lange preserves the idea that the order of one’s evidence should not matter without needing to give up the formal non-commutativity of Jeffrey Conditionalization.

Up until now, we have been treating thought experiment vignettes themselves as the pieces of evidence that people consider when they participate in experiments like the Footbridge-Trolley Experiment (though see footnote 7). But at this point, the nature of participants’ evidence

27 Although we will not focus on it here, the way that Lange describes the evidence also problematizes, we think, how to count the number of pieces of evidence one has. When a child sees her mother laugh and then scowl, does she see and experience a single thing (a laugh-to-scowl) or two things (a laugh and a scowl)? If she experiences two things, does she also experience temporal succession or is each event indexed by time? And if the time index matters to the nature of her experience, isn’t that just to sneak in the supposedly irrelevant information about the order of her pieces of evidence? We have no answers to these questions at present, but surely, anyone interested in defending an order-invariance principle must answer them.
becomes crucial, for there are two ways of formulating the revised normative principle, and it is not clear that the formulations are equivalent. The first formulation is exactly as stated in the revised argument from ordering effects:

[Revised Normative Principle] If judgments about what some thought experiments show exhibit genuine ordering effects, then the judgments about what those thought experiments show are unreliable.

This way of stating the principle is explicit in saying that judgments should be invariant with respect to re-arrangements of the thought experiment vignettes. Plausibly, however, the real epistemological principle at stake should be framed in terms of one’s evidence, along the following lines:

[The Evidential Ordering Principle] Suppose $S_1$ is a sequence of pieces of evidence and $S_2$ is a sequence that differs from $S_1$ only in the order of its elements. If a cognitive process is epistemically reliable, then it should produce the same judgments on the basis of either $S_1$ or $S_2$.

If the vignettes presented to some participant are not themselves the pieces of evidence that the participant has, then the revised normative principle and the evidential ordering principle come apart. Two people might be presented with the same thought experiments in different orders but not thereby be presented with the same pieces of evidence in different orders.

In line with Lange, suppose that an agent’s experience of some vignettes is affected by the order in which those vignettes are presented. If an experience (or its content), rather than a vignette (or some facts about it), is the evidence that an agent has, then the presence of an ordering effect does not necessarily indicate unreliability, even if one thinks that the order of one’s evidence is evidentially irrelevant. Recall that what one actually observes in experiments where genuine ordering effects occur is that the mean of some reports made by participants in
response to a question asked at the end of a sequence of vignettes depends on the order in which the vignettes appear in the sequence. If the evidence that an epistemic agent has consists in the experiences of the agent, then the fact that the agent is sensitive to ordering with respect to the vignettes is not by itself reason to think that the agent exhibits ordering effects with respect to the evidence. We have no idea what the experiences of the participants are like in experiments designed to measure sensitivity to ordering. Maybe the experiences are simply permuted along with the vignettes. But we think it is at least as likely that the quality of the experiences themselves (if such things exist) depends on the order of the vignettes.

Hence, a similar move to the one that Lange makes with respect to Jeffrey Conditionalization is available to critics of the argument from ordering effects in all its forms. If the pieces of evidence that a participant has after considering a sequence of thought experiments is distinct from the thought experiment vignettes themselves—e.g., by consisting in some experiences—then the critic of the argument from ordering effects may retain the principle that what one believes on the basis of his or her evidence should be insensitive to the order in which the pieces that make up that evidence were acquired and yet reject the claim that the skeptical conclusion follows from the principle and the experimental evidence.

Summing up, we think that the revised normative principle is faulty because even genuine ordering effects are not necessarily prima facie vicious. One might endorse the evidential ordering principle, which is (perhaps) still plausible given what we have said so far. But in order to derive skeptical conclusions from experimental results showing that people exhibit a genuine ordering effect for a pair of vignettes on the basis of the evidential ordering

---

28 And that same response problematizes conclusions drawn from the psychological research we surveyed in Section 3 as well. Thanks to Conor Mayo-Wilson for pointing this out.
principle, one would need to endorse an account of evidence on which the evidence an experimental participant has consists in the vignettes themselves, as opposed to the way the participant’s sensory surfaces are stimulated (Quine, 1969) or what the participant is thinking about at the time (Feldman 1988) or an experience the participant has (Lewis 1996) or some beliefs the participant has (Davidson 2001) or some propositions the participant knows (Williamson 2000) or some propositions specially related to how the participant regulates his or her attitudes (Neta 2008).

