FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

REFLECTIVE REASONING FOR REAL PEOPLE:

CLARIFYING THE ROLE OF REFLECTION IN PHILOSOPHY, SCIENCE, AND BIAS

By

NICK BYRD

A Dissertation submitted to the
Department of Philosophy
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

Nick Byrd defended this dissertation on May 29, 2020.
The members of the supervisory committee were:

John Schwenkler
Professor Directing Dissertation

K. Anders Ericsson
University Representative

Michael Bishop
Committee Member

James “Jack” Justus
Committee Member

The Graduate School has verified and approved the above-named committee members and certifies that the dissertation has been approved in accordance with university requirements.

For the philosophers who put up with my scientific tendencies
and the scientists who put up with my philosophical tendencies.

ACKNOWLEDGMENTS

People: Mark Alfano, Adam Arico, Lieke Asma, Jacob Berger, Istvan Berkeley, John Bickle, Mike Bishop, Gunnar Björnsson, Cameron Buckner, Hannah Byrd, Amber Cazell, Dave Chalmers, Cory Clark, Stephen Clarke, Paul Conway, Daniel Coren, Josh Correll, Mike Dacey, Gabriel De Marco, Rodrigo Diaz, Chris Dodsworth, Anders Ericsson, Joe Fraley, Grace Helton, Bryce Huebner, Mike Huemer, Zoe Jenkin, Josh Knobe, Matt Jones, Chick Judd, Jack Justus, Brian Leiter, Jonathan Livengood, Edouard Machery, Eric Mandelbaum, Heather Maranges, Jon Matheson, Michele Merritt, Gordon Pennycook, Valentina Petrolina, Caleb Pickard, Ashby Plant, Ted Poston, Jake Quilty-Dunn, Caleb Reynolds, Luis Rosa, Rob Rupert, John Schwenkler, Eric Schwitzgebel, Nat Stein, Steve Stich, Justin Sytsma, Brian Talbot, Marshall Thompson, Michael Tooley, Jay Van Bavel, Kassidy Velasquez, Jonathan Weinberg, Justin Weinberg, Jake Westfall, Evan Westra, Joe Wilson, Katie Wolsiefer, Mike Zahorec, and anonymous reviewers.

Events. Alabama Philosophical Society conference (2019), American Philosophical Association Eastern conference (2019), Australasian Society for Philosophy and Psychology conference (2018), Buffalo Experimental Philosophy conference (2015), Explaining Religion Workshop (2015), International Association for the Cognitive Science of Religion conference (2018), Midsouth Philosophy Conference (2015), Society for Philosophy and Psychology conference (2019), Southern Epistemology Conference (2018), Southern Society for Philosophy and Psychology conferences (2015, 2017), the Moral and Social Processing Lab meetings.

Funders. University of Colorado at Boulder's Department of Philosophy and Institute of Cognitive Science, Florida State University's Department of Philosophy and Graduate School, the Society of Christian Philosophers via the John Templeton Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

1. Explicating The Concept of Reflection. To understand how 'reflection' is used, I consider ordinary, philosophical, and scientific discourse. I find that 'reflection' seems to refer to reasoning that is deliberate and conscious, but not necessarily self-conscious. Then I offer an empirical explication of reflection's conscious and deliberate features. These explications not only help explain how reflection can be detected; they also distinguish reflection from nearby concepts such as ruminative and reformative reasoning. After this, I find that reflection is not obviously limited to only practical or only theoretical reasoning. The chapter ends with reasons to prefer 'unreflective' and 'reflective' to dual process theorists' labels about systems or types.

2. Bounded Reflectivism & Epistemic Identity. Reflection features centrally in philosophy (e.g., Korsgaard, 1996; Rawls, 1971; Sosa, 1991) and psychology (e.g., Pennycook, 2018). Bounded reflectivism is an empirically adequate model of reflection that explains reflection's capacity to either help or hinder reasoning—e.g., reflective equilibrium vs. reflective polarization (e.g., Kahan et al., 2017). One innovation of the model is epistemic identity: an identity that involves particular beliefs—e.g., religious and political identities. When we feel that our epistemic identity is threatened, we can reflectively defend its beliefs rather than update our beliefs according to the best arguments and evidence. The solution, I argue, is not to suppress epistemic identity but embrace it by appealing to shared, superordinate epistemic identities.

3. All Measures Are Not Created Equal. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? "10 cents" comes to most people's minds immediately. However, upon reflection, we can realize that "5 cents" is the correct answer. There are many reflection tests. Each has its limitations. Some reflection tests are

mathematical—like the bat-and-ball problem (e.g., Frederick, 2005) and others are logical (e.g., Janis & Frick, 1943). So there are concerns that reflection tests track logical and mathematical competence rather than reflection *per se*. Also, some psychologists assume *a priori* that correct answers are reflective and lured answers are unreflective. New evidence raises concerns about the plausibility of these assumptions. I argue that think aloud protocols (Ericsson & Simon, 1998) and process dissociation (Jacoby, 1991) can assuage some of these concerns.

4. Great Minds Do Not Think Alike. Two large studies (N = 1299), one pre-registered, found that many correlations between reflection and philosophical beliefs among non-philosophers replicated among philosophers. For example, less reflective philosophers preferred theism to atheism and utilitarian rather than deontological responses to the trolley problem (Hannikainen & Cova, in prep.; Pennycook et al., 2016; Reynolds, Byrd, & Conway, *forthcoming*). However, philosophical judgments were sometimes better predicted by factors like education, gender, and personality than by reflection test performance. So although some relationships between reflection and philosophy were robust, there is more to the link between reflection and philosophy. Normative implications are also discussed—e.g., how we can infer the quality of philosophical views from their correlations with reflective or unreflective reasoning.

5. What We Can (And Can't) Infer About Implicit Bias (In *Synthese*). Contrary to some philosophers and psychologists, I argue that implicit bias is probably associative. However, I also argue that debiasing is not thereby entirely unconscious and involuntary. Indeed, strong evidence suggests that reflection can change implicitly biased behavior—e.g., conscious and deliberate counterconditioning. This has implications for the science and morality of implicit bias—e.g., evidence suggests that we can reform biases; so. intuitively, we can be responsible for biases.

CHAPTER 1
EXPLICATING THE CONCEPT OF REFLECTION

"The idea that [P] is so intuitive that most will need no more proof than its statement."

—Wenar, 2008

"…we are not forced to act on the desires we happen to find present in ourselves. We have the capacity to take a step back from them and decide whether or not we will endorse them…."

—Korsgaard, 1996

"…it seems reasonable that [P]. …. However, let us reflect."

—Quine, 1951

1.1 What is Reflection?

Some questions prompt an intuitive response. When we ask, "How much should I donate?", our first, intuitive response might be to choose an amount that feels right. However, we might step back and reflect on this intuition. "Should I donate more?" "Must I donate anything?" This reflection can change or confirm our initial intuition. "Upon reflection, I can afford to give more." Or, "Upon reflection, I should donate to another, more effective charity."

Appealing to such intuitions is standard fare in philosophy (Chalmers, 2014; De Cruz, 2014; Kornblith, 1998; Mallon, 2016) and science (Tallant, 2013). Some go as far as to say that

"preoccupation with reflection is, arguably, the Western philosophical tradition's most distinctive feature" (Doris, 2015). Of course, reflection features not only in philosophical theories (e.g., Goodman, 1983; Hursthouse, 1999; Kennett & Fine, 2009; Korsgaard, 1996; Rawls, 1971; Sperber, 1997; Sosa, 1991; Velleman, 1989; 2000; Wallace, 2006), but also in psychological theories (De Neys, 2017; Pennycook, 2018; Pennycook & De Neys, 2019). To understand this literature, we will first need to understand 'reflection'. Alas, philosophers often describe reflection in merely metaphorical or folk psychological terms—e.g., "back up and bring [an] impulse into view" (Korsgaard, 1996). Meanwhile, scientists often describe reflection with various orthogonal concepts—see reviews from Frankish (2010) and Nagel (2014). So if we are going to advance the philosophy and science of reflection, then we need an explication of 'reflection' that is more empirical and precise.

To do this, the present paper examines the use of 'reflection' in ordinary discourse, philosophers' discourse, and its scientists discourse. This investigation will reveal that reflection often involves a couple features: it is deliberate and conscious. Further investigation reveals a distinction between two notions of 'conscious' in discussion of 'reflection': consciousness in general and self-consciousness more specifically. With this two-factor account of reflection and a distinction between general consciousness and self-consciousness, we can begin to parse 'reflection' in terms of dual process theory. The result is an empirically adequate account of reflection that bears on both the philosophy and science of reflection.

1.1.1 Ordinary Language

Ordinary use of 'reflection' reveals a few of its common features. 'Reflection' is used to refer to something that is deliberate, conscious, and sometimes slow.

Deliberate. To begin, consider how 'reflection' is described as contrasting with and inhibiting autonomous thoughts and impulses.

> "Mankind act more from habit than reflection" (Paley, 1785, 26).
>
> "I wish you to pause, reflect, and judge before you decide" (James, 1853, 248).
>
> "You never stop to think—whatever comes into your head to say or do you say or do it without a moment's reflection" (Montgomery, 1908, Chapter 21).

Conscious. Notice also how 'reflection' often implies that we are consciously aware of our reasoning.

> "When I reflect my thought …upon that I have formerly written." (Harington, 1611)
>
> "There is but here and there a man that reflects …and carefully and attentively observes what's doing in his own mind" (Norris, 1704, 121).

Slow. Now consider how 'reflection' seems to require more time than other kinds of reasoning.

> "I have not leisure to reflect, … Or trifle time in thinking" (Congreve, 1797, 36).
>
> "Now just reflect,—meditate for as long time as would soft-boil an egg" (Thompson, 1832, 327).
>
> "I used to have time to think, to reflect, my mind and I" (Keller, 1903, 109).

1.1.2 Philosopher's Language

The use of 'reflection' among philosophers is largely consistent with what we found in the last section. Philosophers talk about reflective reasoning as deliberate, conscious, and slow.

Deliberate. Philosophical accounts of reflection often highlight how reflection is deliberate rather than automatic—sometimes called "autonomous", which will be explained later.

> "The role of reflection is… to step back from the immediate situation, to calculate consequences, to compensate for the immediate force of one desire which might not be the most advantageous to follow…." (Taylor, 1976, 287).

Conscious. Philosophers also highlight the role of conscious representations in reflection.

> "Reflection requires the goal-directed production of a series of …inner speech. … Reflective cognition requires the sequential use of a progression of conscious contents…" (Nagel 2014, 231).

Slow. Philosophers also seem to expect that reflection takes relatively more time than more autonomous and less conscious processes.

> "The supermind is a slow but highly flexible system, which can kick in whenever faster but less flexible basic processes fail to yield a solution. Moreover, because supermental processes are under personal control, we can reflect on them, refine them,…." (Frankish, 2004).

1.1.3 Scientists' Language

Some scientists study reflective reasoning. These scientists also describe reflection as a deliberate, conscious, and often slow phenomenon.

Deliberate. Like ordinary people and philosophers, scientists often contrast reflection with more autonomous processes.

> "…theories and research relating to impulse versus reflection are ubiquitous and prevalent in nearly every sub-discipline" (Deutsch, Gawronski, & Hofmann, 2017).

> "…social cognition and behavior are the outcome of two broad systems of information processing, the reflective system and the impulsive system" (Strack & Deutsch, 2014).

Conscious. Scientists also describe reflection as a process that involves representations of our own reasoning.

> "Analytic cognitive style (the willingness or disposition to critically evaluate outputs from intuitive processing and engage in effortful analytic processing)" (Ross, Pennycook, McKay, Gervais, Langdon, & Coltheart, 2016, p. 300).

> "Decoupling processes enable one to distance oneself from representations of the world so that they can be reflected upon and potentially improved" (Stanovich, 2009).

Slow. Some scientists describe reflection as involving not only deliberate and conscious thought, but also slow thought.

> "Many researchers have emphasized the distinction between two types of cognitive processes: those executed quickly with little conscious deliberation and those that are slower and more reflective" (Frederick 2005, 26).

> "…impulsive subjects complete the [cognitive reflection test] quicker than reflective subjects." (Jiménez, Rodriguez-Lara, Tyran, & Wengstrom, 2017).

1.1.4 Reflection vs. Self-Reflection

Ordinary, philosophical, and scientific uses of 'reflection' revealed that people seem to think of reflection as deliberate, conscious, and relatively slow. But we can do more than just list these features. In addition to these features, there are relevant distinctions. Consider the distinction between reflective reasoning and reflective self-awareness.

Reflection. The first kind of reflection is just a way of reasoning. It is deliberate and consciously represented reasoning (Shea & Frith, 2016). We reason deliberately when we do not simply accept our initial impulses. For instance, when we question our gut reaction to a situation, then we reason deliberately. We reason consciously when we are aware of some of the representations involved in in our thinking. For example, when we are aware of not only the answer to a math problem, but also some of the steps that got us to the answer, we are reasoning consciously.

Unsurprisingly, deliberate and conscious reasoning tends to be slower than more autonomous and unconscious reasoning. Of course, there may be exceptions to this: for instance, a chess expert's reflection about where to move a piece may occur about as fast as their more unreflective intuitions about where to move a piece. Indeed, while there are popular accounts of reflection as slow thinking (e.g., Kahneman, 2011), plenty of evidence suggests that reflective reasoning can be fast (e.g., Bago & De Neys, 2017).

Self-reflection. Note that while reflective reasoning is conscious, it is not necessarily self-conscious. This is because we can consciously reason about things other than our own minds. As John Doris remarks, "'reflection' [can] cover …introspective and extrospective processes." (Doris 2015, 2). For example, when I multiply 13 by 17, I am not consciously representing these numbers *as my* representations of the numbers. I am simply representing the numbers. My so-called self is not represented in the calculation (Horgan & Nichols, 2015). In other words, when we reason reflectively, we may consciously represent something from our mind's perspective, but we are not thereby consciously representing our mind or its perspective. Consider some excerpts about reflection which need not involve self-consciousness.

"Having reflected a little on the Danger which we had…." (Thévenot, 1687, 134).

> "I have sometimes reflected for what reason the Turks…." (Maundrell, 1703, 14).
>
> "'Is it safe to eat beef?' The evidence is inconclusive, but we need to take a view. So, having reflected on the evidence and the risks, we make up our minds on the matter" (Frankish, 2004).

These passages involve representations, but it is not clear that these authors are representing themselves in their reflections.

So what is self-conscious reflection? Imagine a psychologist describes patterns in your thoughts and feelings about your father. As you listen to the psychologist and the thoughts and feelings manifest, you have an epiphany, "I resent my father!" In this case, you seem to reason not only reflectively, but self-consciously: the representations of your thoughts and feelings were clearly represented *as your* thoughts and feelings. Why else would you infer something about yourself from these thoughts and feelings?

Sometimes scholarship is clear about whether reflection involves general consciousness or self-consciousness in particular. Christopher Peacocke's "reflective self-consciousness" is a clear example of self-reflection (2014, Chapter 9). Sosa also uses 'reflection' in ways that clearly involve self-consciousness: "he could always know such properties of his belief by reflection; that is, through mere introspection, memory, and reason" (Sosa, 1991, p. 193). Korsgaard often describes reflection as a form of self-conscious reasoning. "The reflective structure of human consciousness requires that you identify yourself with some law or principle which will govern your choices" (Korsgaard, 1996). Here 'reflection' seems to involve conscious representation of one's own identity.

However, some scholarship is less clear about whether reflection must involve self-consciousness. For example, Kornblith describes reflection as, "Thinking about one's own mental processes from a first-person point of view" (Kornblith, 2012, p. 28). Here it is unclear whether the first-person point of view is itself consciously represented in such reflection. Consider another example from Korsgaard. "I find myself with a powerful impulse to believe. But I back up and bring that impulse into view…. Now the impulse doesn't dominate me and now I have a problem. Shall I believe?" (Korsgaard, 1996). Here the impulse and the question about what to believe seem to be the important conscious representations. It is not clear that representing one's self is as important as representing one's representations. Indeed, the same point can be made without explicit mention of one's self: There is a powerful impulse to believe it, but backing up brings that impulse into view and then the impulse does not seem so dominant and a question remains: should it be believed?

There are two upshots to this discussion of reflection and self-consciousness. Firstly, there is an opportunity for scholarship to acknowledge that reflection and self-reflection are orthogonal. Secondly, the nature and normativity of reflection cannot be inferred solely from the nature or normativity of self-consciousness or related concepts such as self-knowledge. Reflection requires consciousness, but not necessarily self-consciousness. So if we find that "reflection on our beliefs and decisions distorts our view of our own mental processes" (Kornblith, 2019a), then this may be a problem for the normative value of *self*-reflection even though it is not necessarily a problem for the normative value of reflection more generally.

1.2 Reflection As Conscious and Deliberate

A brief overview of how 'reflection' is used in various contexts found that reflective reasoning was often understood to be deliberate, conscious, and relatively slow. However, we admitted that

the relative slowness of reflection might have exceptions. This suggests that reflection has at least two key features: deliberate processing and conscious representation (Shea & Frith, 2016). Investigating additional uses of 'reflection' is worthwhile, albeit beyond the current project's scope, which is to explicate reflection for philosophy and science. To do this, I will explain the meaning and measurement of reflection's deliberateness and consciousness.

1.2.1 Conscious Representation

Reasoning is conscious just in case the reasoner is representing parts of their reasoning consciously (Shea, 2013). When I say that something is represented consciously, I mean that the representation is available for processing at the personal level (Dennett 1969, pp. 90-99). One way to test whether contents are available at the personal level is to see if the contents can be articulated verbally. So if someone can "think aloud", then at least the verbalized contents of their reasoning seem to be consciously represented (Fox, Ericsson, & Best, 2011). We can unpack each part of this explication of 'conscious' below.

Processing at the personal level. Daniel Dennett's personal/sub-personal distinction refers to levels of analysis and explanation. The distinction is between folk psychological states and the mechanistic explanations thereof. So when someone asks, "What is pain?" we can give at least two kinds of answer. One kind of answer simply describes one's pains, how one knows that they are in pain, when one is likely to experience pain, etc. This is explanation at the personal level. Another kind of explanation itemizes the physiological events in the causal chain involved in pain. This is explanation at a sub-personal level. The personal/sub-personal distinction can be applied to explanation as well as it can be applied to mental processing.

The idea is that a person can be somewhat aware of the processing of mental representations—e.g., you are aware of the meaning of this sentence. Moreover, a person can

consciously reason about the representations of which they are aware—e.g., you can think of an objection to what you are reading. And when someone is so aware of processing representations that they can reason about them, then we can say that their representations are available for processing at the personal level. In short, they are conscious.

This, of course, contrasts with unconscious or sub-personal representations: representations that someone is not so aware of that they can reason about them. For instance, we seem unable to consciously process certain stimuli that are presented for only a few milliseconds (Bargh, 1992; Neely, 1977). So if these representations are processed, it would be strange to say that persons consciously process them. Rather, something sub-personal (brains, neurons, etc.) processes them outside the person's conscious awareness. Thus, these are said to be sub-personal or unconscious representations (e.g., Figdor, 2018; Shea, 2018).

Conscious awareness of what? Consciousness has a few meanings in philosophy. For instance, some distinguish between phenomenal consciousness and access consciousness (Block, 1995). Some evidence suggests that phenomenal consciousness—i.e., "what it's like" to be conscious—is not a widespread concept (Sytsma & Ozdemir, 2019). Indeed, the kind of consciousness we have found in ordinary, philosophical, and scientific language has been access consciousness: conscious representations of our thoughts, attitudes, etc. We can distinguish between consciously representing the source of one's reasoning's, the contents of one's reasoning, and the impact of one's reasoning (Figure 1).

Consciousness of content. In the present paper, when I refer to conscious representations, I am referring to the contents of one's reasoning. For the present purposes, conscious contents can, but need not be propositional (Buckner, 2019; Schwitzgebel, 2002, 2010).

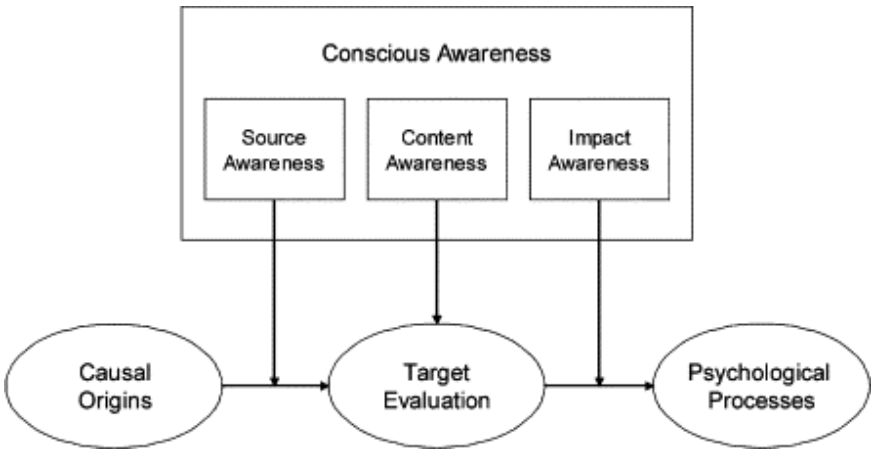

Figure 1. Three distinct targets of conscious representation from Gawronski, Hofmann, and Wilbur (2006).

Consciousness of only source and impact. Think aloud protocols can be crucial for understanding how people complete tasks (Ericsson & Simon, 1980, 1984; Ericsson, 2006). Among other things, such think aloud protocol analysis reveals some of the representations that are conscious during the task, the chronological order of the representations, etc. Of course, there may many unconscious representations involved in completing the task, even when we are thinking aloud. And because participants are not conscious of these representations, they may report about them inaccurately if they report them at all (Doris, 2015). Nonetheless, this is not a reason to think that all verbal reports are so inaccurate (Ericsson & Simon, 1984; Schwenkler, 2018). After all, when participants confabulate an inaccurate story about their unconscious reasoning, we take it for granted they are at least accurately reporting the representation of their confabulation.

One might wonder about cases in which we are conscious only of the conclusion of some reasoning—e.g., when we have an intuition unaccompanied by any explicit reasoning (Mercier & Sperber, 2017; Nagel 2012, Section 1; Sperber, 1997). In that case, the antecedent reasoning is

unconsciously processed. After all, consciousness of only the source or impact of one's reasoning would not constitute the kind of conscious representation that features in reflective reasoning (Shea & Frith, 2016). Even so, processing can be deliberate even when it is not conscious.

1.2.2 Deliberate Processing

Judgment is deliberate just in case it involved inhibiting or considering an alternative to initial, autonomous response(s). We can unpack each part of this explication of 'deliberate' below.

Inhibiting Autonomous responses. John Bargh (1992) introduced the concept of autonomic, which describes a process that, once started, can run to completion without conscious attention. Of course, autonomous processes can be started without deliberation. When something is presented to us, autonomous processing can produce a response without our consciously representing any of that processing. And if we immediately accept that response once we consciously represent the response, e.g., by verbally reporting it as correct, then that response is not deliberate. If however, we inhibit the processing that leads to that response or do not immediately accept the response after it is consciously represented, then doing so is deliberate.

So one common method of observing deliberateness involves giving someone a task that is known to elicit an autonomous response and instructing them to respond incongruently with their autonomous response (Chapter 3). That way, if someone responds incongruently with the known autonomous response, then we can infer that they either inhibited that response or did not endorse it—i.e., we can infer that their response was deliberate.

Verbal report and deliberateness. Concurrent think aloud verbalizations may further clarify if an autonomous response was deliberately inhibited or reconsidered (Byrd, Gongora,

Joseph, and Sirota, *forthcoming*). This is because we can observe deliberateness when people reason aloud. Suppose we give research participants a task and ask them to think aloud as they complete it. Immediately after reading the first question aloud, many participants say, "It's [P]" and begin reading the next question (Szaszi, Szollosi, Palfi, & Aczel, 2017). These participants are not responding deliberately. So they are not responding reflectively. However, some participants finish reading the first question aloud and say something more reflective like, "I think [P], but is that right?" These participants are not immediately accepting their autonomous response, so they are responding deliberately—even if their ultimate response aligns with their autonomous response.

1.2.3 Not Just Classificatory, But also Comparative

In reality, neither the consciousness nor deliberateness of reasoning is binary. That is, it would be difficult to draw a clear and categorical boundary between conscious reasoning unconscious reasoning or between deliberate and autonomous reasoning. After all, we might consciously represent only some part of our consciously representable reflective reasoning (Mercier & Sperber, 2009). And we might only partly inhibit or reject autonomous responses. So the consciousness and deliberateness of reasoning will likely come in degrees.

What I am suggesting is a "comparative" rather than a "classificatory" analysis of reflective reasoning (a la Carnap 1950, §8). A classificatory analysis of reflection is one that allows us to say when a case of reasoning is or is not reflective. For example, X and Y are reflective, but Z is not. A comparative analysis of reflection, on the other hand, allows us to say whether one case is more reflective than another. For instance, X is more reflective than Y.

Perhaps a mature psychology of reasoning will be able to give a quantitative analysis of reflection (ibid.). A quantitative analysis would allow us to index the degree to which cases of

reasoning are reflective. For instance, a quantitative analysis might allow us to quantify the percentage of consciously representable representations that were consciously represented in any given case of reasoning. One might think that because common measures of reflection in psychology do quantify differences in reasoning performance (e.g., scores on the cognitive reflection test, reaction times, etc.), we already have quantitative analyses of reflective reasoning. However, these quantities do not seem to neatly track consciousness or deliberateness as construed herein. And so these common measurements do not allow for the kind of quantitative analysis of consciousness or deliberateness that I have in mind—even if they do quantify other differences in reflective reasoning—e.g., its duration, it success rate, etc.

1.3 Reflection & Its Cousins

Given that reflective reasoning is construed as consciously represented and deliberate, we can use a 2x2 matrix to distinguish reflective reasoning from other forms of reasoning (Table 1, adapted from Shea & Frith, 2016): implicit, ruminative, and reformative.

Table 1. A 2x2 matrix of deliberateness and consciousness used to distinguish reflective reasoning from other kinds of reasoning.

		Processing	
		Less Deliberate	More deliberate
Representation	More conscious	Ruminative	Reflective
	Less Conscious	Implicit	Reformative

1.3.1 Implicit Reasoning

Implicit Reasoning is reasoning that is neither conscious nor deliberate. So it is unreflective. Implicit reasoning might have impacts on our dispositions, judgments, beliefs, decisions, and

behavior independent of the impacts of other kinds of reasoning. For instance, our implicit reasoning might bias our behavior in ways that we do not realize. You might be thinking of measures of how implicit associations allegedly influence our behavior in ways that we do not realize (Greenwald, McGhee, & Schwartz, 1998). We should be careful here. It is one thing to include 'implicit' in the name of a psychometric tool. It is quite another to show that the psychological attribute measured by the tool—as opposed to the tool itself or the process by which the tool measures psychological attributes—is implicit (Gawronski & De Houwer, 2014). So the present paper does not take a position on whether the implicit association test measures something implicit or whether the process by which it measures behavior is implicit.

Given that implicit reasoning is not consciously represented, we might find ourselves disagreeing with it if we become consciously aware of it or its impacts (Greenwald & Banaji, 1995). Further, if we become aware of unreflective reasoning frequently enough, then we might be able to anticipate it based on situational cues that match previous encounters with unreflective reasoning. Moreover, if we anticipate implicit reasoning, then we might deliberately and consciously attempt to inhibit them (Devine, 1989; Byrd, 2019). That is we might challenge our unreflective reasoning by deploying more reflective reasoning. For example, when our behavior is implicitly biased in ways that we reflectively disavow, we can become anxious about contexts in which our behavior might be implicitly biased (Gaertner & Dovidio, 1986). This anxiety might be related to another kind of reasoning: ruminative reasoning (Harrington & Blakenship, 2002).

1.3.2 Ruminative Reasoning

Ruminative reasoning is reasoning that is conscious, but autonomous. There are many modes to rumination. When we find ourselves autonomically and repeatedly thinking about the

great news that we received earlier, we are experiencing positive rumination. Perhaps rumination can be neutral — as in certain cases of conscious mind-wandering (Christoff, Irving, Fox, Spreng, & Andrews-Hannah, 2016). But ruminative reasoning can also be negative. Negative rumination predicts a wide range of maladaptive cognition (Nolen-Hoeksema, Wisco, Lyubomirsky, 2008).

One proposed treatment for negative rumination is a form of reflective reasoning. It is most common in the cognitive behavioral therapy (CBT) tradition. The reasoning is often described as taking one's thoughts to trial. It is an instance of reflective reasoning because it involves deliberately processing conscious representations. In particular, it involves deliberately subjecting one's thoughts to scrutiny. For example, we can question the background assumptions of our ruminative thoughts. One effect of this exercise is the realization that the underlying assumptions of negative rumination are often impoverished or false. Although this reflective realization is helpful in the moment, it might not change our ruminative habits in the long run. That might require reformative reasoning.

1.3.3 Reformative Reasoning

Reformative reasoning is unconscious but deliberative. For instance, one might gradually reform their unconscious stereotypes about another group of people by deliberatively exposing oneself to counterstereotypic representations of people (Byrd, 2019). And these deliberative attempts to intervene on one autonomous social impulses can lead to interactions that decrease one's negative stereotypes about other groups (Davies, Tropp, Aron, Pettigrew, & Wright, 2011; Paolini, Hewstone, Cairns, & Voci, 2004).

It is important to note that the impulse to befriend someone from another group can be deliberate even if we do not consciously represent the subsequent changes in content to our

stereotype(s) about that group. For example, we might deliberately intervene on our autonomous social interaction processes so that we interact with people from a group that we might not otherwise interact with. And if the interaction goes well, then it may ameliorate a negative stereotype about people unlike us. But are do not typically represent the stereotype consciously, so do not consciously notice when the stereotype is updated by interactions with members of its group. We might not even consciously represent the stimulus that updated our stereotype. So even though interaction with someone from another group is deliberate, its impact on reasoning can be unconscious.

Of course, not all reformative unconscious reasoning is so ameliorative. For example, sometimes we deliberately interact only with people who look and think like us. This deliberate decision decreases the probability of positive interactions with people that are different than us. So we miss an opportunity to ameliorate our negative unconscious representations of their group.

1.4 Mapping Reflection Onto Philosophers' and Scientists' Theories

The current two-factor account of reflection as conscious and deliberate has some limitations. It remains to be said how reflective reasoning fits into philosophers' categories of practical and theoretical reasoning. Also, it remains unclear whether reflection maps neatly onto the cognitive scientists' categories of System 2 or Type 2 reasoning.

1.4.1 Practical & Theoretical Reasoning

There are various construals of the distinction between practical and theoretical reasoning. A prominent construal distinguishes one's reasoning about what they ought to do, given one's own circumstances from reasoning about what anyone ought to believe independently of any one person's circumstances (Wallace, 2018). In short, it is a distinction between reasoning about action and reasoning about truth.

Korsgaard's account of reflection is explicitly "practical and not theoretical: it is reflection about what to do, not reflection about what is to be found in the normative part of the world" (1996). Nagel's account of reflection seems to be compatible with theoretical reasoning because its aim is attitudes—like belief—rather than just action: "Reflective cognition requires the sequential use of a progression of conscious contents to generate an attitude" (Nagel, 2014, p. 231). This compatibility suggests that an account of reflection need not, in principle, be an account of only practical or only theoretical reasoning. Indeed, conscious and deliberate reasoning can help us decide what to do as well as what to believe.

1.4.2 Dual Process Theory

There are many dual process theories. Some dual process theories are general, applying to all rational processes (Evans, 2013). Other dual process theories are aimed primarily at one domain—e.g., morality (Baron, 1994; Greene, 2013). Further some dual process theories contain normative claims that others do not (Evans & Stanovich, 2013). Nonetheless, the general idea behind dual process theories is that we can distinguish between at least two ways of reasoning (e.g., De Neys, 2017; Evans, 2019; Evans & Frankish, 2009).

There is some variation in the naming scheme for these two ways of reasoning. Some refer to System 1 and System 2 reasoning (e.g., Kahneman, 2011) and others refer to Type 1 and Type 2 reasoning (e.g., Evans & Stanovich, 2013). If we label dual process theories' duality with 'unreflective' and 'reflective', then we might think that the present two-factor account of reflective reasoning adequately captures what other cognitive scientists mean by System 2 or Type 2. However, a word of caution is in order. For instance, some have argued that some of reflection's cognates—e.g., 'critical thinking'—will not map so neatly onto talk about System or Type 2 reasoning (Bonnefon, 2016). Importantly, reflection is a more general category than

critical thinking: reflection can be used for non-critical ends. Another concern is that system-talk might be misleading: "reflective thinking is not something that happens in a separate system independent of the operation of intuition; reflection is realized in successive cycles of intuitive thought whose outputs are posted to consciousness" (Nagel, 2014, p. 231).

I find 'System/Type 1' and 'System/Type 2' less informative and memorable than 'unreflective' and 'reflective' (Byrd, 2019). For these reasons, I express the dual process theory distinction as a distinction between unreflective and reflective reasoning rather than a distinction between systems or types of reasoning.

1.4.3 Conceptual Engineering

Some have argued that concept explication should be determined, in part, according to utility (Carnap, 1950b; Haslanger, 2012). Some have also pointed out that this utilitarian approach to conceptual engineering sometimes encounter tradeoffs between the utility of a concept and its similarity to the phenomena one wants to explain (Dutilh Novaes, 2018). So, some conceptual engineers have prioritized utility over similarity (Shepherd & Justus, 2014).

To these conceptual engineers, the ultimate test of the current explication of reflection will be its utility to the philosophy and science of reflection. Of course, this test is empirical. Fortunately, there are frameworks for conducting this kind of test (ibid.). The point here is just that the current explication contains hypotheses that are worthy of investigation.

1.5 Conclusion

In the present investigation of ordinary, philosophical, and scientific discourse, 'reflection' was found to refer to thinking that is conscious and deliberate. Subsequent investigation revealed how reflection that is conscious is distinct from and does not entail reflection that is self-conscious. Then empirical explications of reflection's conscious and

deliberate features were provided. These explications helped explain how reflection can be detected. They also distinguished reflection from nearby concepts such as ruminative and reformative reasoning. After this, I found that reflection is not obviously limited to only practical or only theoretical reasoning. I also found reasons to prefer ‘unreflective’ and ‘reflective’ to dual process theorists’ labels containing ‘system’ or ‘type’. With this two-factor, empirical explication of the nature of reflection, we can now turn to the normativity of reflection.

CHAPTER 2
BOUNDED REFLECTIVISM & EPISTEMIC IDENTITY

> Even advanced reflection and training does not insulate one from illusion: … physics graduate students and postdoctoral researchers still experience the characteristic cognitive-perceptual illusions of naïve 'impetus theory' physics….
>
> –Nagel, 2012

> We endorse or reject our impulses by determining whether they are consistent with the ways in which we identify ourselves. […] You are a mother of some particular children, a citizen of a particular country, an adherent of a particular religion, …. And you act accordingly – caring for your children because they are your children, fighting for your country because you are its citizen, refusing to fight because you are a Quaker, and so on.
>
> –Korsgaard, 1996

2.1 What's So Great About Reflection?

You agree to cover the tip for lunch with your friend. You put down an amount of money that feels right. Your friend glances at the money and appears surprised. So you do some calculation in your head, realize that you forgot to factor in your friends' portion of the bill, and add some money to your tip. On the way out of the restaurant, your friend asks you about your political party's latest scandal. You immediately play defense, rehearsing various rationalizations of the scandal. After your monologue, your friend recalls that you criticized the opposing party for the same kind of scandal in the last election.

This story reveals a puzzle about reflective reasoning: reflection often helps, but reflection can also hinder our reasoning. It helps us in everyday circumstances such as double-checking our math. However, it can hinder our reasoning depending on how we use it—e.g., when we use it for partisan rather than impartial purposes. This puzzle about reflection also manifests in the research about reflection. Philosophers and scientists disagree about the role of reflection in reasoning. Some seem to think that reflection is crucial for good reasoning (e.g., Epstein, 1994; Klein, 1998; Korsgaard, 1996; Sosa, 1991). Others are more pessimistic, saying that reflection makes our beliefs "no more reliable" (Kornblith, 2019b). So what gives? Does reflection help our reasoning or not? Or, rather, when does it help and when does it not help?

To address this puzzle about reflection, I offer a middle way between these two possibilities. Reflection does not necessarily make our reasoning good or bad. Rather, reflection is a tool; its costs and benefits depend on its use. For instance, reasoning might be worse when reflection is influenced by partisan goals. In other words, reflection's utility could be bound by factors such as identity. Hence, the name bounded reflectivism.

2.1.1 Theory: Reflectivism & Anti-Reflectivism

Many philosophers take reflection to be crucial for obtaining various intellectual goods. We can call such philosophers reflectivists (Ferrin, 2017). Consider reflectivism's history. "Reflective agency" was said to be important to understanding human action (Kennett & Fine, 2009; Velleman, 1989, 2000; Wallace, 2006). "Reflective endorsement" was said to be necessary for morality to have normative force. (Korsgaard, 1996). "Reflective equilibrium" was said to be necessary to discern and/or justify principles of logic and justice (Goodman, 1983; Rawls, 1971). "Reflective knowledge" was said to be a distinctive capacity of humans that is necessary to understand our beliefs in context and "how they come about" (Sosa, 1991). "Reflective scrutiny"

was said to be necessary for evaluating our ethical outlook from within—as opposed to evaluating neutrally, from the outside, or by merely re-expressing itself (Hursthouse, 1999). "Reflective self-consciousness," was described as a unique and privileged form of self-knowledge. There are probably more instances of reflectivism in philosophy (see Doris, 2015). While an exhaustive catalog of each instance is a valuable historical project, the current paper need only introduce reflectivism's ongoing and wide-ranging influence in philosophy.

Opposing reflectivism, of course, is anti-reflectivism. Anti-reflectivists argue that reflection cannot do or be what reflectivists think. Examples include arguments that reflection cannot be a virtue (Dreyfus, 1986), cannot justify our judgments (Kornblith, 2012), and cannot give us the self-knowledge that many reflectivists imagine (Doris, 2015).

So who is right? The reflectivists or the anti-reflectivists? I want to suggest that both reflectivists and anti-reflectivists get something right. Reflectivists are right to think that reflective reasoning *can* deliver the normative value that reflectivists seek. However, anti-reflectivists are right to think that there is some evidence that reflection *can* either lack this normative value or else undermine it. The middle view that I develop concedes these points, but without as much optimism as reflectivism or as much pessimism as anti-reflectivism.

2.1.2 Evidence Of Good And Bad Reflection

In the wake of the 2016 US presidential election, public discourse in United States was preoccupied with a particular undesirable outcome: the ways in which online fake news was weaponized to influence the US presidential election. Reddit's CEO announced cooperation with federal investigations of the dissemination of fake news on their website and concluded with the belief that "the biggest risk we face as Americans is our own ability to discern reality from nonsense" (Huffman, 2018). Part of the worry was that reasoning is politically partisan.

Some researchers find evidence for such partisan reasoning. For example, multiple studies of over 1000 people found that political conservatism repeatedly predicted that an action was more morally wrong if it was performed by a left-wing than a right-wing agent—$b = 0.33 – 0.41$, $SE = 0.04\text{-}0.58$, $t = 6.87\text{-}9.31$, $p < 0.001$ (Everett et al., 2018, Studies 7a-7d). And in a meta-analysis of over 50 studies involving over 18,000 people, liberals and conservatives rated politically congenial information as more valid, higher quality, or more acceptable than politically uncongenial information— $k = 51$, meta-analytic $r = 0.245$, 95% CI [0.208, 0.280], $p_r < 0.001$, $Q_W = 307.96$, $p_w < 0.001$, Tau = 0.120 (Ditto et. al., 2018). So how might reflection help or hinder such partisan reasoning?

Reflection as a solution. In two studies of over 800 participants, people who performed better on certain reflection tests (Frederick, 2005; Thomson & Oppenheimer, 2016) were significantly more likely to correctly estimate the accuracy of fake news—Study 1: r = -0.3, $p < 0.001$; Study 2: r = -0.26, $p < 0.001$—and significantly less likely to share fake news—Study 2: r = -0.19, $p < 0.01$—even when the source of the news was removed and when headlines aligned with their partisan identity (Pennycook & Rand, 2019b). Across two other studies of about 2000 participants, such reflection predicted more reliance on mainstream news sources over hyperpartisan and fake news—Study 1: $\beta = 0.16$, $p < 0.0001$; Study 2: $\beta = 0.15$, $p < 0.0001$ (Pennycook, & Rand, 2019a). In other words, reflective reasoning was associated with more desirable reasoning in everyday and seemingly high-stakes contexts such as political reasoning. These findings suggest that reflective reasoning is part of the solution to problems with discerning the difference between reality and non-sense.

Reflection as ineffective. Unfortunately, reflective reasoning is not a panacea. Consider the illusory truth effect. Multiple studies of over 1000 people found that encountering false

information repeatedly made people more likely to believe it regardless of their performance on reflection tests—Studies 1, 2, and 5 meta-analytic r = -0.01, 95% CI [-0.09, 0.07]; Studies 5, 6, 7 meta-analytic r = -0.05, 95% CI [-0.10, 0.01] (De Keersmaecker et al., 2019). These findings suggest that some reasoning problems might be immune to reflection.

Reflection as a problem. In fact, reflection might even make matters worse. Multiple experiments found that people who are more likely to reason reflectively are also more likely to reflect in ways that serve their partisan identities. In one experiment, when participants were told that open-minded people who accept climate change are more likely to get the correct answers on the Cognitive Reflection Test (CRT), right-leaning participants were less likely to report that the CRT was valid while left-leaning participants were more likely to report that the CRT was valid (Kahan, 2013). However, this effect of partisanship actually increased among more reflective participants, r = -0.3, $p < 0.01$ (ibid.). In another experiment, participants interpreted fictional studies. More reflective participants were more likely to correctly interpret the findings of studies about rashes. Alas, when interpreting studies about politically-salient topics like gun control policy, right-leaning individuals were more likely to misinterpret evidence that supported bans on pro-conceal-carry policies and left-leaning participants were more likely to misinterpret evidence that supported conceal-carry policies (Kahan, Peters, Dawson, & Slovic, 2017). But, once again, this partisanship effect on reasoning was more dramatic among more reflective participants, r = 0.54, $p < 0.05$. These findings suggest that while reflection can help reasoning in some cases, reflection can also hinder our reasoning—e.g., when deployed for partisan purposes.

Overall, these data suggest that reflection can be both part of the solution, seemingly ineffective, or else part of the problem. With these mixed results, it may seem unclear what we should think about the normative value of reflection. Recall, however, that one of the key factors

in determining whether reflection helped or hindered reasoning was partisanship—or what I call *epistemic identity*.

2.2 Epistemic Identity

Korsgaard famously discusses the roles of practical identity in reflection (1996). Another form of identity is *epistemic identity*: the phenomena of treating certain beliefs as part of one's identity. Suppose that I identify with a religion. If you challenge my religious identity, then I might reflectively defend my religious beliefs rather than dispassionately submit to the best arguments and evidence. Or suppose that you identify with a particular political party—one of those political parties that explicitly codifies its ideological commitments in a party platform that is recited in its public speeches and advertisements. In other words, you identify not only with the party, but its values and beliefs. In this case your political identity is an epistemic identity.

We have already encountered evidence of an effect of epistemic identity on reflection. However, there is also evidence of the impact of epistemic identity on reasoning more generally. For instance, judgments about evidence of global warming are more correlated with self-reported political ideology ("left" vs. "right") in the United States than in any of the other 25 Western, developed countries tested, $d > 0.4$ (Hornsey, Harris, & Fielding, 2018, Figure 1). Moreover, people in the US who identify as Democrats almost unanimously believe the evidence for global warming while only around half of people in the US who identify as Republicans believe this. And this gap in belief about the evidence as a function of political party identification has grown steadily since around 2006 (Pew Research Center, 2017). Given these data, it is unsurprising that identity has become an increasingly common part of reasoning research (e.g., Oyserman & Dawson, 2019; Strohminger, 2018; Van Bavel & Pereira, 2018).

2.2.1 Proximal And Distal Effects Of Epistemic Identity

We can distinguish between proximal and distal influences on action (Mele 2008, Chapter 2). Likewise, epistemic identities can have proximal and distal influences on our reflective reasoning. Epistemic identities have proximal influences on our reflective reasoning when they determine how our reflective reasoning proceeds in any given moment. For example, you might have heard about an outstandingly opinionated and uncivil relative—often described as an uncle who is quick to assert his pet beliefs and loudly dismiss any opposing evidence and arguments (Lynch, 2018). Of course, partisanship can influence reflection in more subtle ways as well.

Distal effects of epistemic identity. Epistemic identities can also have more subtle, longitudinal influences on not only reasoning in general, but reflective reasoning in particular. Our epistemic identities can influence what we seek, what we attend to, what we perceive, and thereby what and how we remember (Kahan, 2017). So when we represent our memories consciously to reason about them deliberately, we may be working with systematically biased priors. For example, members of the sitting president's party will be more likely to seek out, attend to, perceive, and remember the successes and unfair criticisms of the sitting president than members of other parties. And, conversely, members of the sitting president's parties will be less likely to seek out, attend to, perceive, and remember the failures and level-headed criticisms of the sitting president than members of other parties. So the epistemic differences between the two groups are doubled by opposing distal influences on their worldview. This is part of the reason why providing people with more information does not always reduce polarization (Gershman, 2018).

Proximate effects of epistemic identity. The distal influences of epistemic identity can result in proximate impacts of epistemic identity (Van Bavel & Pareira, 2018). For example, it can explain why we might reflectively rationalize our political party's scandals while reflectively criticizing the opposing party for the same kind of scandal: we do not perceive and/or recall the scandals as equivalent. Identity-based processing can also explain why people reject their long-professed ideals when their party rejects them: when a judgment aligns with their current party leader, then they might feel confident enough to not evaluate if the judgment aligns with prior party leaders.

In other words, we might be less likely to notice when our reflective judgments conflict with our ideals and evidence—a proximate effect—because the evidence we recall from memory while reflecting was perceived and/or encoded in partisan ways—a distal effect. Some have argued that such effects of epistemic identity might even influence philosophical discourse (Peters, 2019).

2.2.2 If You Can't Beat Epistemic Identity, Embrace It

Existing solutions to the problem of epistemic identity involve imagining someone else's identity (Hannon, 2020) and activating a superordinate identity (Van Bavel & Pereira, 2018). My suggestion is a combination of these suggestions: to overcome the undesirable impacts of epistemic identities, we will need to appeal to *shared*, *superordinate* epistemic identities.

When a colleague and I reflectively double-down on a political disagreement, then we should stop to ask, "Although we disagree, what *should* we think about this *as scientists*?" Because we both identify as scientists and we agree about the importance of science-based policy, undesired polarization is less likely when we reflect as scientists than when we reflect as political partisans (e.g., Kahan, Landrum, Carpenter, Helft, & Jamieson, 2017).

Depending on the context, shared identities need not be shared by everyone. They may only need to be shared by those involved in the undesirably polarized disagreement. For instance, in a dangerously divisive political climate, appealing to shared national identity—as opposed to, say, a globalist or international identity—might lead to productive agreement (Talaifar & Swann, 2018).

Korsgaard tells us that, "If the problem springs from reflection then the solution must do so as well" (1996). I want to make a similar, but weaker claim about epistemic identity's impact on reflection: these aforementioned findings suggest that epistemic identity can be a problem for reflection, but they also suggest that epistemic identity might be part of the solution.

2.3 The Bounded Reflectivist Model Of Reflection

We now have almost all of the components of the bounded reflectivist model of reflection. Reflection is said to be deliberate and conscious (Shea & Frith, 2016); reflection can deliver goods that reflectivists seek, but reflection can also hinder our reasoning in ways that reflectivists have suggested; and additional factors influence whether reflection helps or hinders our reasoning such as epistemic identity. However, before we can synthesize the bounded reflectivist model of reflection, we will need to account for the antecedents of reflection (Evans, 2007). In other words, what triggers reflection?

2.3.1 The Triggers Of Reflection

Some have proposed that reflection is triggered either by a task (Nagel, 2012) or by how confident one feels about their autonomous response (Koriat, 2019; Mercier & Sperber, 2017; Pennycook, Fugelsang, & Koehler, 2015; Thompson, Prowse Turner, & Pennycook, 2011). Together, these potential triggers can explain performance on reflection tests that do and do not contain lures. Reflection tests are designed to lure us into unreflectively accepting a particular

answer that is, upon reflection incorrect (Pennycook, Cheyne, Koehler, & Fugelsang, 2015). For instance, the well-known bat-and-ball problem lures people to think that $0.10 is the correct answer.

> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (Frederick, 2005)

However, consider a recent variant of the famous bat-and-ball problem.

> A bat and a ball cost 96 cents in total. The bat costs 2 cents more than the ball. How much does the ball cost? (Baron, Scott, Fincher, & Metz, 2015)

This variant, unlike the original, does not lure participants toward a particular incorrect answer. So there is no obvious reason to think that correct responses to this question involve deliberately inhibiting, feeling wrong about, or reflecting on a particular autonomous response.

However, some suggest that reflection can also be triggered by task novelty, stakes, or imagination (Nagel, 2012). Novelty can explain how reflection could be triggered by the non-lured bat-and-ball problem. While the non-lured problem does not elicit a particular autonomous response like its predecessor, reflection might nonetheless be triggered if the problem seems novel, high stakes, or imaginative. For example, reasoners who do not regularly solve mathematical tasks might find the non-lured variant of the bat-and-ball problem novel enough to prompt reflection about it.

2.3.2 Visualizing The Bounded Reflectivist Model

To spare readers the difficulty of imagining all of the model's parts, the triggers, steps, and outcomes of reflection are visualized algorithmically (Figure 2). The algorithm will also be described below, step-by-step.