4.2 Can We Generalize from the Argument?

Resisting the argument from ordering effects by denying that genuine ordering effects are prima facie vicious is, we think, a bold strategy. But perhaps it is too bold. After all, the various examples of genuine ordering effects that we have surveyed from the psychology literature all seem to be cases where our cognition really does go wrong. Consequently, one might think that we should accept some normative principle according to which genuine ordering effects are genuinely bad and resist the skeptical conclusion by rejecting the generalization step, instead. One might agree that all of our cognitive faculties are prone to error in some specific circumstances but deny that such errors are any threat to the reliability of our faculties in general. At the most extreme, one might maintain that skepticism is only warranted in cases where genuine ordering effects have actually been demonstrated to occur. Our cognition is to be regarded as innocent until proven guilty. Such an approach transforms negative experimental critiques based on ordering into local criticisms of specific arguments founded on specific thought experiments, rather than a sweeping methodological point. Alternatively, one might try to characterize some features of the thought experiments that encourage sensitivity to ordering and then limit the extent of the generalization step accordingly. Perhaps vignettes involving
norms are more likely than vignettes not involving norms to elicit genuine ordering effects. Such an approach would make the experimental critique broader and more interesting.

We think that neither strategy is very promising. As we have already noted, sensitivity to ordering owes to a very general feature of human psychology. Thus, it is likely that judgments about every thought experiment—in fact, judgments about every stimulus—will exhibit some sensitivity to ordering in the right experimental conditions. Similarly, sensitivity to ordering has little to do with intrinsic features of the stimuli used in an experiment, and more to do with how they are related to one another and to an underlying task or proposition. For instance, in a now classic paper, Ross (1987) argues that people’s tendency to solve subsequent problems on the basis of prior problems, a tendency which can present itself as an ordering effect, is due to people’s reliance on analogical reasoning. Penn et al. (2006) argue that our ability to reason analogically, which requires the ability to explicitly represent relations symbolically (e.g., Doumas, Hummel & Sandhoffer, 2008; Hummel & Holyoak, 2003), is the feature that distinguishes human cognition from the cognition of non-human primates. If Penn et al. are right (and we are sympathetic to their arguments), then sensitivity to ordering is essential to distinctively human cognition. The generalization step is thus very hard to resist given what we know about how human psychology works. Put another way, sensitivity to ordering is different from errors of cognition owing to failures of heuristics. Whereas heuristics that are extremely useful (e.g., the Availability Heuristic) have been known to fail under specific conditions (e.g., Linda the bank teller), no particular context is required to elicit an ordering effect. Rather, ordering effects arise in virtue of the way the human cognitive system works.

Summing up, resisting the argument at the generalization step looks like a losing strategy. Given what we know about human psychology, we cannot avoid generalizing here if the rest of
the argument goes through. Therefore, we think the bold solution is also the only available one.

If we are to reject the argument from ordering effects, we need to reject whatever normative principle is alleged to make it work.

5. Conclusion

In this chapter, we have called into question two ways to argue against the reliability of intuitive judgments on the basis of their sensitivity to ordering. We began with the observation that the normative principle in the argument from ordering effects is ambiguous between updating effects and genuine ordering effects. We argued that updating effects are not prima facie vicious, and we argued that much of the experimental evidence currently on offer establishes only that intuitive judgments exhibit only these non-threatening updating effects. We then pointed out that even if we grant the empirical conjecture that intuitive judgments exhibit genuine ordering effects, the natural rehabilitation of the argument from ordering effects proves too much, leading to a view practically indistinguishable from global skepticism. One might respond to our arguments by endorsing global skepticism. “Contemporary psychology,” the skeptic might say, “like Hume’s empiricism, shows that the claim that we can have knowledge through our ordinary cognitive faculties is ultimately self-defeating.” But we are not ready to join the skeptic.

NOTE: The proceeding chapter was coauthored with Jonathan Livengood.
Conclusion

In the preceding chapters of this work, I have shown the relevance of empirical research to substantive philosophical questions in epistemology and moral psychology. In Chapters 1 and 2, I showed that criticisms of experimental epistemology pose serious problems for the interpretation of data from survey studies, but that there are other empirical methods that do not succumb to the criticisms that plague surveys. However, I also demonstrated that when care was taken to rule out alternative interpretations for effects found based on survey tasks, these data can also have important implications for traditional epistemological debates (Chapter 3). In Chapters 4 and 5, I demonstrated that understanding the psychological processes involved in moral judgment can help adjudicate between competing moral psychological theories (e.g., Greene et al., 2001; Prinz, 2007), and allow us to understand the role of thought experiments in ethical theorizing. In the final Chapter, I summarized the results from several lines of research in moral psychology to show that when we understand the psychological data on moral judgment, specifically moral judgments about moral dilemmas, we can get traction on questions concerning the reliability of moral judgments and moral intuitions.
References

Chapter 2


Powell, D., Horne, Z., Pinillos, A., & Holyoak, K. J. (Submitted). Justified true belief triggers false recall of "knowing."


**Chapter 3**


Chapter 4

Chapter 5


Chapter 6


**Chapter 7**