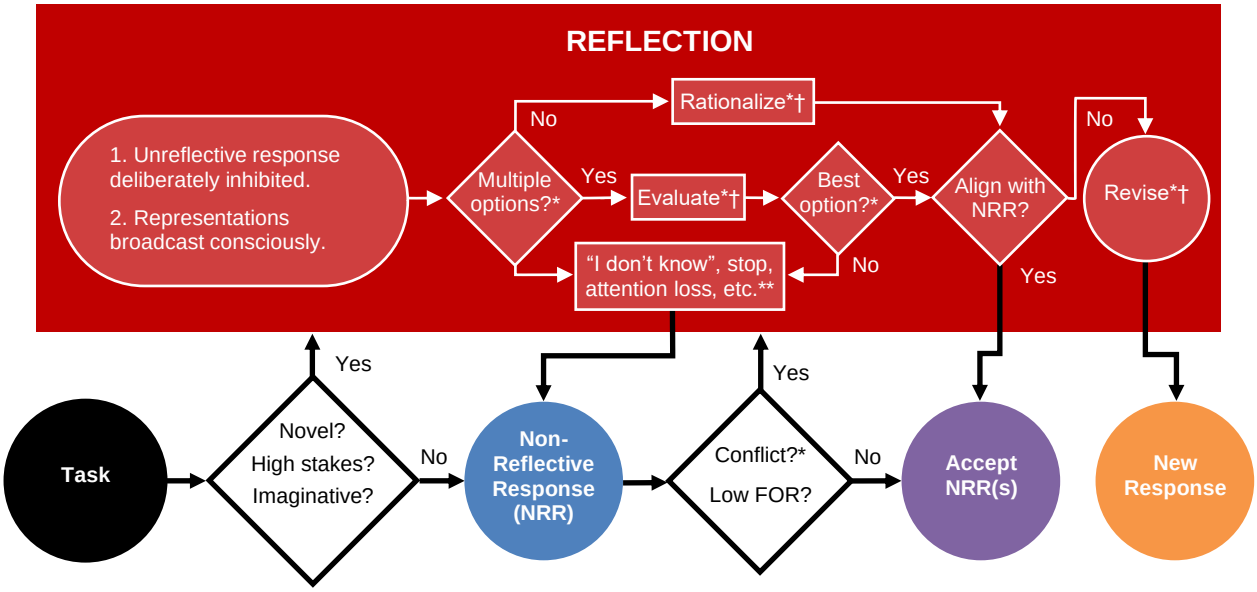

Figure 2. The causes, processes, and outcomes of partisan reasoning, reflective and unreflective. * Can be influenced by factors like epistemic identity. **Can involve new reasons and/or slightly new response even if same spirit of IR. † If interrupted, null response.

When reflection is not triggered. The visualization illustrates how tasks can prompt either an autonomous response or reflection. If the task is not novel, high stakes, or imaginative, then reflection will probably not be triggered. At most, an autonomous response will occur. Whether we accept the autonomous response or reflect on it will depend on whether we detect conflict between the autonomous response and some other response or whether we have a feeling of rightness about the autonomous response. If no conflict is detected and one's feeling of rightness is high, then the autonomous response will probably be accepted. Of course, our acceptance of an autonomous response might be influenced by epistemic identity. After all, we are more likely to unreflectively endorse conclusions that comply with our prior beliefs, even when doing so is logically fallacious (Janis & Frick, 1943).

When reflection is triggered. Alternatively, the task or a feeling about the autonomous responses can trigger reflection. Recall that reflection involves deliberately inhibiting

autonomous responses and consciously representing part of reasoning (Chapter 1). The first step of reflection involves considering options. “P seems true, but is it?” “Might Q be true instead?” What about not-P?” The search for options might end with an empty set of options—e.g., “Cannot be determined”, “I don’t know”—or else get interrupted, leaving the reasoner with, at most, an unreflective response. Alternatively, the reflective search can reveal one or more options. If only one option is considered—either because one reflects in a close-minded manner or for some other reason—then we can reflectively rationalize it (Cushman, 2019; Mercier & Sperber, 2017). If more options are considered—either because one reflects in a more open-minded manner of for some other reason—then we can reflectively evaluate them and decide whether there is a best option. Again, epistemic identity can influence the process of reflection—e.g., during the evaluation phase of reflection.

Evaluation. During reflective evaluation, we might find that an autonomous response was, in fact, the best response. In this case, reflective reasoning would result in accepting the autonomous response—even if for new and/or better reasons than we considered prior to reflection. However, this reflective evaluation might find that an alternative response is superior. In this case, reflection would revise the initial unreflective response to a new response. Of course, Sections 2.1 and 2.3 remind us that we sometimes reflectively evaluate arguments and evidence under the influence of epistemic identities.

2.4 Implications

The bounded reflectivist model operationalizes reflection for scientific inquiry, offers a two-stage account of how reflection can be triggered, and offers an identity-based account of how reflection can help or hinder reasoning. The primary implications of this model of reflection

have to do with the problems of and solutions to partisan reasoning and epistemic identity. However, the model also has implications for reflection tests and the normativity of reflection.

2.4.1 The Validity of Reflection Tests

Many researchers have assumed that correct responses on reflection tests involve overcoming a default (i.e., autonomous) response—hence, the default-interventionist account of analytic reasoning (Evans, 2007; see also Johnson-Laird & Ragni, 2019). But when researchers record participants thinking aloud as they complete these measures of reflection, they often find that participants can immediately respond correctly with no verbal or other evidence of a lured response (Byrd, Gongora, Joseph, & Sirota, *forthcoming*; Szaszi, Szollosi, Palfi, & Aczel 2017). This suggests that the default-interventionist model of reflection cannot account for all reflection test performance.

However, the bounded reflectivist model of reflection can explain these immediate and correct responses on reflection tests. In particular, the model illustrates a route to a successful response that does not involve reflection: when a task is low stakes—as when participants are not rewarded or punished for performance on simple arithmetic questions—and when the task is so familiar that it requires no imagination, participants' autonomous responses might be correct without ever having to reflect.

In addition to explaining otherwise puzzling performance on measures of reflection, the bounded reflectivist model of reflection shows how we can misinterpret performance on reflection tests. After all, the model challenges the standard interpretation of correct responses on measures of reflection: think aloud protocols have revealed that some correct responses do not involve reflection (ibid.). This suggests that we can validate and improve reflection tests with think aloud protocol analysis (Chapter 3).

2.4.2 The Normativity Of Reflection

Recall the reflectivists who argue that reflection is crucial for good reasoning (e.g., Korsgaard, 1996; Sosa, 1991). These reflectivists think that reflection necessarily has normative value. Of course, anti-reflectivists argue against this claim (e.g., Kornblith, 2012; Doris, 2015). Nonetheless, reflectivists might be able to accept the negative conclusions of anti-reflectivists without utterly abandoning the normative value of reflection (Schwenkler, 2018). The bounded reflectivist model of reflection is a more "sensible reflectivism" (ibid.) between reflectivism and anti-reflectivism.

Reflective equilibrium. Bounded reflectivism's middle way between reflectivist and anti-reflectivism involves admitting that reflection's normative value is contingent. Reflection can confer normative value in some conditions and not others depending on factors like epistemic identity. When reflection is motivated by shared, superordinate epistemic identities, reflection might lead to equilibrium. However, when reflection is under the influence of other factors and identities, reflection can lead to undesirable polarization. So, dashing the hopes of reflectivism and deliberative democracy (e.g., Bächtiger, Dryzek, Mansbridge, & Warren, 2018; Goodman, 1983; Rawls, 1971), reflection may not yield the equilibrium that many people seek.

Strategic reliabilism. Of course, I have not predicted all of the conditions in which reflection will help or hinder our normative goals. Research can enumerate the ways reflection is used toward normative goals (Davies, Ives, & Dunn, 2015; Earp et al., *in press*; Savulescu, Kahane, & Gyngell, 2019) and test its efficacy across contexts. The resulting data could feature in a sort of strategic reliabilist account of reflection's normative value—strategic reflectivism: reflection should be deployed in contexts where its benefits have been to reliably outweigh its costs (Bishop & Trout, 2004, 2008; Stich, 1990).

Strategic reflectivism's admission that reflection *can* improve our reasoning is more optimistic than the anti-reflectivism according to which reflection "gives us the illusion that we have subjected our beliefs to a rigorous screening that will improve our epistemic position," when, "In fact, …it achieves no such thing" (Kornblith, 2019b). Similarly, strategic reflectivism's admission that reflection can produce counternormative outcomes is more pessimistic than the reflectivism which sometimes characterizes normativity as "the ability to survive reflection" (Korsgaard, 1996).

Reflective scrutiny. Bounded reflectivism can also address the objectivity problem. Consider how the objectivity problem manifests in metaphilosophy: when we try to justify our metaethical outlook from within the outlook, we end up merely re-expressing the outlook rather than justifying it. Some philosophers suggest that reflection is a solution to this problem (Hursthouse, 1999). However, bounded reflectivism takes a different approach, holding that metaphilosophical scrutiny requires a shared, superordinate epistemic identity.

Naturally, shared, superordinate epistemic identities do not guarantee external objectivity. However, such objectivity may be unachievable (Carnap, 1950a). So, instead, shared superordinate epistemic identities offer desirable consistency with internal objective standards. After all, without such a shared framework, it is not clear how reflection on my metaphilosophical outlook could have any normative force with others.

2.5 Conclusion

This chapter drew on philosophical and psychological models of reflection to propose a new, empirically adequate, model of reflective reasoning and epistemic identity. Of course, interesting questions about reflection and identity remain. For instance, is identity-driven reflection epistemically suspect "motivated reasoning" (Kahan, 2016) or merely an instance of

Bayesian rationality (Tapping, Pennycook, & Rand, 2018)? Further attention to questions like this will certainly advance our understanding and appreciation of reflection and identity.

Nonetheless, the bounded reflectivist model offers value to both philosophers and scientists. It's a clear, algorithmic explication of reflection, featuring empirical predictions about when we will reflect, and some predictions about when reflection will help (vs. hinder) reasoning. Also, bounded reflectivism blazes a middle way between reflectivism and anti-reflectivism by showing how reflection can be valuable—e.g., by producing equilibrium—while also being vulnerable to undesirable influences—e.g., unwanted partisan influence. Finally, bounded reflectivism also yielded implications about the measurement and metaphilosophical role of reflection. Insofar as other accounts of reflection cannot deliver all of these goods, we should prefer the bounded reflectivist model of reflection.

CHAPTER 3
ALL MEASURES ARE NOT CREATED EQUAL

"…the [cognitive reflection test] is intended to measure cognitive reflection, performance is surely aided by reading comprehension and mathematical skills"

—Frederick, 2005

"…think-aloud [verbal] reports provide …data [about] thinking during cognitive tasks"

—Ericsson, 2006

"…process dissociation procedure [can be] used to separate different bases for responding."

–Jacoby, 1991

3.1 Philosophical & Scientific Notions of Reflection

Reflective reasoning is more deliberate and consciously represented than unreflective processes like intuition (Chapter 1). In familiar or low-stakes tasks intuition—i.e., more automatic and unconscious processing—is often sufficient. However, in less familiar, higher-stakes, or imaginative tasks, we might be triggered to reflect on our autonomous and unconscious responses (Chapter 2). Calculating √4, for example, might be so familiar that '2' immediately comes to mind without any reflection. Calculating √40, however, is unfamiliar to most people

and therefore requires some reflection—e.g., consciously representing some numbers, some mathematical principles, and deliberately performing mathematical operations.

There are many tasks that reveal individual differences in the disposition to reflect. Most tasks lure participants toward a particular intuitively appealing yet demonstrably incorrect response. Importantly, the tasks are simple enough that they do not require specialized knowledge. Indeed, a moment's reflection reveals that a lured, intuitively appealing response is incorrect and/or that an alternative response is correct. Thus, participants' lured answers are coded as unreflective and correct answers are coded as reflective. This coding makes it possible to quantify people's disposition to reflect. People who provide more reflective answers on reflection tests are less susceptible to fake news (Pennycook & Rand, 2019b), less religious (Chapter 4; Pennycook, Ross, Koehler, & Fugelsang, 2016; cf. Gervais, van Elk, Xygala, McKay, & Aveyard, 2018), less conspiratorial (Čavojová, Secară, Jurkovič, & Šrol, 2019; Pennycook, Fugelsang, & Koehler, 2015), less likely to endorse pseudo-profound bullshit (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015), less influenced by emotion or disgust (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014b), and less willing to cause harm to if it does not benefit a greater good (Byrd & Conway, 2019).

3.1.1 Some Problems

Some reflection tests are mathematical (e.g., Frederick, 2005), some are logical (Janis & Frick, 1943), and some are not obviously mathematical or logical (e.g., Thomson & Oppenheimer, 2016; Sirota, Kostovičová, Juanchich, Dewberry, & Marshall, 2018). So, some measures of reflection confound reflection with other domain-specific skills. For example, mathematical reflection tests may partially confound reflection with—among other things—numeracy (e.g., Campitelli & Labollita, 2010).

Moreover, reflection tests confound reflection with reaction time (e.g., Stupple, Pitchford, Ball, Hunt, & Steel, 2017) and intuitive bias (e.g., Pennycook, Cheyne, Koehler, & Fugelsang, 2015). As a result, common measures of reflection do not necessarily measure reflection *per se*. They also measure some combination of supporting abilities. Likewise, the correlates of reflection do not necessarily correlate with reflection *per se*. They might also correlate with some of reflection's supporting abilities. So, there are ambiguities about what is measured by reflection tests and what correlates with reflection test performance. These ambiguities reinforce growing concern about reflection tests and its correlates (e.g., Byrd & Conway, 2019; Rouder, Kumar, & Haaf, 2018).

3.1.2 Some Solutions

This paper points to two methods that can validate and improve measurement of reflection: verbal report protocols (Ericsson & Simon, 1998) and process dissociation (Jacoby, 1991). Of course, these two methods need not be deployed on all reflection tests. So, this paper also explains when and why verbal report protocols and process dissociation should be implemented.

3.2 Opportunities To Improve

Reflection is a staple of good reasoning in the history of ideas. For instance, philosophers have thought that reflection is necessary for genuine moral judgment (Kennett & Fine, 2009), justice (Rawls, 1971), knowledge (Sosa, 1991), logical induction (Goodman, 1983), normativity (Korsgaard, 1996), practical reasoning (Velleman, 1989), self-knowledge (Peacocke, 2014), and self-scrutiny (Hursthouse, 1999).

Philosophers often describe reflection with folk psychological metaphors. For example, "I find myself with a powerful impulse to believe. But I back up and bring that impulse into view

and then I have a certain distance. Now the impulse doesn't dominate me" (Korsgaard, 1996, p. 93). Nonetheless, philosophers' notions of reflection can be mapped on to scientists' explications of reflection. Consider the famous bat-and-ball problem: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" (Frederick, 2005). Participants often impulsively report a particular wrong answer, detect some sort of conflict, and then begin reflecting (Chapter 2): "Let's see! $1.10 minus $1 is 10 cents… Wait. that's wrong! This should be solved as an equation…" (Szaszi, Szollosi, Palfi, & Aczel, 2017, p. 218). So many psychologists describe correct responses on reflection tests with 'reflective'. Likewise, many psychologists describe lured responses (e.g., "10 cents") with 'unreflective' (Pennycook, Cheyne, Koehler, & Fugelsang, 2015).

Alas, there is more to reflection than metaphorical or empirical explications. After all, it is not at all clear what it means to 'back up' and 'bring an impulse into view'. Moreover, some people seem to be able to arrive at the correct answer to problems like the bat-and-ball problem without any reflection: "It's 5 cents!" (Szaszi, Szollosi, Palfi, & Aczel, 2017, p. 218; see also Bago & De Neys, 2019). Also, while reflection tests do not fully reduce to tests of other constructs (Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Patel, 2017; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016), reflection tests do partially confound reflection with other factors such as numeracy (Campitelli & Gerrans, 2014) and general cognitive ability (Toplak, West, Stanovich, 2011). So although our understanding of reflection tests is not entirely wrong, it is clearly incomplete.

3.3 Indirect Measures of Reflection

There are various measures of reflection with important differences. However, all measures of reflection share two design features. First, each of their items is designed to be

straightforwardly solvable with a moment's reflection. However, and second, each of their items is designed to lure participants toward a particular incorrect response. In short, measures of reflection are designed to pit unreflective responses against reflective responses.

3.3.1 Indirect Logical Measures of Reflection

Some reflection tests are logical. A typical logical reflection test item is a syllogism—i.e., two premises and a conclusion. Logical reflection tests typically ask participants if a particular conclusion follows from two premises. The conclusions in question are either believable or unbelievable, so that participants will be lured into (incorrectly) responding according to the believability of the conclusion rather than the logical validity of the syllogism. Evaluating syllogisms according to believability rather than logical validity is known as belief bias.

Belief bias tests. People do not tend to reflect on the logical validity of syllogisms. Instead, they tend to endorse syllogisms based on the believability of their conclusions (e.g., Janis & Frick, 1943). Revealing this belief bias involves up to four kinds of syllogism (Table 2).

Table 2. Four kinds of syllogism in Belief Bias tests.

Incongruent	**Congruent**
Believable and invalid	Believable and valid
Unbelievable and valid	Unbelievable and invalid

Incongruent syllogisms. Two kinds of syllogism pit unreflective responses against reflective responses. We can call these incongruent syllogisms because their believability is incongruent with their logical validity. The first kind of incongruent syllogism is believable, but

invalid. For example, “All flowers have petals. Roses have petals. If these two statements are true, can we conclude from them that roses are flowers?” (Markovits & Nantel 1989). Our prior beliefs about roses produce a strong impulse to accept the conclusion. However, a moment’s reflection shows that, logically, we should not conclude from these premises that roses are flowers. The second kind of incongruent syllogism is unbelievable, but valid. For example, “All vehicles have wheels. Boats are vehicles. If these two statements are true, can we conclude from them that boats have wheels?” (De Neys & Franssens, 2009). Our prior beliefs about boats produce a strong impulse to reject the conclusion that boats have wheels. However, a moment’s reflection shows that, logically, we should infer that conclusion from the premises.

Congruent syllogisms. The other kinds of syllogism in belief bias tests are not, by themselves, measures of reflection because they do not pit unreflective responses against reflective responses. In these syllogisms, reflective and unreflective responses are congruent because believability and logical validity are congruent. One kind of congruent syllogism is believable and valid. For example, “All business owners are rich. Bill Gates is a business owner. If these two statements are true, can we conclude from them that Bill Gates is rich?” (Baron, Scott, Fincher, & Metz, 2015). Our prior beliefs about Bill Gates produce a strong impulse to accept that conclusion and, logically, we should draw that conclusion. Another kind of congruent syllogism is unbelievable and invalid. For example, “All cats are furry. Rabbits are furry. If these two statements are true, can we conclude from them that rabbits are cats?” (ibid.). Our prior beliefs about cats and rabbits produce a strong impulse to reject that conclusion and, logically, we should reject that conclusion. So, we are likely to respond correctly to congruent syllogisms regardless of whether we reflect about them.

Syllogisms without lures. One might want a control condition in which participants evaluate syllogisms without prior beliefs about certain words biasing responses. This can be accomplished with pseudowords—a.k.a., nonsense or "wug" words (Berko, 1958). An example of a syllogism with such pseudowords is as follows: "All laloobays are rich. Sandy is a laloobay. If these two statements are true, can we conclude from them that Sandy is rich?" (Baron, Scott, Fincher, & Mets, 2015). Even though we have no strong impulse to believe anything in particular about lalloobays, a moment's reflection will remind us that the conclusion follows logically from the premises.

3.3.1 Indirect Mathematical Measures of Reflection

Some widely used measures of reflection employ math problems. These mathematical measures of reflection often take the form of mathematical syllogisms: they contain two mathematical premises, and they ask participants about what follows from them. And, like other measures of reflection, these math problems usually employ lures that pit unreflective responses against reflective responses.

Base rate neglect. People often do not reflect on relevant information in straightforward math problems (Bar-Hillel, 1980). The evidence for this is based, in part, on responses to basic questions about probability.

Incongruent base rates. When participants are explicitly told that in a sample of 100 people, 30 are engineers and 70 are lawyers, they systematically ignore these prior probabilities when asked about the probability that any one individual is an engineer—but only when the individual is described as a stereotypical engineer. That is, people tend to account for the 30-70 base rates when they receive this prompt: "Suppose now that you are given no information whatsoever about an individual chosen at random from the sample. The probability that this man

is one of the 30 engineers in the sample of 100 is ____ %" (Kahneman & Tversky, 1973, p. 241). Most people correctly determine the probability that this random individual from the sample is an engineer—i.e., 30%. However, people tend to ignore the 30% base rate of engineers when given individuating information about someone in the population, such as the following (Kahneman & Tversky, 1973, p. 241).

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> The probability that Jack is one of the 30 engineers in the sample of 100 is ____ %

Most people tend to overestimate the probability that Jack is an engineer relative to the 30% base rate of engineers, $t(169) = 3.23$, $p < 0.01$ (Kahneman & Tversky, 1973, p. 241).

So, when people have no prior beliefs about an individual, they correctly use base rates to determine the probabilities of the individual's profession. However, when people have prior beliefs about both an individual and about the relevant base rates, their prior beliefs about the individual produce a powerful impulse to determine the probability of an individual's profession according to how well the individual's description represents a stereotype of that profession—i.e., the representativeness bias (Kahneman & Tversky, 1974). Of course, a moment's reflection will remind us that the probability that Jack is one of the engineers in the sample cannot be higher than the 30% base rate of engineers in that sample.

Congruent base rate items. Sometimes base rates and individuating information are congruent. That is, a non-stereotypic description of an engineer is congruent with a low base rate of engineers in a population. When there is no such conflict between base rates and individuating

information, people determine probabilities more accurately and quickly (Pennycook & Thompson, 2012; Pennycook, Trippas, Handley, & Thompson, 2014). Like congruent belief bias tests, this indicates that reflection may not be needed when an unreflective bias does not conflict with prior beliefs. Also, like congruent belief bias tests, congruent base rate problems are not, by themselves, tests of reflection because they do not involve conflict between reflective and unreflective responses.

Cognitive reflection test. People also tend not to reflect about relevant information in more basic math problems. The evidence for this comes from a short and widely-used math test known as the Cognitive Reflection Test (Frederick, 2005).

Incongruent CRT. The original cognitive reflection test (CRT) was designed to lure test-takers into a particular incorrect response to math problems. Recall its famous bat-and-ball problem , "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" (Frederick, 2005, p. 26). This question lures participants into answering "10 cents". However, a moment's reflection shows that this response cannot simultaneously satisfy both of the premises of the problem.

CRT without lures. Like other tests of reflection, there are versions of the test that do not lure participants toward particular incorrect responses. For instance, "A bat and a ball cost 96 cents in total. The bat costs 2 cents more than the ball. How much does the ball cost?" (Baron, Scott, Fincher, and Metz, 2015). A correct response to this item probably requires some reflection from the average research participant. However, the item is less likely to require the participant to overcome a particular incorrect response than its lured counterpart, given that there are significantly more unique incorrect responses on the non-lured version than the lured version (Byrd & Conway, 2019).

3.3.2 Other Measures Of Reflection

Other measures of reflection employ neither logic nor math, but they are nonetheless straightforward cognitive tasks that pit unreflective responses against reflective responses. There are well-known examples of this like the Stroop Task (Table 3) as well as more recent items.

Table 3. Examples of a congruent and an incongruent Color Stroop Task

Congruent List	Incongruent List
Red	Blue
Blue	Purple
Green	Green
Purple	Red
Blue	Green

Stroop task. The Stroop Task is a visuo-verbal task which requires participants to inhibit or reject an autonomic response (Stroop, 1935; cf. Comalli, Wapner, & Werner, 1962). Participants are instructed to report the color names in a list of color words. Importantly, some color words are congruent with their font color—e.g., if the font color of 'red' is red—and some color words are incongruent with their font color—e.g., if the font color of 'red' is purple. Most people have a strong impulse to say the word itself rather than the color of words' font. Of course, one can overcome this impulse with a moment's reflection. So success on the Stroop Task seems to require reflection even though it is not obviously a mathematical or logical task.

Miscellaneous measures of reflection. There are probably many ways to measure reflection with tasks that are neither transparently mathematical nor transparently logical. The

structure of these questions would be as follows: they would refer participants to relevant information that participants are lured into ignoring (e.g., a premise, a base-rate, a color word, etc.). For example, “Ann’s father has a total of five daughters: Lala, Lele, Lili, Lolo, and ____. What is the name of the fifth daughter?” (Krizo, 2012; Sirota et al., 2018). The pattern in this list names provokes a strong impulse to think of ‘Lulu’, however, a moment’s reflection reminds us that Ann must be Ann’s father’s fifth daughter. Jonathan Baron and Edward Royzman have devised similar measures of reflection that do not seem to require logic or math:

> On the side of a boat hangs a ladder with six rungs. Each rung is one foot from the next one, and the bottom rung is resting on the surface of the water. The tide rises at a rate of one foot an hour. How long will it take the water to reach the top rung?
>
> ☐ 5 hours
>
> ☐ 6 hours
>
> ☐ Never

This item produces a strong—and potentially mathematical—impulse to think that it will take “5 hours” for the water to reach all of the next 5 ladder rungs. However, a moment’s reflection about the relationships between boats and tides reminds us that boats rise with tides—so, the water will never reach the top rung. In both examples, the reflection that leads to the correct response is not transparently mathematical or logical—even if it can be formalized mathematically and logically after the fact.

3.4 Verbal Report Protocols

One way to get more information about how people reason is to collect verbal reports. Verbal reports have been used in empirical studies of how people complete tasks since as early

as the first half of the 20th century, before the time of audio recorders (e.g., Bulbrook, 1932; Duncker, 1926; Maier, 1931; Watson, 1920). In the second half of the 20th century, the use of verbal reports to understand human reasoning was challenged when researchers repeatedly found that verbal reports often bely peoples' actual reasoning processes, as indicated by controlled features of peoples' environment and by people's behavior (e.g., Goethals & Reckman, 1973; Johansson, Hall, Sikström, & Olsson, 2005; Nisbett & Bellows, 1977; Nisbett & Wilson, 1977b; Rosenfeld & Baer, 1969; Wilson & Nisbet, 1978; cf. Lieberman, 1979; Petitmengin, Remillieux, Cahour, & Carter-Thomas, 2013; Smith & Miller, 1978; White, 1980). In the late 20th century and early 21st century, evidence suggested that these problems with introspection and verbal reports are overcome by—among other things—asking people to *narrate* or *recall* rather than *explain* their decisions (Ericsson & Simon, 1980, 1984; Ericsson, 2018).

So what do verbal reports have to do with reflection tests? Recall that reflection is more consciously represented and deliberate processing (Chapter 1). And recall that conscious representations can be explicated in terms of what one can articulate (ibid.). Thus, by definition, what someone reports verbally when asked to think aloud, is a conscious representation. So, whatever participants report verbally (and deliberately) while completing a task is, by definition, some evidence not only of reflective reasoning, but the contents thereof. In other words, verbal reports are a uniquely useful source of data about whether and how people reflect during a task.

3.4.1 Verbal Reports Reveal Nuance

The aforementioned non-verbal measures of reflection are indirect. They measure and quantify differences in behavior which indicate, indirectly, whether (or how much) someone was reflecting on a task. One benefit of indirect measures of reflection is that they do not necessarily direct individuals' attention to the fact that reflection is being measured. More direct measures—

such as concurrent verbal reports can draw participants' attention to the fact that reflection is being observed. Nonetheless, participants' verbal reports may reveal something that indirect measures cannot.

Graded rather than mere categorical differences. One thing that verbal report protocols might reveal is gradation in reflection. For example, some people might consciously represent more than what others consciously represent. One way to detect this would be to test whether some participants verbally report more reasoning content than other participants. So, verbal reports offer a way to empirically distinguish degrees or extensiveness of reflection. This might have benefits over many reflection tests' artificially categorical distinction between so-called reflective and unreflective responses.

A categorical distinction proposes a definite boundary between two concepts, whereas a comparative distinction merely proposes a relative difference between two concepts (Carnap 1950, Section 3 to 8). So, a categorical distinction between reflective and unreflective reasoning would imply that reflection (or its measurement) is an all-or-nothing affair: A response to a reasoning task either involves reflection or it does not.

However, a comparative distinction between reflective and unreflective reasoning allows for the more nuanced possibility that one response can involve more or less reflection than other responses. For instance, two participants may respond correctly—and, by definition, reflectively—on various tasks that indirectly measure reflection. However, one participant might consciously represent more of the information. Or they might represent the same information but process it more deliberately. For example, one might double-check their solution on a test of reflection by solving the problem a second time with another method, thereby representing up to two times as much content as the person who solves the problem only once. Alternatively, one

conscious representation of a problem might have more content than another conscious representation of a problem (See "Representation" in Table 4). By recording only participants' test responses, one ignores the potential individual differences in relative reflection that might explain otherwise puzzling patterns in data.

Table 4. Two reflective, but non-identical representations and responses to the bat-and-ball problem: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"

Representation	Premise 1: x + y = $1.10 Premise 2: y = x + $1.00 Conclusion: x = ?	x + (x+ $1.00) = $1.10
Processing	Conjecture 1: x = $0.10 […] Conjecture *N*: x = $0.05	2x = $1.10 – $1.00 2x = $0.10 x = $0.05

False negatives. In many tests of reflection, participants might arrive at so-called unreflective responses even though they consciously represented and deliberately processed some relevant information (e.g., Hoover & Healy, 2019). For instance, one can commit the base rate fallacy by consciously representing and deliberately processing all relevant information, except the relevant base rate. According to standard protocols for measuring reflective reasoning indirectly, responses that neglect such relevant information are, by definition, unreflective. In fact, these responses are merely less reflective than the so-called reflective response. So, standard coding procedure is vulnerable to this false negative problem in which reflective responses are coded as unreflective. Verbal reports have the potential to show that seemingly unreflective (i.e., particular incorrect) responses actually involved some reflection.

Skill. False negatives may also be the result of the fact that many measures of reflection are confounded with domain-specific skills. For example, the CRT is designed to be a measure of reflective disposition, but even its designer expected it to confound reflection with "reading comprehension and mathematical skills" (Frederick, 2005, p. 35). Many researchers have been concerned about the latter confound: numeracy (Campitelli & Gerrans, 2014). Numeracy is often described as a disposition to comprehend and process mathematical information (Reyna, Nelson, Han, & Dieckmann, 2009). So, insofar as measures of reflection require familiarity with mathematics, those measures can confound reflection with numeracy. Fortunately, plenty of research shows that while mathematical measures of reflection often correlate with numeracy, they may not be mere numeracy tests (Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Patel, 2017; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; cf. Erceg, Galic, & Ružojčić, 2020; Patel, 2017).

Verbal reports have been crucial to understanding whether and how numeracy could play a role in poor performance on tests of reflection (e.g., Szaszi, Szollosi, Palfi, & Aczel, 2017). Consider the verbal report of one participants attempt to solve the bat-and-ball problem: "[10 cents] …but I'm not sure… If together they cost $1.10 and the bat costs $1 more than the ball, the solution should be 10 cents. I'm done" (Ibid, p. 218). According to standard protocols for measuring reflective reasoning indirectly, this '10 cents' response is, by definition, unreflective. However, this participant's verbal report shows that their '10 cents' response is clearly reflective: The participant inhibited their initial impulse—"but I'm not sure"—and reported reflecting about all of the relevant mathematical information—"together they cost $1.10 and the bat costs $1 more". Nonetheless, the participant processed the mathematical information incorrectly. So, this

verbal report shows that the standard coding procedure of labeling this particular '10 cents' response produces a false negative.

False positives. The aforementioned indirect measures of reflection also have a false positive problem. For instance, some participants are so familiar with a test that they can perform well on it without reflecting, thus leading to falsely coding an unreflective response as a positive case of reflection.

Test familiarity. Some measures of reflection like the CRT are so widely used that one might worry about participants becoming so familiar with the test that they memorize the answers and thereby change what the test measures (Chapter 4). To prevent this familiarity problem, many versions of the test have been created (Ackerman & Zalmanov, 2012; Finucane & Gullion, 2010; Baron, Scott, Fincher, & Metz, 2015; Oldrati, Patricelli, Colombo, & Antonietti, 2016; Primi, Morsanyi, Chiesei, Donati, & Hamilton 2016; Shtulman & McCallum, 2014; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2014; Trémolière, De Neys, & Bonnefon 2014). Later research found that the familiarity problem does not completely invalidate the CRT: performance on the CRT was robust even after repeated exposure to the test (Białek & Pennycook, 2018; Meyer, Zhou, & Frederick, 2018; Stagnaro, Pennycook, & Rand, 2018; Welsh & Begg, 2017).

Nonetheless, consider how verbal report could reveal whether test familiarity has produced false positives. Imagine that a participant says the following when presented with the famous bat-and-ball problem: "Oh, this is that trick question about the bat and the ball. This one fooled me the first few times I saw it. Then I looked up the answer and memorized it. The answer is '5 cents'." According to standard protocols for indirectly measuring reflective reasoning, the '5 cents' response is, by definition, reflective. In fact, this participant's '5 cents'

response is unreflective. Their verbal report clearly illustrates that the most important aspects of the problem were not consciously represented and no impulse was deliberately inhibited. Rather, the participant's answer was simply retrieved from memory. So, standard coding procedure would label this unreflective response as reflective—a false positive. Thus, by recording participants' verbal reports, one can determine with better accuracy whether seemingly reflective responses—e.g., '5 cents'—are the result of actual reflection or mere test familiarity.

Double trouble. Of course, the test familiarity problem is more than a thought experiment. Test familiarity has been shown to predict performance on measures of reflection as well as other variables that commonly predict such performance (e.g., Chapter 4). So, if variance in reflection performance is shared not only by factors of interest—like religiosity (e.g., Pennycook, Ross, Koehler, & Fugelsang, 2016) or moral judgment (Cova et al., 2018)—but by test familiarity, then research on reflection in the absence of verbal report has two false positive problems: falsely positive measurements of reflection and falsely positive correlations with reflection.

Differences in representational content. Even if two people solve one and the same problem mathematically, one might represent different information than the other. For instance, one might represent the bat-and-ball problem logically. To do this, one might imagine a pair of mathematical premises involving two unknown variables and then test conjectures about the variables by trial-and-error, until a conjecture that forms valid mathematical syllogism is found (Table 4, left column). Alternatively, one might represent the bat-and-ball problem as an algebraic equation with only one unknown variable and then solve for that variable (Table 4, right column). While these differences in representation and processing are not detected with standard protocols for measuring reflection, they might be detected via verbal report.

3.4.2 Reactive vs. Non-reactive Verbal Report Protocols

Think aloud protocol analyses have not only improved various measures, but also revealed some of the damning evidence about introspection. For instance, people are more likely to forget or even fabricate during posthoc or explanatory verbal reports than concurrent or recollective verbal reports (e.g., Ericsson & Simon, 1998; Russo, Johnson, and Stephens, 1989). A famous criticism of introspection and verbal report is motivated by only posthoc explanatory verbal reports (e.g., Nisbett & Wilson, 1977a). So, reflection research that treats verbal reports as data might pre-empt common problems by adopting better verbal report protocols.

Reactive verbal report protocols. There are many cases of poor *post hoc* introspective verbal report involving the evaluation of consumer goods. For example, in one study, passersby are asked which of a set of identical products was the "best quality" (Nisbett & Wilson, 1977a, p. 243). After making their choice, participants were asked to explain the reason for their choice. Although no participants reported basing their decision on the location of the product and some even denied it when asked directly, most participants chose the "right-most" item—even though it was identical to the items to its left (Nisbett & Wilson, 1977b, pp. 243-44). Findings like these have convinced some that people are not only unaware of their own reasoning processes, but prone to fabricate demonstrably false explanations of their reasoning processes. Naturally, this kind of finding casts doubt on the accuracy of verbal reports and the legitimacy of treating verbal reports as data. Of course, these findings do not settle the question of whether the illegitimacy of verbal reports is a necessary feature of all verbal reports or merely an accidental feature of particular verbal report protocols.

Non-reactive verbal report protocols. The idea of discrediting all verbal reports might strike some people as so absurd that they need no more than hear the idea in order to reject it.

Others, however, might want to investigate the matter empirically. Specifically, they will want to know if there are any circumstances in which judgments and decisions are well explained by verbal reports rather than contradicted by them (á la Nisbett et al.).

Early on, psychology researchers asked participants to "think aloud" while solving tasks (e.g., Benjafield, 1969). However, early think aloud studies were often under-powered. So one might want to hear about evidence from larger think aloud studies. Fortunately, larger think aloud studies have found consistency between behavior and verbal report. For example, a meta-analysis of 94 studies and over 3500 participants found that participants' behavior was not altered by giving verbal reports (Fox, Ericsson, & Best, 2011). This meta-analysis also revealed a potential explanation of past discrepancies between behavior when participants gave verbal report and when they did not: asking participants' to *explain* their decisions had a significant impact on participants' task performance while merely asking participants to *report* their reasoning while deciding had no impact on participants' task performance—even though all participants reported concurrently. This indicates that asking people to concurrently *explain* their reasoning is not the same as asking them to concurrently *report* their reasoning.

Research also reveals a similar distinction in retrospective think aloud reports. Participants have been more credibly aware of how their decisions were manipulated by initially unnoticed factors when they were asked to *remember how* they made a decision than when they were asked to *explain why* they made a decision (e.g., Petitmengin, Remillieux, Cahour, & Carter-Thomas, 2013).

So recent and well-powered findings indicate that the seeming invalidity of verbal reports is not a necessary feature of verbal reports, but of particular verbal report protocols. After all, concurrent think aloud verbalizations and carefully executed retrospective think aloud

verbalizations have been consistent with and illuminating of corresponding task performance (Ericsson, 2003). The key to successful verbal report protocols, then, seems to be not asking participants to *explain why* they made a particular decision, but asking participants to *narrate how* they decide in real time or else to *recall what* they remember thinking during their decision process (Ericsson & Simon, 1984).

3.4.3 Verbal Reports Can, But Do Not Necessarily Improve Measures of Reflection

The take-away for studying reflection and its correlates, then, is that verbal reports are demonstrably capable of revealing nuances about whether and how people reflect that common measures of reflection do not detect. Nonetheless, revealing this nuance can beget misleading results if demonstrably inferior verbal report protocols are used.

3.5 Process Dissociation

Recall measures of reflection reviewed so far: first, task-based measures of reflection—mathematical, logical, etc.—and, second, measures of the verbal reports that accompany these tasks. The latter measures have helped to reveal nuance in the data gathered by the former measures. One might wonder if other methods can detect additional nuance about whether and how people reflect. In what remains, I will explain how process dissociation is one such method.

Process dissociation (PD) is useful for untangling interactions between multiple processes that might jointly contribute to behavior (Payne & Bishara, 2009). The procedure was developed to dissociate automatic uses of memory from more intentional uses of memory (Jacoby 1991). However, the procedure is content agnostic. So, for example, PD has also been used to dissociate various processes involved in implicit bias (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005) and moral reasoning (e.g., Conway & Gawronski 2013; Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017).

3.5.1 Most Measures of Reflection Are Already Well-Suited For Process Dissociation

Recall that many measures of reflection involve congruent and incongruent items—e.g., belief bias syllogisms in which the believability of its conclusion is either congruent or incongruent with the logical validity of the syllogism. PD analyzes the performance differences across both congruent and incongruent items to reveal how two processes (e.g., belief bias vs. reflection) independently contribute to reasoning. This involves presenting people with one set of tasks in which the two processes elicit the same kind of response (e.g., congruent belief bias syllogisms) and another set of tasks in which the two processes elicit different responses (e.g., incongruent belief bias syllogisms). Each participants' response pattern for congruent items and incongruent items can be turned into two frequentist probabilities. For example, the frequency with which participants judged congruent, invalid syllogisms to be invalid and incongruent valid syllogisms to be valid (i.e., the reflective parameter) and the frequency with which participants judged both congruent, invalid syllogisms and incongruent valid syllogisms to be invalid (i.e., the unreflective parameter). These independent parameters for each process are visualized with the processing tree in Figure 3. Process dissociation could add two more columns of response patterns to this processing tree (e.g., both "congruent, valid" and "incongruent, invalid"), but these have been left out for the sake of space.

3.5.2 Dissociating Responses To The CRT

Existing research has already fruitfully deployed process dissociation on mathematical tasks like the CRT (e.g., Ferreira, Mata, Donkin, Sherman, & Ihmels, 2016). For example, recall the CRT's bat-and-ball problem: "A bat a and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" In this task, reflective and unreflective processing are incongruent: we have an impulse toward a particular response that, upon reflection, is wrong.

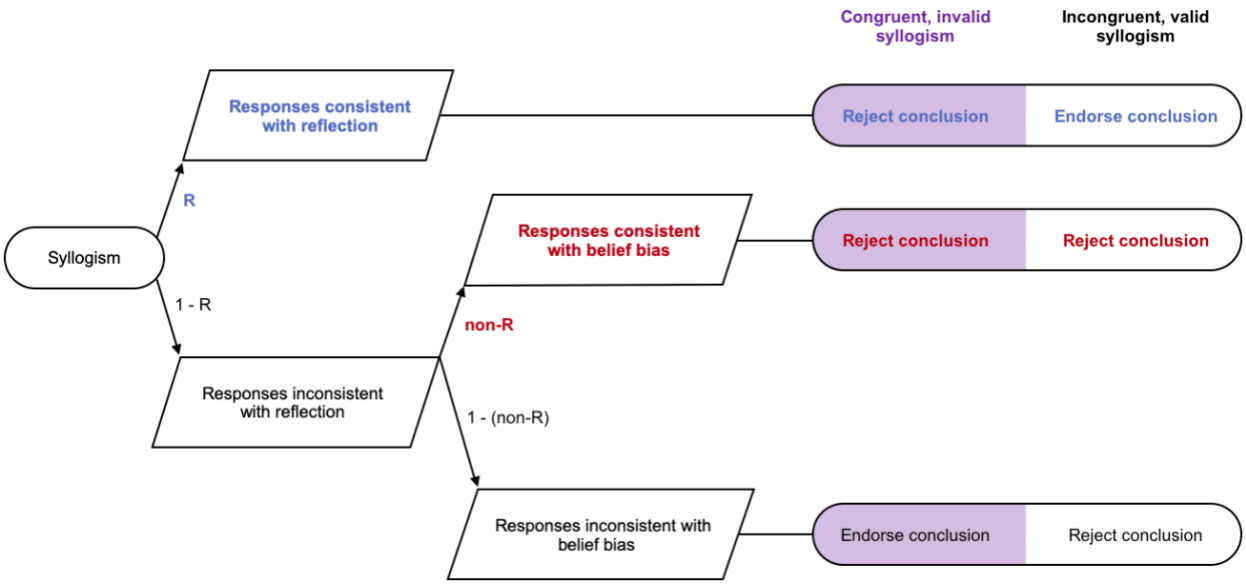

Figure 3. An example of process dissociation employed on logical measures of reflection. The processing tree illustrates the reflective and unreflective processes underlying responses to congruent, invalid and incongruent, valid syllogisms.

A congruent version of the original bat-and-ball problem would be designed to lure participants toward the correct response rather than a particular incorrect response: "A bat and a ball cost $1.10 in total. The bat costs $1.05. How much does the bat cost?" (adapted from McPhetres, 2018). This congruent version of the bat-and-ball problem is so straightforward that it can be solved without much, if any, reflection—even though it can, in principle, be solved via careful reflection. And so, on the congruent version of the bat-and-ball problem "5 cents" is consistent with both reflective and unreflective processing.

By applying PD to both congruent and incongruent versions of CRT items, one can derive parameters of reflective and unreflective processing. To do so would involve calculating frequentist probabilities for each participant using a processing tree. Mathematically, the parameters can be calculated using formulas 1 and 2 below (adapted from McPhetres, 2018).

$$R = p(\text{unreflective answer} \mid \text{congruent}) - p(\text{unreflective answer} \mid \text{incongruent}) \quad (1)$$

$$\text{Non-R} = p(\text{unreflective answer} \mid \text{incongruent}) / (1\text{-R}) \quad (2)$$

Thus, a participant who responded unreflectively on 6 out of 10 congruent CRT items (0.6) and unreflectively on 3 out of 10 incongruent CRT items (0.3) would have an R parameter of 0.3 and a non-R parameter of 0.43 (adapted from McPhetres, 2018):

$$R = 0.6 - 0.3 = 0.3$$

$$\text{Non-R} = 0.3 / (1\text{-}0.3) = 0.43$$

Notice that these parameters do not add up to 1. So, they are not perfectly anti-correlated. Indeed, they are often only weakly correlated (McPhetres, 2018). This is supposed to be one of the rationales of PD: it has the potential to provide independent parameters of otherwise dependent parameters. However, the CRT already had the potential to provide independent parameters of reflective and unreflective processing without the help of PD. So, one might wonder whether PD versions of the CRT are overall better than previous CRT measures of reflection.

Prior CRTs already confer independent parameters. Reflective and unreflective responses were already assessed independently on the original CRT. This is because, unlike many other measures of reflection (e.g., logical measures), CRT responses are not binary. Rather, there are an infinite amount of possible responses for every CRT item. Only two of the infinite possible responses are coded. That is, reflective responses (e.g., "5 cents") and unreflective responses (e.g., "10 cents") on the original (i.e., incongruent) items are distinguishable from all other possible responses to the item. As a result, the frequency of reflective responses and unreflective responses, while often anti-correlated, are not perfectly anti-correlated (Chapter 4; Byrd & Conway, 2019). Reflective and unreflective responses on the

CRT are only perfectly anti-correlated when CRT responses are erroneously coded—e.g., if not just one particular incorrect response per item is coded as unreflective, but all possible incorrect responses per item are labeled unreflective. When coded correctly, the incongruent versions of the CRT provide independent parameters of reflective and unreflective processing. So, one might wonder whether the PD version of the CRT has any net benefit over existing (incongruent) versions of the CRT.

Costs and benefits of using process dissociation on the CRT. Using the PD version of the CRT affords a few benefits. First, it provides another way of obtaining independent parameters for reflective and unreflective processing. Second, it provides new CRT items, which might help overcome the aforementioned familiarity problem of the CRT (Chapter 4; McPhetres, 2018). One cost, however, is that the PD version of the CRT requires at least twice as many items—the version being presented herein originally contained 28 items (McPhetres, 2018). So, given that the existing versions of the CRT—including the original 3-item version (Frederick, 2005)—already allow independent parameters to be calculated and that CRT performance seems to be robust to repeated exposures (Białek & Pennycook, 2018; Meyer, Zhou, & Frederick, 2018; Stagnaro, Pennycook, & Rand, 2018; Welsh & Begg, 2017), the PD approach will need to justify its use of more CRT items. Nonetheless, some have argued that scoring reflection tests by merely summing unreflective responses and reflective responses treats reflection test scores as factors, which may not be appropriate in some cases (McNeish & Wolf, 2019). In those cases, PD could be a more appropriate alternative.

3.5.3 Process Dissociation Can, But Does Not Necessarily Improve Measures of Reflection

The take-away for studying reflection is that PD can but does not necessarily reveal additional nuance about whether and to what degree people reflect while completing tests of

reflection. For binary measures of reflection—i.e., measures whose responses can be exhaustively classified as either reflective or unreflective—reflective and unreflective responses are necessarily perfectly anti-correlated. So, PD can reveal additional nuance in binary measures of reflection by introducing the possibility of more than two response patterns and thereby, the possibility of assessing reflective and unreflective processing independently.

However, for non-binary measures of reflection—i.e., measures whose response patterns cannot be exhaustively classified as either reflective or unreflective—reflective and unreflective responses are not necessarily perfectly anti-correlated. As a result, non-binary measures of reflection are already capable of independently assessing reflective and unreflective processing. So, prior to further investigation, it is not yet clear that PD always reveals additional nuance—or reveals additional nuance more efficiently—than existing non-binary measures of reflection.

3.6 Conclusion

There are many ways to measure whether and how people reflect during a task. Research would do well to acknowledge the limitations of each of these measures such as how many widely used measures of reflection partially confound other factors—e.g., numeracy—with reflection and how many measures of reflection are prone to false positives and false negatives, depending on participants familiarity with the measure—e.g., the original 3-item CRT—or its content—e.g., logic, math, etc. One way that research can overcome the limitations of widely used measures of reflection is to collect verbal reports and employ process dissociation. However, not all verbal report protocols are appropriate and not all measures of reflection need to be revised to allow for process dissociation. Evidence currently recommends think aloud protocols and elicitation interviews to reduce the rates of falsely positive and falsely negative assessments of reflection. Likewise, evidence is currently compatible with the recommendation

to create process dissociation measures of only binary measures of reflection since non-binary measures of reflection already allow for process dissociation. With all of these methods combined, there are many possible ways to measure whether and how people reflect. The current chapter shows how not all of these measures are created equal.

CHAPTER 4
GREAT MINDS DO NOT THINK ALIKE

Many philosophers already accept that both intuition and reflection play crucial roles in philosophy. On the one hand, philosophers admit that they appeal to intuition to motivate some of the premises in their arguments (e.g., Chalmers, 2014; Climenhaga, 2018; De Cruz, 2014; Kornblith, 1998; Mallon, 2016). Indeed, some philosophers think that intuitions are *prima facie* justified (e.g., Huemer 2006, 2007). On the other hand, reflective thinking is considered both an important part of philosophical inquiry (e.g. Rawls, 1971) and an essential dimension of philosophically significant topics such as knowledge (Sosa 1991), logic (Goodman 1983), metaethics (Hursthouse, 1999; Korsgaard, 1996; Sidgwick, 1874/1962), and self-knowledge (Peacocke, 2014), to name a few. So both unreflective intuition and more reflective judgment feature centrally in philosophy.

The difference between more reflective and less reflective (or intuitive) reasoning has also been studied extensively by experimental psychologists and economists, who have developed various tasks that distinguish more reflective reasoning from less reflective reasoning. These reflection tests have two key features (Byrd, *under review*). First, they lure participants toward a particular incorrect response. Second, they can be solved correctly with a moment's reflection. For example, if a bat and a ball cost $1.10 in total and the ball costs $1.00 more than the ball, then how much does the ball cost? This well-known bat-and-ball problem lures people toward incorrectly answering "10 cents" even though a moment's reflection can reveal that the correct answer is "5 cents" (Frederick, 2005).

Some evidence suggests that cognitive reflection tests measure a domain-general disposition. For instance, people who reason more reflectively about math have also reasoned more reflectively about logic (Byrd & Conway, 2019), Newtonian physics (Gette & Kryjevskaia, 2019), and moral side effects (Pinillos, Smith, Nair, Marchetto, & Mun, 2011). Likewise, people who reason less reflectively have exhibited other philosophical tendencies such as atheism (Pennycook et al., 2016; cf. Gervais et al., 2018), social conservatism (e.g., Deppe et al., 2015; Yilmaz & Saribay, 2016; Yilmaz, Adil, & Iyer, 2019; cf. Alper & Yilmaz, 2019; cf. Price-Blackshear, Sheldon, Corcoran, & Bettencourt, 2019), and acceptance of utilitarian sacrifices (Patil, et al, 2020; Reynolds, Byrd, & Conway, *in prep.*; cf. Byrd & Conway, 2019). Hence, dispositions to reflect may be general enough to explain individual differences in philosophical beliefs.

However, the degree to which reflection correlates with philosophical beliefs is not yet well understood empirically. While the link between reflection and philosophical judgment seems fairly robust (Pennycook, Ross, Koehler, & Fugelsang, 2016; Reynolds, Byrd, & Conway, *in prep.*; Reynolds, Makhanova, Ng, & Conway, 2019), it is not found in all populations (e.g., Gervais et al., 2018). One population whose beliefs and reflection test performance has yet to be thoroughly studied is academic philosophers. Also, existing research often focuses on judgments about particular thought experiments rather than general philosophical beliefs. Alas, subtle differences in wording of thought experiments can cue readers toward certain judgments in unexpected ways (Cullen, 2010; Fischer & Engelhardt, 2019). For example, judgments about free will varied depending on whether thought experiments included ‘psychologists’, ‘mind’, and ‘thoughts’ vs. ‘neuroscientist’, ‘brains’, and ‘neural processes’ (Nahmias, Coates, and Kvaran, 2007). So it is possible that intuitions about particular philosophical thought experiments do not

track the more considered philosophical beliefs that may have motivated the thought experiments (Mandelbaum & Ripley, 2012; Sinnott-Armstrong, 2007).

This chapter employs empirical methods to address these gaps in the literature (Knobe & Nichols, 2007). Two studies report how reflection test performance correlated with philosophers' philosophical beliefs. The main finding is that many links between reflection and philosophical beliefs tendencies among non-philosophers replicate among philosophers, but some of these links are better explained by factors such as culture, education, gender, and personality.

4.1 Introduction

Whereas many philosophical beliefs have correlated with reflection test performance, some correlations between reflection and philosophical beliefs have depended on factors such as personality. Also, experimentally manipulating reflection's impact on philosophical judgments has proven to be more challenging than observing correlations between reflection test performance and philosophical judgments. Reviewing these literatures will help explain the design of the current research.

4.1.1 Correlates Of Philosophical Beliefs

Philosophical beliefs have been found to correlate not only with individual differences in reflection, but also with culture, education, and personality.

Culture. As the scope of psychological research expanded beyond Western, educated, industrialized, rich, and democratic countries (a.k.a., WEIRD countries), researchers realized that many of their findings may not generalize to other countries (e.g., Henrich, Heine, & Norenzayan, 2010). Philosophers have also found cross-cultural variation in people's philosophical judgments (e.g., Machery, Mallon, Nichols, & Stich, 2004; van Dongen, Colombo, Romero, & Sprenger, 2019). Moreover, the correlation between reflection and philosophical

beliefs seemed to vary between WEIRD and non-WEIRD societies (e.g., Gervais et al. 2018; Yilmaz & Alper, 2019). Indeed, the moral side effect effect that has been found in WEIRD societies seems to have *reversed* in two traditional, non-WEIRD cultures (Robbins, Shepard, & Rochat, 2017). Cross-cultural differences in responses to moral dilemmas have also been found (e.g., Winskel & Bhatt, 2019). This evidence suggests that correlations among philosophers' beliefs and reflection might vary between citizens of WEIRD and non-WEIRD nations.

Education. Some philosophical judgments have been found to vary by class. For instance, differences in class have correlated with differences in moral judgments about thought experiments (e.g., Haidt, Koller, & Dias, 1993) and epistemic judgments about thought experiments (e.g., Weinberg, Nichols, & Stich, 2001). However, not all researchers have been able to replicate these findings (e.g., Kim & Yuan, 2015; Seyedsayamdost, 2015). One possible explanation of these mixed findings about class may have to do with education, which is often treated as a socioeconomic indicator (e.g., Lee, Kawachi, Berkman, & Grodstein, 2003). The idea is that educational differences might explain some otherwise class-based differences in philosophical judgment. For example, reflection often correlates with certain philosophical views (e.g., Yilmaz, Adil, & Iyer, 2019) and advanced education in philosophy correlated with greater reflection (Livengood, Sytsma, Feltz, Scheines, & Machery, 2010), suggesting that the classes of people who pursue advanced education in philosophy might prefer certain philosophical views.

Personality. Early studies found that intuitions about philosophical thought experiments correlated with personality (Feltz & Cokely, 2008), leading some to argue for greater restriction on appealing to intuition in philosophy (Feltz & Cokely, 2012). Later, higher powered studies confirmed that even philosophers' judgments about thought experiments correlated with personality (Holtzman, 2013). Eventually, a meta-analysis of 25 studies found robust support for

this personality-philosophy relationship among non-philosophers (Feltz & Cokely, 2019). However, other researchers find the personality-philosophy link is not cross-culturally stable among non-philosophers (Alper & Yilmaz, 2019). Although philosophical correlates of personality may vary by culture, this literature also suggests that extraversion might correlate with compatibilism about free will (Feltz & Cokely, 2008; 2019) and openness might correlate with religious and political beliefs (Gerber, Huber, Doherty, & Dowling, 2011; Saroglou, 2002).

4.1.2 Mediators of Reflection's Correlates

Past work has also found that correlates of reflection sometimes depend on gender and personality. So although variables like reflection have correlated with philosophical judgements, this relationship may be at least partly explained by other factors.

Reflection, morality, and gender. Reflective responses on the Cognitive Reflection Test (Frederick, 2005) sometimes correlate with the so-called utilitarian response to moral dilemmas (e.g., Paxton, Ungar, Greene, 2012; Baron, Scott, Fincher, & Metz, 2015; Byrd & Conway, 2019). However, some have found that this reflection-morality correlation is detected in men, but not women (e.g., Capraro & Peltola, 2018).

Reflection, personality, and gender. Reflection has also been found to correlate with openness (r = -0.08), conscientiousness (r = -0.06), extraversion (r = -0.14), agreeableness (r = -0.08), and neuroticism (r = -0.06) (Yilmaz & Saribay, 2016; Rand, 2019). However, reflection's correlations with agreeableness and conscientiousness were not significant when controlling for gender (Rand, 2019).

4.1.3 The Current Research

This paper aims to study the relationships between philosophers' reflection and philosophical beliefs as well as its mediators. One approach to these relationships is

experimental. For instance, one could randomly assign participants to reflection-inducing, reflection-inhibiting, as well as control conditions and then compare philosophical beliefs between the conditions (e.g., Shenhav, Rand, & Greene, 2012). However, researchers frequently find that inducing reflective and unreflective reasoning is more difficult than earlier work suggested (ibid.; Deppe et al., 2015; Enke et al., 2020; Meyer et al., 2015; Thompson et al., 2013). Also, while momentary changes in reflection may have an impact on one-off intuitions about novel thought experiments, such changes may have little or no impact on more considered philosophical beliefs—e.g., the philosophical beliefs of academic philosophers. Nonetheless, if changes in the disposition to reflect can have a longer-term impact on philosophers' philosophical beliefs, then we will find that philosophers' disposition to reflect will correlate with some of their philosophical beliefs. Likewise, if there is no such long-term impact of reflection, then it will be difficult to reliably detect correlations between philosophers' reflection test performance and their philosophical beliefs. For these reasons, observational studies will be more suitable than experiments in preliminary investigations of the relationships between philosophers' reflection their philosophical beliefs.

Given the existing evidence and theory, we can form a few general hypotheses about this observational research.

> *The reflective convergence hypothesis*: like non-philosophers, more reflective philosophers' will tend toward certain philosophical beliefs—and not just certain intuitions about thought experiments (e.g., Rawls, 1971; Sidgwick, 1874/1962).
>
> *The control hypothesis*: some of the individual differences in philosophers' philosophical beliefs will be better explained by factors such as culture, education in philosophy, and personality than by reflection test performance (e.g., Gervais et al., 2018).

The mediation hypothesis: even when philosophers' reflection test performance reliably correlates with their preferences for certain philosophical beliefs when controlling for other factors, some of this relationship will be mediated by factors such as training in philosophy or correlates of reflection (e.g., Baron, Scott, Fincher, & Metz, 2015).

In addition to these research-based priors, the current research can test some empirical assumptions about philosophers.

The ceiling effect assumption: academic philosophers are so reflective that there will not be enough variance in their reflection test performance to correlate with variance in their philosophical beliefs (Huemer, 2014).

The intuition-belief alignment assumption: Reflection will correlate with both intuitions about particular thought experiments such as the trolley problem as well as the corresponding philosophical views that are thought to motivate these thought experiments such as deontological and utilitarian metaethics (cf. Kahane, Everett, Earp, Farias, Savulescu, 2015).

Two large studies—one preregistered replication and extension—found support for these hypotheses, but not for these assumptions. Pre-registration, data, and R syntax are available from the Open Science Framework: osf.io/a98ck/. APA and IRB ethical guidelines were followed for all studies and analyses.

4.2 Studies

4.2.1 Method

Participants. Past meta-analyses suggest that the expected philosophy-reflection effect size would be $r \cong .18$ (Pennycook et al., 2016) or else $r \cong .2$ (Gignac & Szodorai, 2016). To

obtain 99% power to detect the more conservative r = .18, GPower suggested a sample size of 558 participants (Faul, Erdfelder, Lang, & Buchner, 2007).

In Study 1, LeiterReports.typepad.com recruited 979 respondents. An *a priori* decision was made to exclude all incomplete surveys (n = 382) and patently insincere surveys (n = 3)—i.e., one participant reported that their ethnicity was "Handsome" and other participants reported that they were citizens of the countries "Texas" or "Narnia"—resulting in a final sample of 594 (106 identified as women, 483 as men, and 5 as other; 485 identified as White, 37 as Multiethnic, 23 as Asian, 22 as Hispanic or Latino, 4 as Black, 3 as Caribbean, 1 as American Indian or Native American, and 19 as other ethnicity).

Study 2 attempted to improve upon the representativeness of Study 1 by recruiting from more and more varied participant pools and to obtain at least 80% power to detect the average philosophy-reflection correlation found in Study 1 (r = .1) by securing GPower's suggested sample size of 782 participants (Faul, Erdfelder, Lang, & Buchner, 2007). LeiterReports.typepad.com, DailyNous.com, and the PHILOSOP listserv recruited 745 respondents while Amazon Mechanical Turk simultaneously recruited 225 respondents. Again, an a priori decision was made to exclude all incomplete surveys (n = 263) and insincere surveys (n = 2)—i.e., two participants reported that they randomly selected answers—leaving a final sample of 705 (mean age = 36.5; 163 identified as women, 534 as men, and 8 as other; 548 identified as White, 65 as Asian, 43 as Multiethnic, 16 as Black or African American, 2 as American Indian or Native American, and 31 as other ethnicity).

Procedure and materials. Data for Study 1 were collected in 2014 using Qualtrics (qualtrics.com) in accordance with University of Colorado IRB protocol #13-0678 (Byrd, 2014). Data for Study 2 were collected in 2019 using Qualtrics in accordance with Florida State

University IRB Protocol #2018.25325. Participants in Study 1 completed measures of philosophical beliefs, reflection, class, and culture.

Philosophical beliefs . Following Chalmers and Bourget (2014), all participants were asked to report their agreement with philosophical views and their decision about the Trolley Problem (Foot, 1995; Thomson, 1986) by completing the 20-item PhilPapers survey—Newcomb's paradox was omitted. To limit jargon (Knutson & Presser, 2010), participants that reported not having or not being a candidate for a Ph.D. in philosophy received jargon-free translations (Table 5). Translations were developed in cooperation with philosophy graduate students and faculty at the University of Colorado (see also Yaden, *under review*).

Table 5. Example of how responses to "Do you have (or are you a candidate for) a Ph.D. in philosophy?" determined PhilPapers survey question that participants received.

Yes (Study 1 N = 328; Study 2 N = 280)	**No** (Study 1 N = 267; Study 2 N = 427)
Free will: incompatibilism or compatibilism?	If every event in the universe is determined, do you think it is possible that there could be free will?
__ Accept incompatibilism	__ Accept no
__ Lean toward incompatibilism	__ Lean toward no
__ I don't know	__ I don't know
__ No inclination	__ No inclination
__ Lean toward compatibilism	__ Lean toward yes
__ Accept compatibilism	__ Accept yes

Personality. Following prior surveys of philosophers (Holtzman, 2013), all participants were asked to complete a brief, validated, and widely-used Big-Five personality assessment (Gosling, Rentfrow, & Swann Jr., 2003) by rating their agreement with statements such as, "I see myself as disorganized, careless" on a 7-point scale ranging from *Disagree strongly* to *Agree strongly*.

Reflection. Following prior work (e.g., Shenhav, Rand, & Greene, 2012, Studies 1 and 2), all participants were asked to complete the 3-item Cognitive Reflection Test (Frederick, 2005) by answering questions like, "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" Responses were typed into a blank text box. An unreflective parameter was computed by summing participants' lured responses—ignoring all non-lured, incorrect responses (Chapter 3). A reflective parameter was computed by summing participants' correct responses.

Culture. All participants in Study 1 were asked to report their country of citizenship. Table 6 shows how these responses were sorted as WEIRD or non-WEIRD countries based on earlier work (e.g., Klein et al., 2018, Yilmaz & Alper, 2019). Study 1 was unable to recruit the quantity of participants from non-WEIRD nations that would be required to detect a small effect of culture. Without a reason to expect Study 2 to overcome this limitation, the citizenship question was not added to Study 2.

Education. Participants reported whether they had (or were a candidate for) a Ph.D. in philosophy. Response options were bivalent: yes or no.

Participants in Study 2 completed most of the measures from Study 1 as well as the following items in order to improve upon the measures in and clarify the results of Study 1.

New reflection test. The original cognitive reflection test is so widely used that many participants are already familiar with its questions and answers (Stieger & Rips, 2016). Indeed, in 2014, Study 1 found that 28% participants reported familiarity with CRT questions and 13% reported familiarity with CRT answers. CRT performance is also confounded with general math test performance (e.g., Erceg, Galic, & Ružojčić, 2020; Patel, 2017).

Table 6. List of countries and their sample sizes from Study 1.

Non-WEIRD Nations		WEIRD Nations	
Country	**N**	**Country**	**N**
Brazil	10	Australia	12
China	2	Canada	47
Costa Rica	2	Chile	1
Dominican Republic	1	Denmark	1
Ecuador	2	Finland	1
Hong Kong (China)	2	France	4
India	6	Germany	9
Lithuania	1	Hungary	1
Malaysia	1	Iceland	1
Mexico	2	Israel	1
Nepal	1	Italy	5
Peru	1	New Zealand	6
Romania	2	Norway	2
Russia	1	South Korea	2
Singapore	1	Spain	1
Taiwan (China)	1	Sweden	7
Turkey	1	Switzerland	2
		The Netherlands	4
		United Kingdom	58
		Unites States	398

To overcome these confounds with familiarity and numeracy, Study 2 participants completed fourteen newer reflection test items that are less familiar (Sirota et al., 2018) and mostly non-mathematical before they completed the original CRT in Study 2. Three mathematical reflection test prompts asked questions such as, “If it takes 2 nurses 2 min to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?” (Baron, Scott, Fincher, & Metz, 2015). Nine logical reflection test items such as, “All flowers have petals. Roses have petals. If these two statements are true, can we conclude from them that roses are flowers?” (Markovits & Nantel, 1989), in which the logical validity of the syllogisms are incongruent with the believability of their conclusions, thereby luring participants into evaluating syllogisms according to believability and not logical

validity—a.k.a., belief bias. Study 2 also completed two verbal reflection items from, among others, Sirota and colleagues (2018) including, "Ann's father has a total of five daughters: Lala, Lele, Lili, Lolo, and ____. What is the name of the fifth daughter?" Reflective and unreflective parameters were computed by summing correct and lured answers, respectively.

To test for reflection while controlling for variables that are thought to be confounded with reflection (e.g., Attali & Bar-Hillel, 2020; Campitelli & Labollita, 2010; cf. Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Szaszi, Szollosi, Palfi, & Aczel, 2017), Study 2 participants also completed more general measures of logical competence, numeracy, and preferences for open-minded thinking.

Logic test. Study 2 featured seven logical syllogisms. Three included quantifiers (e.g., 'some' or 'all') as follows: "In a box, some red things are square, and some square things are large. What can we conclude? (a) Some red things are large. (b) All red things are large. (c) We can't conclude anything about red things and large things" (Johnson-Laird & Bara, 1984). The other four syllogisms lacked quantifiers but contained unfamiliar pseudowords such as "All laloobays are rich. Sandy is a laloobay. If these two statements are true, can we conclude from them that Sandy is rich?" (Baron, Scott, Fincher, & Metz, 2015). Correct responses to these questions were summed for each participant.

Numeracy. To measure mathematical competence, Study 2 participants were asked to complete the 4-item Berlin Numeracy Test with questions such as, "Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? (a) 5 out of 50 throws, (b) 25 out of 50 throws, (c) 30 out of 50 throws, or (d) None of the above" (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Correct responses to these questions were summed for each participant.

Open-minded thinking. Next participants in Study 2 completed the 10-item Actively Open-Minded Thinking scale (Baron, 2018) by rating their agreement with statements such as "People should revise their beliefs in response to new information or evidence" on a 5-point scale ranging from *Completely disagree* to *Completely agree*. After reverse-coding items accordingly, Actively Open-Minded Thinking (AOT) ratings were summed.

Age. Some evidence suggests that age correlates with reflection (Hertzog, Smith, & Ariel, 2018) and personality (Harris, Brett, Johnson, & Deary, 2016). So participants reported their age by selecting their year of birth. Age was computed by subtracting birthyears from 2019.

4.2.2 Results

Descriptive statistics. Table 7 reports, sample sizes, ranges, means and standard deviations of reflection test performance. Like past findings (e.g., Livengood et al., 2010), philosophers were more reflective than other samples.

After Study 1 was conducted, a replication and extension was pre-registered. To replicate the univariate and multivariate regression analyses of Study 1, Study 2 pre-registered the same analyses as well as conditional mediation analyses. Specifically, if multivariate analysis of Study 2 found that reflection test performance predicted philosophical beliefs above and beyond other factors, then mediators of the remaining links between reflection test performance and philosophical beliefs would be tested.

***Univariate regression analysis*.** Aligning with the reflective convergence hypothesis, philosophers' reflection test performance correlated with beliefs about language, God, and the Trolley problem for all reflection tests in both studies—largely aligning with existing evidence about non-philosophers (Byrd & Conway, 2019; Hannikainen & Cova, *in prep*.; Jaquet & Cova,

Table 7. Descriptive statistics about reflection test performance in these and other studies.

Unreflective Responses On The Original 3-item CRT	***N***	**Range**	**Mean**	**Std. Deviation**
LeiterReports.com (Study 1)	594	0-3	0.42	0.72
Philosophers and mTurk Workers (Study 2)	705	0-3	0.58	0.87
Yale students (Meyer et al., 2015)	786	0-3	0.62	0.88
mTurk Workers (ibid.)	5191	0-3	1.27	1.25
New Haven residents (ibid.)	263	0-3	1.38	1.2

Reflective Responses On The Original 3-item CRT	***N***	**Range**	**Mean**	**Std. Deviation**
Philosophers (Study 1)	594	0-3	2.34	0.90
Philosophers and mTurk Workers (Study 2)	705	0-3	2.26	1.00
Yale students (Meyer et al., 2015)	786	0-3	1.98	1.08
mTurk Workers (ibid.)	5191	0-3	1.39	1.22
New Haven residents (ibid.)	263	0-3	1.01	1.25

New 14-item Reflection Test — Math, Logic, and Verbal	***N***	**Range**	**Mean**	**Std. Deviation**
Philosophers' and mTurk Workers' unreflective responses (Study 2)	705	0-14	3.85	2.95
Philosophers' and mTurk Workers' reflective responses (Study 2)	705	0-14	9.82	3.2

under review; Pennycook, Ross, Koehler, & Fugelsang, 2016; Reynolds, Byrd, & Conway, *in prep*.). Correlations between reflection test performance and beliefs about internalism (vs. externalism) about moral judgment, consequentialist (vs. deontological) metaethics, and empiricism (vs. rationalism) about knowledge were not detected in either Study 1 or Study 2. All zero-order correlations between philosophical beliefs and reflection test performance from both studies are reported in Table 8.

The most and largest correlations between reflection test performance and philosophical beliefs were detected using the new, less mathematical 14-item reflection test in Study 2.

Table 8. Correlations between philosophical beliefs and unreflective or reflective responses to the original 3-item as well as less familiar, less mathematical reflection tests in both studies.

Study 1 (2014, N = 594)		**Study 2 (2019, N = 705)**				
Unreflective (3-item CRT)	Reflective (3-item CRT)	Unreflective (3-item CRT)	Reflective (3-item CRT)	Unreflective (New CRT)	Reflective (New CRT)	PhilPapers Items, coded -2 to 2 (Bourget & Chalmers, 2014)
-.06	.06	**-.14*** **	**.17*** **	**-.29*** **	**.29*** **	**1. Mind**: Anti-physicalism (-) or physicalism (+)?
-.05	.02	**-.13*** **	**.15*** **	**-.22*** **	**.22*** **	**2. Mental Content**: Internalism (-) or externalism (+)?
.10*	-.05	**-.09***	**.12****	**-.13*** **	**.17*** **	**3. Language**: Russellianism (-) or Fregeanism (+)?
.07	-.04	.01	-.03	**.09***	**-.08***	**4. Analytic-Synthetic Distinction**: No (-) or Yes (+))?
-.01	.04	**-.11****	.07	**-.11****	**.11****	**5. Time**: A-theory (-) or B-theory (+)?
-.02	.03	-.00	.03	-.07	**.08***	**6. Laws of nature**: Humeanism (-) or non-Humeanism (+)?
.03	-.05	-.06	.06	**-.09***	.09*	**7. Justification**: Externalism (-) or internalism (+)?
-.09*	**.08***	-.04	.03	**-.11****	**.11****	**8. Free Will**: Incompatibilism (-) or compatibilism (+)?
.09*	**-.08***	**.24*** **	**-.26*** **	**.35*** **	**-.35*** **	**9. God**: Atheism (-) or theism (+)?
.06	-.04	-.05	.06	**-.08***	**.08***	**10. Meta-ethics**: Moral anti-realism (-) or moral realism (+)?
-.00	.04	-.03	.02	.01	-.02	**11. Moral Judgment**: Internalism (-) or externalism (+)?
.00	-.06	.01	-.03	-.01	.00	**12. Ethics**: Consequentialism (-) or deontology (+)?
-.03	.03	**-.12*** **	**.15*** **	**-.23*** **	**.24*** **	**13. Politics**: Libertarianism (-) or egalitarianism (+)?
-.05	.04	**-.19*** **	**.21*** **	**-.21*** **	**.23*** **	**14. Science**: Anti-realism (-) or realism (+)?
-.01	.03	-.07	**.08***	**-.09***	**.09****	**15. Abstract Objects**: Nominalism (-) or Platonism (+)?
-.05	.02	-.03	.03	-.03	.03	**16. Knowledge**: Empiricism (-) or rationalism (+)?
.03	-.03	**.09***	**-.11****	**.11****	**-.10****	**17. Metaphilosophy**: Naturalism (-) or non-naturalism (+)?
-.13**	**.14*** **	**-.12*** **	**.15*** **	**-.18*** **	**.19*** **	**18. Trolley Problem**: Straight (-) or turn (+)?
-.10*	.07	-.07	.07	**-.16*** **	**.17*** **	**19. Personal Identity**: Physical (-) or psychological (+)?

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Nonetheless, even the small correlations between this test and philosophical beliefs aligned with past work on non-philosophers. For instance, Study 2 detected a univariate correlation between reflection test performance and non-naturalism (e.g., Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012).

The univariate correlations from Study 1 were smaller than the anticipated effects of r = .18. In fact, the first study's correlations were as small as $r = 0.08$. Nonetheless, posthoc power analysis suggests that Study 1 had 62% power to detect its smallest correlations of $r = 0.08$ (Faul, Erdfelder, Lang, & Buchner, 2007). Indeed, GPower suggests that a sample size of 964 participants would have been required for 80% power to detect such small effects (ibid.). Study 2 overcomes this limited power not only by recruiting a larger sample, but by detecting larger correlations than were anticipated in its power analysis. Nonetheless, twice as many correlation tests were conducted in Study 2. To correct subsequent analyses for the 76 comparisons in Study 2, only univariate correlations that crossed the Bonferroni-corrected α threshold ($p = 0.001$) will be mentioned in further analyses of Study 2 (Institute of Education Sciences, 2017).

The size of univariate correlations detected in both studies ranged from small to large relative to other individual differences research (Gignac & Szodorai, 2016). Nonetheless, raw correlation coefficients and categorical labels like 'small' and 'large' do not answer all questions about the nature and magnitude of the reflection-philosophy correlations reported in aforementioned correlation tables. So Figure 4 supplements these tables with visualizations of the significant univariate correlations detected in Study 1 and Study 2.

Multiple regression analysis. Correlational analysis of Study 1 and Study 2 found that reflection test performance correlated with various philosophical beliefs. However, those

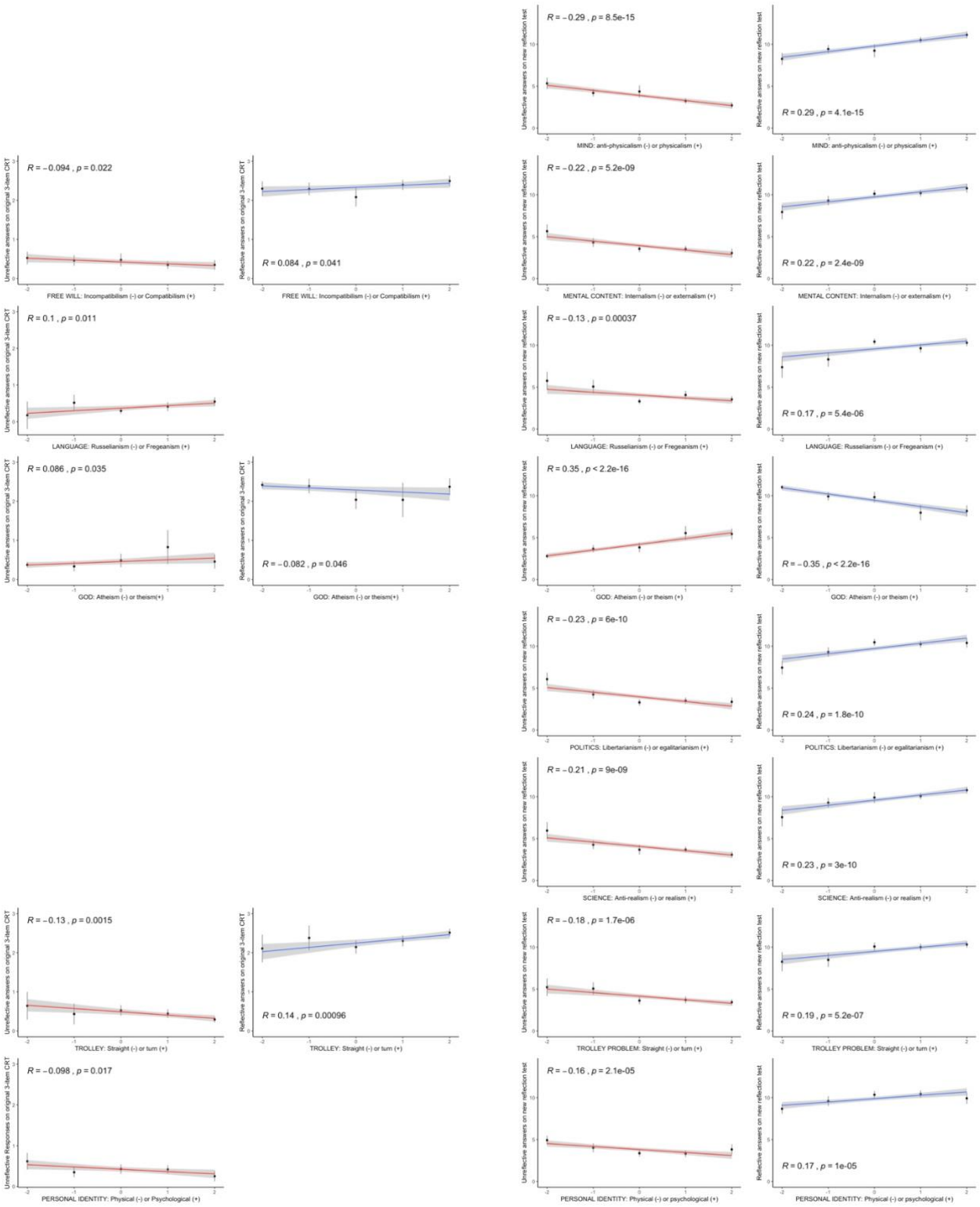

Figure 4. Significant reflection-philosophy correlations detected in Study 1 ($p < 0.05$, N = 594, left) and Study 2 ($p < 0.001$, N = 705, right) with 95% confidence intervals (grey bands).

correlations did not account for other variables that sometimes correlate with unreflective and reflective reasoning such as prior familiarity with the CRT (e.g., Stieger & Reips, 2016), philosophical training (e.g., Livengood, Sytsma, Feltz, Scheines, & Machery, 2010), gender (e.g., Frederick, 2005), WEIRDness (e.g., Yilmaz & Alper, 2019), and personality (Yilmaz & Saribay, 2016). So it is not clear whether philosophical judgments correlate with reflection *per se* or with other variables that correlate with reflection. To test the relationship between reflection and philosophical judgments while controlling for these other variables, multiple regression was employed. To reduce researcher degrees of freedom, only those PhilPapers items that correlated with reflection test performance in univariate analyses were subjected to multiple regression. Standardized regression coefficients for Studies 1 and 2 are reported in Tables 9 and 10**.**

Table 9. Standardized multiple regression coefficients predicting philosophical beliefs (-2 to 2 a la Bourget & Chalmers, 2014) from all measures in Study 1 (N = 594). Each column is a separate multiple regression analysis.

	Language Russelianism or Fregenaism?	**Free Will** Incompatibilism or compatibilism?	**God** Atheism or theism?	**Trolley Problem** Straight or turn?	**Personal Identity** Physical or psychological?
Unreflective (CRT)	**.14****	-.06	.05	-.04	-.12†
Reflective (CRT)	**.15****	.01	-.03	.06	-.06
Familiar with CRT	**-.11****	**.11****	.02	.03	-.01
Philosophy Ph.D.	**-.55*****	**.17*****	-.04	**.11****	**.28*****
Gender (M - W)	.03	.02	.00	**-.16*****	.06
WEIRD nation	-.06†	**.11****	-.04	.07†	.02
Extraversion	-.04	-.03	.01	**.09***	-.01
Agreeableness	.05	.05	**.11***	-.03	.06
Conscientiousness	.00	-.00	.02	**-.09***	-.02
Stability	.03	.05	-.03	.05	-.00
Openness	.03	**.11****	**-.10***	**.09***	-.01
Combined Adjusted R_2	**.33*****	**.07*****	**.013†**	**.08*****	**.09*****

Note: † p < .1, * p < .05, ** p < .01, *** p < .001

Reflection test performance. Further aligning with the reflective convergence hypothesis, reflection test performance in both studies correlated with philosophers' beliefs about language,

Table 10. Standardized multiple regression coefficients predicting philosophical beliefs all measures in Study 2 (N = 705). New measures in Study 2 are below the dashed line. Each column represents a separate multiple regression analysis.

	Mind Anti-physicalism or physicalism?	**Mental Content** Internalism or externalism?	**Language** Russellianism or Fregeanism?	**God** Atheism or theism?	**Politics** Libertarianism or egalitarianism?	**Science** Anti-realism or realism?	**Trolley Problem:** Straight or turn?	**Personal Identity** Physical or psychological?
Unreflective responses	-.11	-.08	**.41***	**.43***	-.16	.28	.02	.01
Reflective responses	.09	-.04	**.54****	.25	.13	**.44***	-.01	.00
Philosophy Ph.D.	.01	.06	**-.53*****	-.02	**.27*****	.05	.08†	**.28*****
Gender (M – W)	-.02	-.02	-.00	.07†	-.02	-.01	.02	.03
Extraversion	.00	-.04	-.03	.06†	-.04	-.05	**.08***	-.04
Agreeableness	-.04	**-.11****	.03	**.08***	**.13*****	.01	**-.10****	-.04
Conscientiousness	-.05	-.02	-.01	-.01	.05	-.02	-.03	-.00
Stability	-.02	-.03	.05	**.09***	-.04	.04	-.01	.07†
Openness	-.02	**.08***	.01	-.02	-.02	-.01	.01	.02
Logical Competence	-.05	.10†	**.17*****	.02	**.16*****	-.06	**.06**	**.15****
Numeracy (BNT)	.05	.08†	-.01	-.05	-.05	.07	.08†	-.07†
Open-minded Thinking	**.15****	-.01	**.21*****	**-.22*****	.07	**.10***	**.16*****	-.02
Age	.03	.05	.06†	-.02	-.06	.05	.00	-.02
Combined Adjusted R_2	**.09*****	**.08*****	**.33*****	**.17*****	**.15*****	**.06*****	**.07*****	**.10*****

Note: † p < .1, * p < .05, ** p < .01, *** p < .001

God, and science above and beyond other factors. Less reflective test performance correlated with preferences for theism over atheism while controlling for all other factors in Study 2—further aligning with existing meta-analyses (e.g., Pennycook, Ross, Koehler, & Fugelsang, 2016). More reflective test performance correlated with preferences for scientific realism over scientific anti-realism independent of other factors in Study 2. In both studies, preferences for Fregeanism about language correlated with *both* unreflective *and* reflective test performance over and above other factors—for more about this, see the mediation analyses.

Other variables. Variables other than reflection test performance explained more of the variance in other philosophical beliefs than reflection *per se*. In both studies more predictive variables included having or being a candidate for a Ph.D. in philosophy and personality—partially aligning with past work on non-philosophers (Feltz & Cokely, 2019; Hannikainen et al.,

2019). In Study 1, prior familiarity with reflection test answers, gender, and citizenship in a WEIRD country were more predictive of philosophers beliefs than reflection test performance—in line with studies of non-philosophers (Alper & Yilmaz, 2019; Friesdorf, Conway, & Gawronski, 2015; Machery, Mallon, Nichols, & Stich, 2004; Stieger & Reips, 2016; Strimaitis, 2017; van Dongen, Colombo, Romero, & Sprenger, 2019; Welsh & Begg, 2017). In Study 2, logical competence and preferences for open-minded thinking predicted more variance in philosophers' beliefs than reflection, which also aligns with prior research on non-philosophers (Byrd & Conway, 2019; Baron, Scott, Fincher, & Metz, 2015). Neither age nor numeracy predicted a significant amount of variance in philosophers' philosophical beliefs beyond the other factors.

Mediation analysis. Both studies found that reflection test performance predicted philosophical beliefs above other factors like gender and education in philosophy. This satisfies the pre-registered conditions for mediation analysis of Study 2. Another reason for mediation analyses was the finding that having a Ph.D. in philosophy often predicted variance in philosophical beliefs better than reflection. One explanation of this is an education mediation hypothesis: education in philosophy mediated the relationship between reflection and certain philosophical beliefs. After all, greater reflection correlated with having a Ph.D. in philosophy—$r = 0.38$, $p < 0.001$—having a Ph.D. in philosophy correlated with certain philosophical beliefs—such as atheism, $r = -0.16$, $p < 0.001$. So, to fulfill the pre-registered analytic strategy and clarify the role of philosophical education in the reflective convergence hypothesis, mediation analyses were conducted on data from Study 2 (*a la* Baron & Kenny, 1986). Monte Carlo power analysis for indirect effects running 1000 replications and 20000 draws per repetition suggested that Study 2 had 97% power to detect such indirect effects (Schoemann,

Boulton, & Short, 2017). To avoid the problems of CRT familiarity revealed by Study 1, reflection test performance is based on the new (rather than the original) reflection test items.

Mediation of the reflection-language relationship. Curiously, multiple regression in Study 1 and Study 2 found that both unreflective and reflective responses correlates with a preference for Fregeanism about language over Russellianism about language. However, only reflective answers correlated with a preference for Fregeanism in the univariate correlational analysis of Study 2. One explanation of the seeming reversal of the correlation between reflection and beliefs about language is that having a Ph.D. in philosophy—the best predictor of views about language from multiple regression analysis—mediated the relationship between reflection and beliefs about language. Mediation analysis confirmed this (Figure 5). Reflective responses correlated with having a Ph.D. in philosophy and having a Ph.D. in philosophy correlated with Russellianism about language. Nonetheless, after accounting for this indirect effect, the direct relationship between reflective responses and beliefs about language was a correlation with Fregeanism—in line with the initial correlational analysis—supporting both a reflective convergence hypothesis and a mediation hypothesis about reflection and beliefs about language: more reflective philosophers tended toward Fregeanism about language and some of this direct correlation was mediated by having a Ph.D. in philosophy.

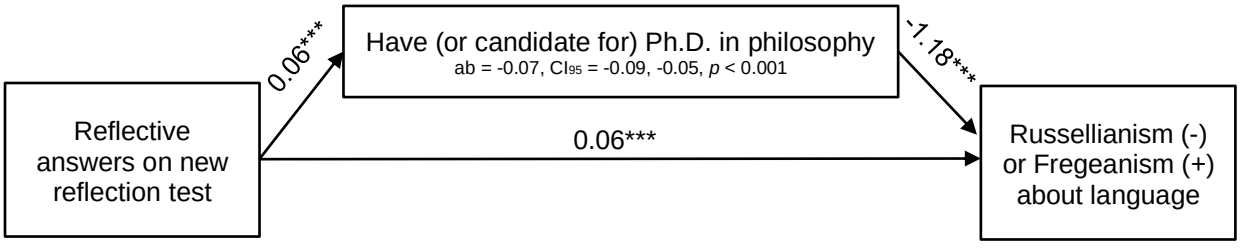

Figure 5. In 1000 bootstrapped samples, a significant average unstandardized indirect effect of Ph.D. status was detected, as was a significant average direct effect of reflective responses, b = -0.06, CI_{95} [0.03, 0.09] $p < 0.001$.

Mediation of the reflection-God relationship. The relationship between unreflective responses and belief in God was not mediated by an indirect effect through Ph.D. status (Figure 6). Unreflective responses negatively correlated with having a Ph.D. in philosophy but having a Ph.D. in philosophy did not correlate with beliefs about God when controlling for reflective answers on the new reflection test in Study 2. Nonetheless, after controlling for the potential indirect effect, there remained a significant direct relationship between reflection and atheism.

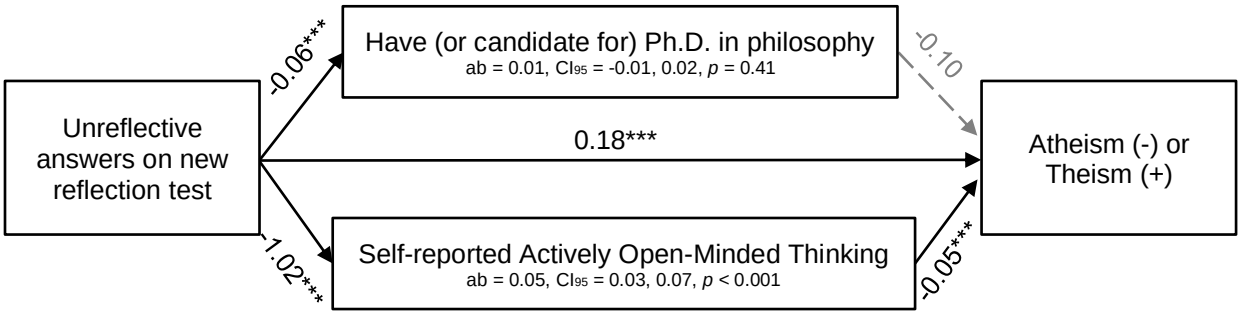

Figure 6. In 1000 bootstrapped samples, a significant average unstandardized indirect effect of Ph.D. status was not detected. However, in 1000 more separate bootstrapped samples, a significant average unstandardized indirect effect of self-reported Actively Open-Minded Thinking of was detected. In both mediation analyses, the average direct effect of reflective responses was $b = 0.18$, CI_{95} [0.15, 0.22] $p < 0.001$.

This suggests that if the relationship between reflection test performance and belief in God is mediated by a variable from the multiple regression analysis, then that variable is something other than Ph.D. status. The correlate of reflection test performance that explained the most variance in beliefs about God in multiple regression analysis was self-reported preferences for actively open-minded thinking (AOT). To test whether AOT mediated the reflection-God relationship a separate mediation analysis was conducted. The relationship between unreflective responses and belief in God was mediated by an indirect effect through AOT (Figure 6). Unreflective responses correlated with lower AOT and AOT correlated with atheism. Moreover,

after controlling for the indirect effect of AOT, the remaining direct relationship between unreflective responses and theism remained significant. This supports both a reflective convergence hypothesis and a mediation hypothesis about reflection and atheism: more reflective philosophers tended toward atheism and part of this direct correlation was mediated by self-reported preferences for actively open-minded thinking.

Mediation of the reflection-science relationship. The relationship between reflection and beliefs about science was not mediated by an indirect effect through Ph.D. status (Figure 7). Reflective responses correlated with having a Ph.D. in philosophy but having a Ph.D. in philosophy did not correlate with beliefs about science when controlling for reflective responses on the new reflection test in Study 2. However, after accounting for an indirect effect, the direct relationship between reflective responses and beliefs about science was a correlation with scientific realism.

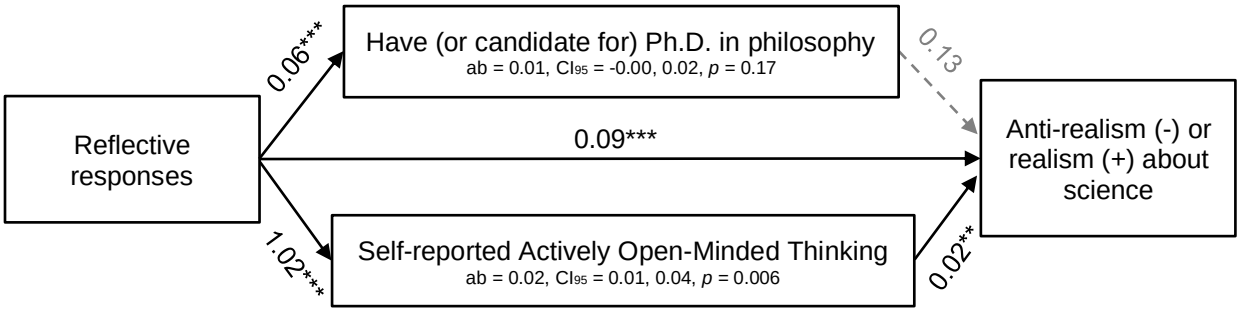

Figure 7. In 1000 bootstrapped samples, a significant average unstandardized indirect effect of Ph.D. status was not detected. However, in 1000 additional bootstrapped samples, a significant average unstandardized indirect effect of self-reported Actively Open-Minded Thinking of was detected. In both mediation analyses, the average direct effect of unreflective responses was b = 0.09, CI_{95} [0.01, 0.04] $p < 0.001$.

Again, this suggests that if the relationship between reflection test performance and beliefs about science is mediated by a variable from the multiple regression analysis, then that

variable is something other than Ph.D. status. The only other significant predictor of variance in beliefs about science from the multiple regression analysis was self-reported preferences for actively open-minded thinking (AOT). To test whether AOT mediated the reflection-science relationship another mediation analysis was conducted. The relationship between reflective responses and beliefs about science was mediated by an indirect effect through AOT (Figure 7). Reflective responses correlated with higher AOT and AOT correlated with scientific realism. Moreover, after controlling for the indirect effect of AOT, the remaining direct relationship between reflective responses and scientific realism remained significant. This also supports both a reflective convergence hypothesis and a mediation hypothesis about reflection and scientific realism: more reflective philosophers tended towards scientific realism and some of this direct correlation was mediated by self-reported preferences for actively open-minded thinking.

Overall, mediation analyses found support for the mediation hypothesis: either having a Ph.D. in philosophy or reporting preferences for actively open-minded thinking partially mediated the three significant reflection-philosophy correlations that survived multiple regression analysis in Study 2. Nonetheless, once these three mediation paths were accounted for, the direct reflection-philosophy correlations remained. So mediation analyses also found more support for reflective convergence hypotheses for philosophers' beliefs about language, God, and science.

4.3 General Discussion

In two large studies of mostly philosophers (total N = 1299), less reflective participants tended toward different philosophical beliefs than more reflective participants. Some of these correlations between reflection and philosophical beliefs remained when controlling and testing for mediation by other relevant factors. Also, correlations were detected with intuitions about

particular thought experiments, but not with the general philosophical views that may motivate those thought experiments. So these data amount to support for the reflective convergence, the control, and the mediation hypotheses, but not the ceiling effect and intuition-belief alignment assumptions.

4.3.1 Reflection & Metaphilosophical Orientation

The philosophical beliefs that most reliably correlated with philosophers' reflection test performance across univariate, multivariate, and mediation analyses of both studies were beliefs about language, God, and science. In particular, reflective responses correlated with Fregeanism about language and scientific realism whereas unreflective responses correlated with theism.

The current research did not pre-register factor analyses. However, prior research on philosophers found that preferences for theism clustered together with similarly "anti-naturalist" philosophical preferences such as non-naturalism (Bourget & Chalmers, 2014). So one hypothesis about the present findings is that less reflective philosophers have non-naturalist or supernaturalist metaphilosophical inclinations. This metaphilosophical hypothesis enjoys some support from the present univariate correlations detected in Study 2. This unreflective tendency toward metaphilosophical non-naturalism would find support from prior work on non-philosophers finding that unreflective responses correlated with belief in God (e.g., Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014a; Pennycook, Cheyne, Koehler, & Fugelsang, 2013; Pennycook et al., 2016; Zuckerman, Lee, Lin, Hall, 2019; cf. Yilmaz & Isler, 2019), paranormal beliefs (e.g., Aarnio & Lindeman, 2005; Erceg, Galić, & Bubić, 2019; Gianotti, Mohr, Pizzagalli, Lehmann, & Brugger, 2001; Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012), and ontological confabulation (Mækelæ, Moritz, & Pfuhl, 2018). Of course, this metaphilosophical hypothesis is just that: a hypothesis. Although compatible with some of the present evidence, it is

not necessarily the best or only explanation of the present evidence. Ultimately, the soundness of this metaphilosophical hypotheses depends on the results of further research.

Philosophical beliefs that most reliably correlated with reflective responses were Fregeanism about language and scientific realism. This fits less neatly with the aforementioned factor analyses, which found an "anti-realist" component including both Fregeanism about proper names and realism (rather than anti-realism) about science (Bourget & Chalmers, 2014). So it less clear how earlier research on philosophers' beliefs yields a coherent hypothesis about why more reflective philosophers preferred both Fregeanism about language and scientific realism.

4.3.2 Reflection & Cultural Defaults

The metaphilosophical hypothesis is limited by—among other things—the fact that some of the beliefs that correlated with unreflective test performance are not obviously non-naturalist —e.g., scientific anti-realism. So even if the metaphilosophical orientation hypothesis is supported by future evidence, additional hypotheses might fruitfully supplement them. One such supplemental hypothesis is a cross-cultural difference hypothesis (e.g., Henrich, Heine, & Norenzayan, 2010; Kitayama et al., 2019).

Prior work finds that the relationship between reflection and atheism varies across cultures (e.g., Byrd & Sytsma, *in prep*.; Gervais et al., 2018). These and other cross-cultural findings raise the question of whether reflection can decrease acceptance of a culture's default philosophical beliefs. The idea is that just as reflection can involve questioning one's initial impulse on a reflection test, reflection might also involve questioning the predominant philosophical beliefs of one's culture. This culture-interventionist hypothesis would predict that reflection will correlate with atheism in predominantly theist societies, but not in societies where

atheism is the predominant religious orientation. Thus, culture might be able to explain reflection-philosophy correlations that the metaphilosophical hypothesis cannot explain on its own.

Of course, it may be that both metaphilosophical and cultural tendencies are related to philosophical tendencies (Baron, 2020). There is growing support for such dual-inheritance hypotheses about philosophical beliefs among non-philosophers (e.g., Gervais, Najle, Schiavone, & Caluori, *in prep*.). So subsequent research might profitably extend this literature by testing dual inheritance hypotheses about the philosophical beliefs of academic philosophers.

4.3.3 Normativity & Appealing To Reflection

Some psychologists argue for the normative superiority of certain philosophical beliefs by appeal to their correlations with reflective reasoning (e.g., Baron, 1994, Greene, 2013). The idea is that reflection leads to better judgments. For instance, more reflective people are less likely to believe fake news (Pennycook & Rand, 2019b), less likely to accept conspiracy theories (Ćavojová, Secară, Jurkovič, & Šrol, 2018; Pennycook, Fugelsang, & Koehler, 2015a), less likely to affirm pseudo-profound bullshit (Čavojová, Brezina, & Jurkovič, 2020; Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015; Pennycook & Rand, 2019a), less influenced by emotion or disgust (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014a), less susceptible to the sunk cost fallacy (Ronayne et al., 2020), and less likely to fall prey to misinformation about their own eyewitness memory (Greene, Maloney-Derham, & Mulligan, 2020). So if more reflective people tend toward some philosophical beliefs over alternatives, then—according to the appeal to reflection—those philosophical beliefs are probably superior. While recent evidence undermines some utilitarians' appeals to reflection (Byrd & Conway, 2019), one might

wonder if the appeal to reflection could be marshalled for or against other philosophical beliefs (cf. Easton, 2018).

There is at least one empirical obstacle for such a broad appeal to reflection: it is not obvious how dispositions to overcome faulty impulses about simple mathematical and logical questions would lead to normatively superior reasoning in academic philosophy. This challenge to the appeal to reflection resembles a challenge to the appeal to expert intuition (e.g., Clarke, 2013; Weinberg, Gonnerman, Buckner, & Alexander, 2010). What the appeal to reflection currently lacks is an empirically adequate account of how someone's reflection about contrived questions about bats and balls can determine that their philosophical beliefs also benefitted from reflection (De Neys, 2020). Existing attempts to find empirical support for such appeals to reflection have been less than encouraging. For example, experimentally increasing reflection may have little impact on philosophical judgments (Colaço, Kneer, Alexander, & Machery, 2016; Deppe et al., 2015; Meyer et al., 2015; Shenhav, Rand, & Greene, 2012) aside from increases doubt about one's existing beliefs (Yilmaz & Isler, 2019).

There are also philosophical obstacles to the appeal to reflection. For instance, some philosophers have argued that some philosophical beliefs are "properly basic" and justified independently of reflective reasoning (e.g., Plantinga, 1967; cf. De Cruz, 2014). Indeed, some of the beliefs that correlated with reflection test performance in the current research—e.g., theism—are precisely the beliefs that some philosophers take to be so basically justified. These philosophers are not alone in thinking that some beliefs are less subject to the norms assumed by reflection tests: Children and adults in the US seem to agree (e.g., Heiphetz, Spelke, Harris, & Banaji, 2013; Liquin, Metz, & Lombrozo, 2018; Metz, Weisberg, & Weisberg, 2018). Thus, religious believers may not treat some beliefs as requiring the kind of reflection that some

atheists and agnostics demand. So correlations between unreflective thinking and theism may not entail that religious believers are less reflective as much as they entail that religious believers endorse a different epistemology.

Hence, appeals to reflection incur the burden of empirically ruling out alternative hypotheses about correlations between reflection test performance and philosophical beliefs. The present research does not overcome this burden. Rather, it merely identifies some philosophical beliefs that could be relevant to the debate about appealing to reflection.

4.4 Conclusion

The current research largely replicated and clarified the relationships between reflection and philosophical beliefs. The reflective convergence hypothesis was largely supported: philosophers' reflection test performance correlated with many of the philosophical beliefs that have correlated with non-philosophers' reflection test performance—beliefs about God, science, and politics—as well as other philosophical beliefs—beliefs about language, mind, mental content, and personal identity.

However, after controlling for other factors such as age, culture, gender, numeracy, preferences for open-minded thinking, personality, and philosophical training, correlations between reflection and beliefs about mental content, time, politics, and science, and judgments about the trolley problem were no longer detected. This supports the control hypothesis.

Moreover, the mediation hypothesis was supported: when reflection did correlate with philosophical beliefs over and above other factors, education or preferences for actively open-minded thinking partially mediated those remaining reflection-philosophy correlations.

Contrary to the intuition-belief alignment assumption, while reflection correlated with judgments about thought experiments such as the trolley problem, it did not correlate with the more general philosophical views thought to motivate those thought experiments such as deontological or consequentialist metaethics. This affirms existing distinctions between considered endorsement of general philosophical views and one-off intuitions about particular thought experiments (Cullen, 2010; Wilkenfeld, 2020). After all, accepting utilitarian tradeoffs in particular sacrificial moral dilemmas may not entail general endorsement of utilitarian metaethics (Kahane, Everett, Earp, Farias, & Savalescu, 2015; Conway, Goldstein-Greenwood, Polacek, & Greene, 2018).

Finally, contrary to the ceiling effect assumption, there are some robust relationships between reflection and philosophers' philosophical beliefs even after controlling for other factors and looking for mediators. Indeed, some individual differences between philosophers' beliefs were reliably explained by the fact that philosophers did not think alike about—among other things—reflection tests.

CHAPTER 5
WHAT WE CAN (AND CAN'T) INFER ABOUT IMPLICIT BIAS

This chapter was published in Synthese (Byrd, 2019) prior to defense.

> The imagination is influenced by associations of ideas; which, …are not easily altered.
>
> David Hume (1983)

Imagine nutrition scientists discover that bodyweight can be changed not only by calorie ingestion and consumption, but by other factors. When science columnists catch wind of these findings, they write up pieces with titles like "Why Calories Don't Matter", arguing that gaining and losing weight is not predicated on "any" caloric processes. Some columnists go as far as to recommend that the received, thermodynamic view of bodyweight be abandoned. Obviously, the science columnists' conclusions do not follow. The scientists did not demonstrate that changes in bodyweight are not predicated on any caloric processes. Rather, the scientists demonstrated that some weight changes are not predicated on "only" caloric processes. That finding is consistent with the idea that bodyweight is predicated on caloric processes, even if not fully. This paper cautions against the science columnists' any-only mix-up when thinking about implicit bias: the mistake of concluding that implicit bias is not predicated on any instances of a particular process when the evidence merely shows that implicit bias is not predicated on only instances of that particular process.

Discussions of implicit bias are increasingly common. Debate moderators ask presidential candidates about implicit bias (Blake, 2016), Fortune 500 companies close thousands of stores in order to teach their employees about implicit bias (Meyer, 2018), and philosophers worry that implicit bias poses epistemic threats to philosophy (e.g., Saul, 2013a, 2013b; Peters, 2019). Nonetheless, some are skeptical about the existence of implicit bias or the efficacy of corporate implicit bias training (e.g., McCoy, 2018). So, academics try to remind the public about evidence of implicit bias (e.g., Payne, Niemi, & Doris, 2015; Jost et al., 2009) and successful debiasing (e.g., Carley, 2018). Philosophers of mind have taken this evidence seriously, arguing that these debiasing findings undermine the received view of implicit bias (e.g., Mandelbaum, 2016) and demand new solutions to implicit bias (e.g., Huebner, 2016; Madva, 2017).

Given these stakes in philosophy and in public discourse, one will want to take every opportunity to be careful about what they infer about implicit bias from debiasing experiments. This paper explains how to identify methodologically sound debiasing experiments and determine what they tell us about implicit bias. Section 5.1 explains and distinguishes nine views of implicit bias. Section 5.2 explains how to (and how not to) draw inferences from debiasing experiments. Then, Section 5.3 reviews influential debiasing experiments, highlighting differences in methodological quality along the way. Section 5.4 explains what follows from the strongest evidence, using the inference principles from earlier sections. Of course, a paper this size cannot carefully examine every debiasing experiment. So, Section 5.4 also explains what would follow if forthcoming or overlooked debiasing experiments' findings differ from the findings considered herein. The primary conclusion is that up to three views of implicit bias are compatible with current and future evidence: associationism, interactionism, or minimalism. A

secondary conclusion is a sort of reflectivism about implicit bias. These conclusions imply that both the received view and more recent non-associationist views of implicit bias are incompatible with strong evidence. Reviewing some of the literature on implicit bias will help explain how these conclusions follow.

5.1 Implicitly Biased Behavior

The most well-known measure of implicitly biased behavior is the Implicit Association Test (IAT for short). The IAT is a categorization task. Various versions of the test measure various modes of implicit biases in behavior. For example, the Race IAT measures differences in responses to racial stimuli. This paper will focus on the Race IAT, but its analysis can be fruitfully applied to other versions of the IAT and other indirect measures of bias.

The IAT includes multiple phases of categorization. In the first phase of the Race IAT, participants press buttons on a keyboard to categorize words into one of two categories: GOOD or BAD. Then participants categorize faces with either white or black racial features into one of two categories: WHITE or BLACK (Figure 8). This much is fairly straightforward.

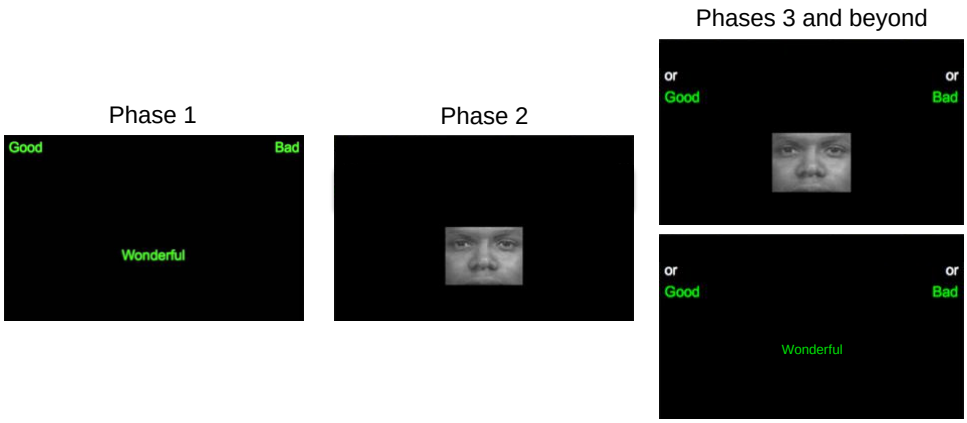

Figure 8. Phases of the Implicit Association Test: word categorization, face categorization, and word-and-face categorization.

In each subsequent phase, participants categorize either faces or words, one at a time, into composite categories: In one phase, the composite categories might be BLACK/GOOD or WHITE/BAD and in the following phase, composite categories might be WHITE/GOOD or BLACK/BAD. It is in these latter phases with composite categories where interesting patterns emerge. Most participants' categorization accuracy and response latencies reveal a preference for white facial features over black facial features (e.g., Greenwald, McGhee, & Schwartz, 1998). That is, participants are quicker to pair black facial features than white racial features with composite categories containing BAD. And, likewise, participants are quicker to pair white facial features with composite categories containing GOOD.

It is not uncommon to detect such implicit Pro-White biases in the behavior of those who explicitly express Pro-Black preferences (e.g., Gaertner & McLaughlin, 1983). While this does not suggest that people are unaware of their own biases (Gawronski, Hofmann, & Wilbur, 2006; Gawronski, forthcoming), it does suggest that behavior can be biased in ways that are not consciously endorsed or even in ways that are consciously disavowed.

Naturally, this disconnect between implicit biases in behavior and more explicit attitudes might raise questions about whether there is a disconnect between implicit biases and behaviors besides button-pressing (Greenwald, Andrew, Uhlmann, & Banaji, 2009; Greenwald, Banaji, & Nosek, 2015; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2015). In short, one might wonder about the validity of measures like the IAT. The virtue of the IAT is its ability to accurately quantify error rates and reaction times and other indirect measures of attitudes and behavior in controlled settings (Jost, 2018). More ethologically valid measures of implicit biases in behavior make quantification, timing, and control more challenging—e.g., implicit biases in resume evaluation (e.g., Tyler & Mccullough, 2009) and seating distance (Sechrist & Stangor, 2001).

Fortunately, the present paper's analysis will apply to debiasing according to any indirect measure of biases in behavior. So, concerns about the validity of the IAT undermine the present investigation only if these concerns generalize to all indirect measures.

The name 'Implicit Association Test' advertises how implicitly biased behavior was initially thought to be predicated on associations (Greenwald et al., 1998). Consequently, this associative view of implicit bias became the received view of implicit bias among philosophers (e.g., Gendler, 2008a, 642; 2008b, 577). Philosophers describe associations as "pairs of thoughts [that] become associated based on […] past experience" (Mandelbaum, 2017). Accordingly, the associative explanation of the Race IAT findings is roughly as follows: people experience White racial features paired with positive valences more than negative valences and they experience Black racial features paired with negative valences more often than positive valences. This conditioning results in associations between White racial features and positive valences or Black racial features and negative valences. These associations explain why unendorsed preferences for certain racial stimuli would manifest on tasks like the IAT.

However, this associative view of implicit bias has become controversial. Some argue that implicit bias is belief-like (Mandelbaum, 2013; cf. Madva, 2015) and that implicit bias is "not predicated on any associative structures or processes" (Mandelbaum, 2016, p. 629). Others argue that while implicit bias might be belief-like, such beliefs are nonetheless dispositional (Schwitzgebel, 2002, 2010; cf. Quilty-Dunn & Mandelbaum, 2017). Yet others argue that implicit bias is less like belief and more like a patchy endorsement (Levy, 2015) or a trait (Machery, 2016). And, coming full circle, some admit that implicit bias might be associative after all, even if only in part (e.g., De Houwer, 2006; Del Pinal & Spaulding, 2018, Huebner, 2016; Gawronski & Bodenhausen, 2014). Some background theory and evidence will explain

why anyone would want to abandon the received, associative view of implicit bias for other views.

5.1.1 Dual Process Theory

Consider the dual-process theory of cognition. The theory distinguishes between at least two types of processes with labels such as ‘Type 1’ and ‘Type 2’ (e.g., Evans & Stanovich, 2013; Table 1) or ‘System 1’ and ‘System 2’ (e.g., Evans, 2009, Table 2.1; Frankish, 2010, Table 1). To make it easier to remember what these labels describe, this paper will borrow more informative labels for each type of processing: Type 1 processes will be labeled ‘unreflective’ and Type 2 processes ‘reflective’ (à la Pennycook, Cheyne, Koehler, & Fugelsang, 2015; Strack & Deutsch, 2004). Some common dual-process distinctions are found in Table 11.

Table 11. Dual process descriptions

Unreflective (Type 1)	**Reflective (Type 2)**
associative	non-associative
fast	slow
automatically processed	deliberately processed
not consciously represented	consciously represented

Of course, one need not buy all the common dual process distinctions—at least, not without qualification. Indeed, one might be suspicious of binary distinctions in psychology more generally (Newell, 1973). Fortunately, one need not accept all common or binary dual-process distinctions in order to accept the conclusions of this paper. Consider two examples of common dual-process theory distinctions that need not be accepted without qualification.

Start with the associative vs. non-associative distinction. Explaining behavior in terms of associations is about as old as philosophy (Anderson & Bower, 1980, p. 9), so many construals of associations have accumulated. Hume thought that associations operate automatically and unconsciously (Hume, 1978).

> Tis evident, that the association of ideas operates in so silent and imperceptible a manner, that we are scarce sensible of it, and discover it more by its effects than by any immediate feeling or perception.

Some cognitive scientists have adopted such Humean construals of associations. For example;

> When a response is produced solely by the associative system, a person is conscious only of the result of the computation, not the process. Consider an anagram such as ‘involnutray' for which the correct answer likely pops to mind associatively (involuntary) (Sloman, 1996, 6).

However, the Humean construal of associations is controversial. Indeed, there are plenty of reasons to think that associations can cross the conscious/non-conscious divide (Dacey, 2016; Devine, 1989; Fridland, 2017, 2019; Hahn, Judd, Hirsch, & Blair, 2014). Because of this, some have cautioned against inferring either that cognitive processing is necessarily associative because it is automatic or unconscious or that it is necessarily automatic and unconscious because it is associative (Mandelbaum, 2016, p. 647; cf. Hütter & Sweldens, 2018). Importantly, this implies that the associative vs. non-associative distinction could be orthogonal to the unreflective vs. reflective distinction (*contra*, for example, Strack & Deutsch, 2004). This chapter takes that possibility seriously, as I explain below.

Consider the distinction between fast and slow processing (e.g., Kahneman, 2011), which is also controversial. Seemingly reflective reasoning is sometimes fast (Bago & De Neys, 2017; De Neys & Pennycook, 2019). For this and other reasons, many cognitive scientists seem to reject a definite distinction between fast and slow processing (Krajbich, Bartling, Hare, & Fehr, 2015; Pennycook, Fugelsang, Koehler, & Thompson, 2016; Sun, 2016). However, one can admit that the boundary between fast and slow is vague while maintaining that there is a range of response times within which mental representations are unlikely to be available for conscious control or even explicit endorsement (Posner, Snyder, & Solso, 1975).

At this point, a critic of dual-process theory might begin to question the existence or utility of a dual-process distinction (Melnikoff & Bargh, 2018). However, the critic should remember that the absence of a clear categorical dual-process distinction does not show that dual-process distinctions are altogether illegitimate (Pennycook, Neys, Evans, Stanovich, & Thompson, 2018). A categorical distinction proposes a clear boundary between two concepts, whereas a comparative distinction merely proposes a relative difference between two concepts (Carnap 1950, Sections 3 to 8). So, dual-process theorists have explicated some dual-process distinctions comparatively rather than categorically (e.g., Evans & Stanovich, 2013, pp. 229-231). That brings us to the two dual-process distinctions employed in this paper.

First, this paper will employ the common distinction between reflective and unreflective processing. However, this distinction will be comparative rather than categorical. Reflective processing is more consciously represented and deliberately processed while unreflective processing is less consciously represented and more automatically processed (Shea & Frith, 2016). Cognition is more conscious when participants are more aware of, more able to articulate, and/or more able to process it at the personal level (ibid.). Cognition is more deliberate when it

involves more interruption of or less acceptance of the output of automatic processing (Bargh, 1992; Fridland, 2016; Moors & De Houwer, 2006). This explication of reflection will be familiar to anyone who is aware of the famous cases of reflection from philosophy and psychology: someone finds their first intuition plausible, but steps back for a moment to consider their intuition, and then either endorses the intuition or arrives at a new response (e.g., Frederick, 2005; Korsgaard, 1996).

Second, I will employ a categorical distinction between associative and non-associative processing. Before I describe this categorical distinction, two caveats are in order. First, while processing is either associative or non-associative, attitudes and behavior may not be so binary. Indeed, one of the morals of this present paper will be that one and the same behavior can be influenced by both associative and non-associative processes. Second, there is an emerging literature which disputes what associative processing can and cannot do (e.g., Buckner, 2018; cf. De Houwer, 2018). Since that debate has yet to resolve, I will grant a conventional notion of associative processing and point interested readers toward the unfolding debate (Corneille & Stahl, 2018). Conventionally, cognitive processing is associative if it can be well-described by stimulus-response phenomena such as conditioning or counterconditioning (à la Mandelbaum, 2016). Conditioning and counterconditioning involve repeatedly activating two representations until activating one representation also activates the other representation. This explication of associations captures the kind of processing that might be involved in the behavior that is measured by the Race IAT. For example, a racial association might be formed as follows. For whatever reason, someone repeatedly experiences BLACK MALE paired with DANGER. These experiences create and strengthen an association between the concept representation (BLACK MALE) and the negatively valenced representation (DANGER). Once the association is formed,

the mere activation of BLACK MALE activates the negative valence DANGER. That automatic activation of negative valence is supposed to explain the often-unendorsed reflexive biases that manifest during the Race IAT.

A 2x2 matrix can be constructed to sort cognition according to the two distinctions just explained (Figure 9). The boundary between the left and right sides of the matrix separates associative from non-associative processing. The fuzzy boundary between the top and bottom separates more reflective from less reflective processing.

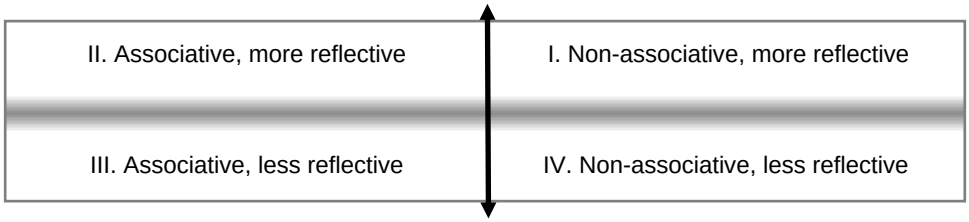

Figure 9. Matrix distinguishing four modes of cognition.

One might think that this deviates from dual-process theory since it proposes four processes. In reality, this merely proposes that two common dual-process distinctions are orthogonal. Besides, this more-than-two quadrant approach to dual-process theory is already common among cognitive scientists (e.g., Evans, 2009; Stanovich, 2009; Gawronski & Bodenhausen, 2014; Shea & Frith, 2016). Further consideration of these approaches and dual-process theory goes beyond the scope of the present investigation. The existing conceptual space need only represent key differences between the views of implicit bias under consideration in the present paper.

5.1.2 Nine Views of Implicit Bias

The matrix just described allows us to classify views of implicit bias based on whether they predicate implicit bias on (1) associative or non-associative processing as well as (2) more or less reflective processing. This produces a 3x3 matrix of nine categories of views about implicit bias (Table 12).

Table 12. A matrix of up to nine modes of cognitive processing on which implicit bias could be predicated.

Implicit bias is predicated on…	**Associative processing**	**Associative or non-associative processing**	**Non-associative processing**
More reflective processing	Reflective associationism about implicit bias	Reflective interactionism about implicit bias	Reflective non-associationism about implicit bias
More or less reflective processing	Associationism about implicit bias	Interactionism about implicit bias	Non-associationism about implicit bias
Less reflective processing	The received view of implicit bias	Unreflective interactionism about implicit bias	Unreflective, non-associationism about implicit bias

In the left column are associationist views of implicit bias. In the bottom, left cell is the received view of implicit bias, claiming that implicit bias is predicated on associations that are processed less reflectively. The cell above the received view of implicit bias refers to a more capacious associationism about implicit bias, according to which implicit bias is predicated on associations, but allowing that associations can be processed more reflectively or less reflectively.

In the middle column are interactionist views about implicit bias. Interactionism about implicit bias claims that implicit bias can be predicated on associative and non-associative processes. Interactionists about implicit bias can, in principle, claim that implicit bias is predicated on more reflective processing (top center), less reflective processing (bottom center), or some combination thereof (middle, center).

In the far-right column are non-associationist views about implicit bias. Non-associationism about implicit bias can be described by statements like, "implicit biases are not predicated on any associative structures or associative processes" or "the structure of implicit bias is not, after all, underwritten by associations" (Mandelbaum 2016, pp. 629, 637). Similar to associationism and interactionism about implicit bias, non-associationism about implicit bias can vary depending on whether it predicates implicitly biased behavior on more reflective processing, less reflective processing, or some combination thereof.

Before we determine how to evaluate these views of implicit bias, a few words of clarification are in order. First, determining who has defended each kind of view is a worthy historical project, but that is beyond the scope of the present investigation. Second, this matrix might not classify all possible views of implicit bias. Third, the matrix spares some of the details of the views that it classifies. For example, the matrix's interactionist views specify that associative and non-associative processes interact, but do not specify how these processes interact—the latter is discussed in the next section and by Gawronski & Bodenhausen (2014). Nonetheless, the matrix visualizes various answers to an ongoing question in the debate about the nature of implicit bias: Should the received view of implicit bias be abandoned for a more centrist, more far-right, or more reflective view of implicit bias?

5.2 Inferences From Debiasing Experiments

Determining the kind(s) of processing on which implicit bias is (and is not) predicated involves determining the kinds of processing on which debiasing is (and is not) predicated. In other words, views of implicit bias depend—at least in part—on what can be inferred from debiasing experiments. So, we need to determine what can be inferred.

5.2.1 Existing Inferences from Manipulation

One way to proceed is to follow precedent. The debiasing literature contains at least two inferential principles for determining the types of processing on which implicit bias is (and is not) predicated. One common inference in the debiasing literature assumes the following principle of affirmation.

> *Affirmative Manipulation Principle*. S is predicated on P-type processing just in case a P-type manipulation changes S.

The affirmative manipulation principle seems to feature in hypothetical deductions that implicit bias is propositional.

> …if you find two negatives making a positive, what you've found is a propositional, and not an associative, process. […] When a person you don't like dislikes another, you tend to like that other person. When a person you don't like dislikes another, you tend to like that other person. So, a negative valence when combined with a negative valence somehow results in a positive valence. The 'somehow' […] is sensible on a propositional theory. (Mandelbaum 2016, pp. 640-641)

The point is not that two positives making a negative cannot be well-explained by associative processes. Like many claims about what counts as an associative process (e.g., De

Houwer, 2018), that point is controversial (Toribio 2018b; see also Cone, Mann, & Ferguson, 2017). For example, Gawronski, Walther, and Blank (2005) do not endorse that claim when reporting that multiple positives made a negative. Also, that claim relies on a dichotomy between associative and propositional processing that might be false given what is already known about human and animal psychology (Buckner 2018, 3-6). Hence, the point is just that if we agree that an experimental manipulation involves certain types of processing, then we should agree that any phenomena changed by that manipulation are predicated on those types of processing. I will grant the legitimacy of the affirmative inference principle in this paper.

Another common inference in the debiasing literature assumes the following principle of negation.

> *Negative Manipulation Principle*. S is not predicated on P-type processing just in case S is manipulated by a non-P-type manipulation or S is not manipulated by a P-type manipulation.

The negative manipulation principle seems to be at work in arguments for non-associationism about implicit bias. For example;

> … if AIB [associationism about implicit bias] is true, then no logical or evidential interventions should directly work to change implicit attitudes. […] If there are [such] interventions that reliably work to counteract implicit bias […], then we have evidence that the structure of implicit bias is not, after all, underwritten by associations. […. And] a logical intervention did in fact have an impact on participants' implicit attitudes. (Mandelbaum 2016, pp. 635, 637, 645)

The conclusion is that "we have evidence that the structure of implicit bias is not, after all, underwritten by associations" (ibid.) The idea is that logical or evidential manipulations are non-associative. So, according to the negative manipulation principle, if non-associative manipulations change implicit biases, then those implicit biases are not predicated on associations.

However, there are problems with the negative manipulation principle. One problem is that the negative manipulation principle leads to something like the science columnists' any-only mix-up: the mistake of thinking that implicitly biased behavior cannot be predicated on any associative process because implicitly biased behavior is not predicated on only associative processes.

5.2.2 Manipulation Is Not Enough for Negative Inference

To avoid the science columnists' any-only mix-up about implicit bias, the negative manipulation principle will need to be replaced with a more circumspect principle, like the one below.

> *Negative Intervention Principle*. S is not predicated on P-type processing just in case both P-type manipulations or measurements and non-P-type manipulations or measurements are employed and, empirically, only non-P-type manipulations cause a change in S.

Unsurprisingly, the difference between the negative manipulation principle and the negative intervention principle has to do with the difference between manipulation and intervention. Not all philosophers or social scientists distinguish manipulation from intervention—indeed, many use the words interchangeably. So, a definition of 'intervention' is in order: process P intervenes on S only when the change in S is caused by only P. Or, in causal

graph terminology, P intervenes on S only when it "breaks all other arrows directed into" S (Woodward, 2016).

Manipulations, on the other hand, change S in a way that can involve multiple causes. So, manipulations do not show that a change was caused by only one process. However, interventions show both: that something is changed and that the change was caused by only one process. So, while manipulations are necessary for intervention, they are not sufficient for intervention because interventions are a subset of manipulations. By conjunction, we are less likely to detect interventions than manipulations. This can be illustrated by imagining the development of a debiasing research program.

Exploratory experiments. At first, researchers just want to see if any manipulation whatsoever can cause debiasing. So, researchers do not design their experimental manipulations according to the theoretical likelihood that they involve a particular type of processing such as associative or non-associative processing. Rather, researchers design their manipulations based on anecdotes or intuitions about how debiasing works. Eventually, the researchers find that certain experimental conditions reduce implicit biases in behavior significantly more than control conditions.

Because the researchers do not have strong theoretical reasons to think that their manipulation involved only associative or only non-associative processing, there are at least six viable interpretations of their debiasing results (Figure 10, adapted from Figure 1 in Perugini, Richetin, & Zogmaister, 2010; see also Madva, 2015, Section 6, and Brownstein, 2018 for more discussion of this interpretive difficulty). Only two of these six interpretations involve a single type of processing intervening on implicitly biased behavior (a and b). The other four interpretations involve two kinds of processing jointly manipulating implicitly biased behavior

(c, d, e, and f). So, exploratory experiments cannot arbitrate between the aforementioned views of implicit bias.

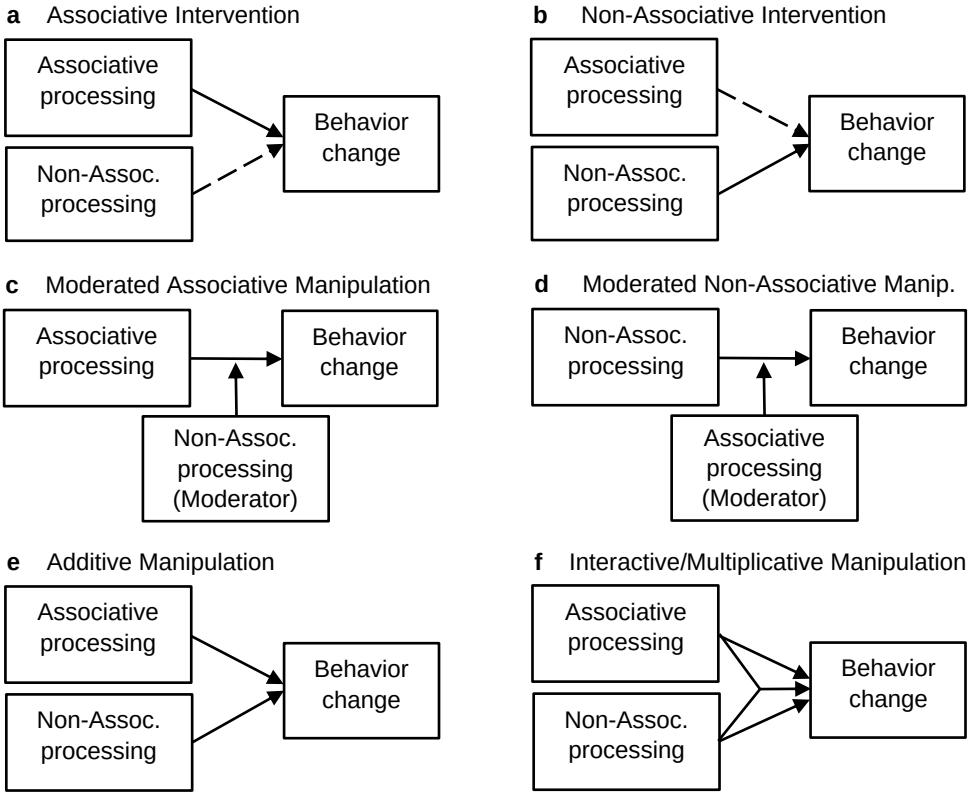

Figure 10. Intervention patterns (a and b) vs. manipulation patterns (c, d, e, and f) adapted from Perugini and colleagues (2010, Figure 1). NOTE: dashed lines denotes a broken causal arrow.

Univariate follow-up experiments. After the exploratory experiments, researchers decide to manipulate something that is widely accepted to involve only one type of processing. They try to manipulate implicitly biased behavior via associative processing by presenting participants with counterstereotypes (i.e., counterconditioning). Analyses of their data reveals that their associative manipulations repeatedly reduced implicitly biases in behavior compared to their control groups.

A few things follow from these results. First, these univariate findings negate only one of the six interpretations from Figure 10 (i.e., b). Second, via the affirmative manipulation principle, researchers can infer that implicitly biased behavior is predicated on associative processing. However, they cannot infer that implicitly biased behavior is not predicated on non-associative processing. Indeed, the researchers did not measure (or manipulate) non-associative processing. So, they cannot analyze the impact of non-associative processing. Therefore, while univariate follow-up experiments tell us something about the nature of implicit bias, they hardly settle the debate about the nature of implicit bias.

Multivariate follow-up experiments. After our univariate experiments, researchers decide to detect interventions on implicitly biased behavior via either associative or non-associative processing. So, they design two kinds of manipulations. The associative manipulation attempts to change implicit biases in behavior by presenting participants with counterstereotypes (i.e., counterconditioning). The non-associative manipulations attempt to change implicit biases in behavior in a way that is compatible only with non-associative processing—see Mandelbaum (2016) and De Houwer (2018) for discussions of such manipulations. Then the researchers randomly assign participants to, associative manipulation conditions, non-associative manipulation conditions, or a control group.

Now, imagine what we can infer if all multivariate follow-up experiments find that only one kind of manipulation condition reduces implicitly biased behavior significantly more than the control condition. First, we can negate all but one of the interpretations from Figure 10 (i.e., either a or b would remain). Second, we can infer—via the negative intervention principle—that implicitly biased behavior is not predicated on any of one type of processing—i.e., that it is

predicated entirely on another type of processing. Thus, multivariate follow-up experiments have the potential to settle the debate about the nature of implicit bias.

However, imagine what we can infer if both manipulation conditions reduce implicitly biased behavior significantly more than the control condition. First, we can negate only two of the six interpretations in Figure 10 (i.e., a and b). Second, via the affirmative manipulation principle, we can infer that implicitly biased behavior can be predicated on both associative and non-associative processing. In other words, while univariate follow-up experiments allow us to infer something about the nature of implicit bias, they do not settle the existing debate about the nature of implicit bias.

5.2.3 What to Infer from Null Results

Another possibility is that debiasing experiments find no manipulations that result in long-lasting, reliable, and significant reductions in implicit bias compared to controls. This is not a far-fetched possibility. Some meta-analyses find that experimental manipulations of implicit bias are "relatively weak" (Forscher et al., 2018).

Weak or even null results in debiasing experiments are to be expected if the average correlation between two IAT scores from the same person, i.e., test-retest reliability, is not high (e.g., $0.45 < r < 0.63$ in Bar-Anon & Nosek, 2014; average $r = 0.54$ in Gawronski, Morrison, Phill, & Galdi, 2017). One might think that this would show that the IAT is an unreliable measure. However, IAT scores could have high internal consistency (e.g., 0.83 and 0.88, ibid.) even if test-retest reliability is low. So, one interpretation of these findings could be that the IAT reliably measures something that is not highly stable over time (Gawronski, 2019; Jost, 2018). Insofar as that is right, it will be more difficult to detect changes in implicit bias that are the result of debiasing manipulations rather than the result of ordinary instability in implicit bias.

Further, insofar as that is right, it will be more difficult to infer anything about the nature of implicit bias from debiasing experiments alone.

5.3 Debiasing Experiments

Philosophers sometimes cite debiasing experiments in favor of their view about implicit bias. As I will argue below, the evidential value of these experiments varies. A few experiments are under-described, making their evidential value indeterminable. Other experiments are adequately described but do not address several methodological concerns, mitigating their evidential value. Only some experiments are scrupulous enough to constitute strong evidence. Naturally, one should infer views of implicit bias from the strong evidence.

5.3.1 Under-described Evidence

Mandelbaum mentions a debiasing experiment which found that variations in argument strength can manipulate implicitly biased behavior (2016, p. 640). The experimenters presented an unspecified quantity of undergraduates with either strong or weak reasons in favor of a new policy to integrate more black professors at their university (Briñol, Petty, & McCaslin, 2009, p. 293). The strong reasons were as follows: “the number and quality of professors would increase with this program (without any tuition increase) [and] the number of students per class would be reduced by 25%” (ibid., 294). The weak reasons were as follows: “the program would allow the university to take part in a national trend and with the new professors, current professors might have more free time to themselves” (ibid.). After participants were presented with these strong or weak reasons for the policy, they were given the Race IAT.

Briñol and colleagues found that participants who were presented with strong reasons for the pro-Black policy were more positive toward Black facial features than participants who received weak reasons. The authors consider this an associative manipulation. However, some

argue that this finding would be difficult to explain via only associative processing (e.g., Mandelbaum, 2016). If that is right, then Briñol and colleagues' manipulation would be non-associative. Alas, even if we grant that, we do not yet have enough information about the finding to know if we should infer anything from it. While this experiment's design has promise, its sample size and other descriptive statistics (e.g., the p-value and the effect size) are not reported—cf. Horcajo, Briñol, & Petty, 2010 for similar experiments with descriptive statistics about non-racial stimuli. Historically, not reporting such relevant details has been common in some social sciences (McCloskey & Ziliak, 1996). However, unless or until the details of such under-described experiments are provided, their evidential significance is indeterminable (ibid.).

5.3.2 Mitigated Evidence

Mandelbaum also references a debiasing experiment which found that differences in peer disagreement can manipulate implicitly biased behavior (2016, p. 641). The experimenters sorted about 50 undergraduate psychology students into a low-bias group and a high-bias group based on their level of racial bias (Sechrist & Stangor, 2001). Once sorted, each group was randomly sorted into two more groups: the high-consensus group was told that 81% of their peers agreed with their judgments about race, the low-consensus group was told that 19% of their peers agreed with their judgments about race. After receiving their peers' feedback, participants were asked to wait in a chair in the hallway just outside the experiment room. The hallway was staged with seven chairs, side-by-side. A black research confederate, "who was unaware of the experimental condition of the participants, sat in the seat closest to the door of the experimental room" (ibid., p. 647; see also Figure 11). In short, students had to choose how close to sit to a black peer right after finding out that either most or few of their peers agree with their racial judgments.

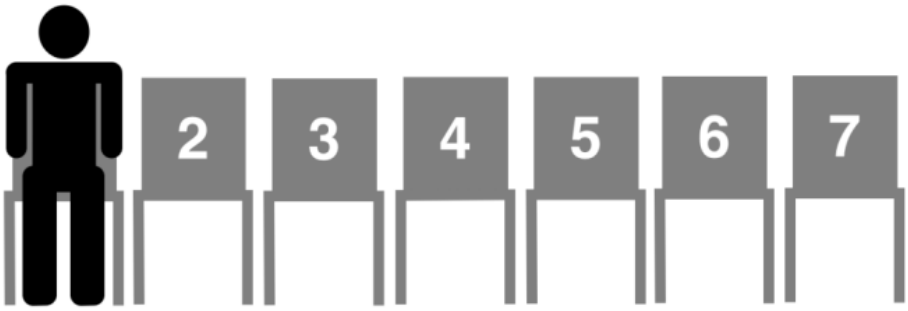

Figure 11. Measure of implicit racial bias from Sechrist and Stangor (2001).

Sechrist and Stangor found that highly biased participants in the high-consensus group sat further away from their black peer than their counterparts in the low-consensus group, $F(1, 50) = 5.65$, $p < 0.05$, suggesting that normalizing the biases of high-bias individuals increases their biased behavior. Lowly biased participants in the high-consensus condition sat closer to their black peer than their counterparts in the low-consensus group, $F(1, 50) = 3.22$, $p < 0.07$, suggesting that normalizing the biases of low-bias individuals decreases their biased behavior.

It is no doubt important to investigate whether this peer feedback manipulation is associative or non-associative, reflective or unreflective. However, even if the nature of this manipulation were discovered, there are a few reasons to resist basing one's view of implicit bias on this particular experiment. First, the analysis might be underpowered, given the sample size of this experiment. A common rule of thumb for sufficient statistical power is to have a minimum of about 50 participants per experimental condition (Simmons, Nelson, & Simonsohn, 2013, 2018)—around 4 times the quantity of participants in each of the aforementioned conditions. A proper power analysis could reveal whether this finding is, in fact underpowered, but that would require more information than is reported—e.g., the standard deviations of seating distances in each group. In lieu of a proper power analysis, some researchers recommend estimating power as

follows: p = .05 → power ≈ .5; p = .01 → power ≈ .75; p = .005 → power ≈ .8; and p = .001 → power > .9 (Greenwald, Gonzalez, Harris, & Guthrie, 1996). Suffice it to say that the seating distance findings are not well powered, according to this estimation. Worse, recent replication attempts suggest that if this estimation errs, it errs on the side of overestimation (Camerer et al., 2018; Open Science Collaboration, 2015). Power aside, one might still be concerned about the statistical significance of the finding. It is either insignificant or marginally significant even according to the older, lower, and now controversial p-value threshold of 0.05 (Benjamin et al., 2018). Also, the reliability of the clever seating measure is not reported. Given the aforementioned psychometric concerns about the IAT, one would want to be reassured about the psychometric validity of this seating measure before accepting the implications of experiments that employ it.

Madva mentions debiasing experiments showing that approach-avoidance behaviors sometimes change implicit racial biases (2017, p. 151). Earlier research had found that repeatedly performing approach behaviors towards photographs of Black people and avoidance behaviors towards photographs of White people resulted in a relative preference for the White people on the race IAT, *F*s(2, 41-69) = 2.93-9.09, *p*s = 0.004-0.06 (Kawakami, Phills, Steele, & Dovidio, 2007, 960-966, Experiments 1 through 4). However, more recent research found that this associative approach and avoidance manipulation did not change race IAT performance compared to controls, $F(1, 60) = 0.0441$, $p = .83$ (Van Dessel, De Houwer, Roets, & Gast, 2016, Experiment 1). Surely failures to replicate earlier findings can constitute progress in the debiasing literature. Indeed, Van Dessel and colleagues claim that this failure to replicate the four earlier findings challenges the idea that implicit biases are "(exclusively) the result of automatic associative learning processes" (Van Dessel et al., 2016, e2)—notice that their

qualificatory use of ‘exclusively’ avoids the science columnists’ any-only mix-up. Nonetheless, Van Dessel and colleagues admit that their null result is “not reliable enough to be treated as conclusive evidence” (ibid., e12) perhaps in part because, as they admit, these studies involve “very small sample[s] of participants” (ibid., e5)—on average, 58 total participants in each of Kawakami and colleagues’ and Van Dessel and colleagues’ experiments about implicit racial bias. So, as Van Dessel and colleagues suggest, more research is required to understand how their manipulations change implicit bias and, thereby, whether implicit bias is predicated exclusively on automatic associations, more reflective associations, or even non-associations.

One might wonder whether the limitations of these experiments apply to other debiasing experiments cited in the debate about the nature of implicit bias. Evaluating the rigor of all relevant debiasing experiments is a worthy inquiry, but it goes beyond the scope of the present paper. The point is just that one should be hesitant to infer a view of implicit bias from the mitigated evidence that is sometimes cited in the debate about the nature of implicit bias.

5.3.3 Strong Evidence

Madva (2017) also mentions more recent, larger, and more methodologically rigorous experiments. One found long-term debiasing while the other found short-term debiasing.

In-person, long-term Debiasing. In one of the experiments, 91 non-Black introductory psychology students (67% female, 85% White) were randomly assigned to either a control or an experimental condition after they took the Race IAT (Devine, Forscher, Austin, & Cox, 2012). Both groups completed the Race IAT and typed their results into a computer that explained their results. Then the control group was dismissed but was told that they would need to fill out questionnaires at two points later in the semester. The experimental group was presented with “a

45-minute narrated and interactive slideshow" that educated participants about implicit bias and trained them in five debiasing strategies (ibid., p. 7).

Devine and colleagues found that the experimental manipulation significantly reduced Race IAT results compared to the control groups, F (88) = 7.95, p = 0.006 (Devine et al., 2012, p. 8). And this reduction in implicit bias was maintained (i.e., was not significantly different) 4 and 8 weeks later — F(88) = 0.67, p = 0.42 (ibid.). These findings suggest that certain strategies can manipulate implicitly biased behavior for extended periods of time.

The strategies that Devine and colleagues' participants learned are as follows:

A. *Stereotype replacement*. Identify the stereotypes that inform our responses and replace them with responses that are not based on stereotypes (Monteith, 1993).

B. *Counter-stereotypic imaging*. Imagine counter-stereotypical exemplars when a stereotype is activated (Blair et al., 2001).

C. *Individuation*. Focus on the individual features of someone rather than the stereotypes about them (Brewer, 1988).

D. *Perspective taking*. Imagine the first-person perspective of a member of a stereotyped group rather than the stereotypes about their group (Galinsky & Moskowitz, 2000).

E. *Increasing opportunities for contact*. Seek out positive experiences with members of other groups rather than let oneself imagine stereotypically negative experiences with members of that group (Pettigrew & Tropp, 2006).

Some of these strategies clearly involve conditioning or counterconditioning negative associations—e.g., counter-stereotypic imaging and increasing opportunities for contact (Helton,

2017). Conditioning and counterconditioning are associative manipulations (Mandelbaum 2016, p. 635).

Notice also how much deliberate processing of conscious representations (i.e., reflection) is involved in these strategies: representing a stereotype as a stereotype, imagining not just a stereotype but a counter-stereotype, focusing on individual rather than group-level features, imagining the first-person experience of someone with racial features that are different from one's own racial features, and interact positively with people that are negatively stereotyped. So, some of these manipulations involve not only associative processing, but reflective processing.

Nonetheless, some of these debiasing strategies are not known to be purely associative or purely non-associative—e.g., individuation. So, just like in the imagined exploratory experiments, we are left unable to interpret the roles that associative and non-associative processing play in some of these debiasing strategies.

Online, short-term debiasing. In another experiment, around 5000 participants from 17 universities in the United States were randomly assigned to 1 of 9 debiasing conditions or a control condition, after they took the Race IAT (Lai et al., 2016, Study 2). Immediately after completing their condition's requirements, participants took a post-test Race IAT. And two to four days after the experiment, participants completed a second post-test Race IAT.

Lai and colleagues found that eight of the nine debiasing conditions significantly reduced implicitly biased behavior more than the control condition, Fs(1, 1000-1045) = 6.16-286.73, ps = 0.001-0.013 (ibid., pp. 1009-10). Alas, when participants retook the Race IAT two to four days later "[n]one of the interventions had significantly reduced IAT scores relative to control" (Ibid., p. 1010). So, like the previous experiment, the debiasing strategies changed implicitly biased behavior. Yet, unlike the last experiment, the changes did not last.

The debiasing conditions employed by Lai and colleagues debiasing conditions were as follows:

F. *Vivid counterstereotypic scenario*. Read a vivid story about a White villain and a Black hero (Dasgupta & Greenwald, 2001) and keep that in mind during the post-manipulation IAT.

G. *Counterstereotypic IAT*. Practice 32 trials of the IAT in which Black is paired with Good and White is paired with Bad, including some famously positive Black figures such as Oprah and some famously negative White figures such as Hitler (Joy-Gaba & Nosek, 2010).

H. *Competition with shifted group boundaries*. Play a simulated dodgeball game in which one's own teammates are Black and play well and one's opponents are White and play poorly.

I. *Shifting group affiliations under threat*. Read a vivid story about the threat of postnuclear war in which one's closest friends are Black and helpful and one's enemies are White.

J. *Priming multiculturalism*. Read a pro-multiculturalism excerpt, summarize it in one's own words, and list reasons that multi-culturalism improves group relations (Richeson & Nussbaum, 2004).

K. *Evaluative conditioning*. Observe 20 Black faces paired with positive words and 20 White faces paired with negative words.

L. *Evaluative conditioning with Go/No-Go task*. Press a button when a Black face is paired with a positive word, do not press a button with a Black face is paired with a

negative word, and count the number of Black-positive pairings (Nosek & Banaji, 2001).

M. *Implementation intentions*. Learn that one can override bias by thinking of conditional intentions like, "If I see a Black face, then I will respond by thinking 'good'" (Gollwitzer, 1999).

N. *Faking the IAT*. Learn about Pro-White biases on the IAT and how to intentionally manipulate one's response times in order for the test to detect a Pro-Black bias (Cvencek, Greenwald, Brown, Gray, & Snowden, 2010).

All conditions except the priming multiculturalism condition clearly involve pairing racial stimuli and valences—i.e., conditioning or counterconditioning. So, at least 8 of Lai and colleagues' manipulations seem to be associative (Mandelbaum 2016, p. 635). Whether the priming multiculturalism condition counts as purely non-associative will be controversial given the disagreement about whether so-called logical and evidential manipulations involve associative processing (e.g., Briñol et al., 2009 vs. Mandelbaum, 2016). Until such disagreement is resolved or at least uncontroversial and until it is clear that priming multiculturalism can produce lasting changes in implicit bias, experiments that prime multiculturalism are exploratory debiasing experiments and therefore unable to determine whether implicit bias is associative or non-associative.

Notice also that some of Lai and colleagues' associative manipulations seem to involve deliberate processing of conscious representations—e.g., pre-emptively thinking "Black = Good" (2016, pp. 1005). This suggests that some of Lai and colleagues' associative manipulations involved reflection.

Duration of experiment & debiasing. The evidence for long-term debiasing was mixed. Devine and colleagues found that counterconditioning-like protocols produced short- and long-term changes in implicitly biased behavior, but Lai and colleagues found that counterconditioning resulted in only short-term changes. Three details about this mixed evidence are worth emphasizing.

First, consider the conflict about long-term findings. Given that Lai and colleagues' sample was much larger and was collected from multiple locations, its population is more representative, and its analysis confers greater statistical power. So, if one had to bet on the likelihood of long-term debiasing via counterconditioning across populations, then one should bet against them—until further debiasing experiments suggest otherwise, of course. Nonetheless, we might wonder if this conflict is only apparent. That is, perhaps long-term debiasing works in certain subsets of the population—like Devine and colleagues' non-Black, mostly-White undergraduate sample—even if long-term debiasing does not work, on average, across the population as a whole. Of course, this is an empirical hypothesis.

Second, the conflict about long-term findings might be related to the duration, frequency, or even context of the counterconditioning manipulations. Notably, previous work found that 5 minutes of counterconditioning in a controlled setting did not significantly change implicitly biased behavior, but four blocks of 96 trials of counterconditioning did (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000). Similarly, Lai and colleagues found that short, online debiasing protocols did not lead to long-term changes, but Devine and colleagues found that teaching students how debiasing works in everyday social settings did. Taken together, one might hypothesize that long-term debiasing is more likely with further counterconditioning (Van

Dessel, De Houwer, Roets, & Gast, 2016, e12)—and not just in lab settings, but in everyday social settings (Madva, 2017).

Third, notice a consistency between Devine and colleagues' and Lai and colleagues' findings: more or less reflective counterconditioning changed implicitly biased behavior, even if only briefly. This finding is also consistent with earlier work (e.g., Olson & Fazio, 2006, Experiment 2; Rydell & McConnell, 2006).

5.4 Critical Discussion

With the inferential principles and the evidence on the table, we are now prepared to determine whether the received view of implicit bias should be abandoned for more centrist or far-right views (Table 12). I will argue that the received, unreflective, associationist view of implicit bias should be abandoned for a more general associationist view of implicit bias, given the strong evidence just considered. However, I will concede that if certain evidence exists, now or in the future, even this more general associationist view of implicit bias should be abandoned for either interactionism about implicit bias or minimalism about implicit bias—views that will be explained below.

5.4.1 Associationism & Reflectivism About Implicit Bias

The strongest evidence found that conditioning or counterconditioning can change implicitly biased behavior—even if only briefly. Conditioning and counterconditioning are widely accepted to be associative manipulations (Mandelbaum 2016, p. 635).

Associationism. Via the affirmative manipulation principle, one can infer that implicit bias is at least partly associative. It may be tempting to conclude from this that associationism about implicit bias is true and, therefore, that non-associationism about implicit bias is false. However, that relies on the problematic negative manipulation principle that leads to the science

columnists' any-only mix-up: in this case, the mistake of concluding that implicit bias is not predicated on any non-associative processes when the evidence merely shows that something is not predicated on only non-associative processes.

Reflectivism. To test for positive evidence that implicit bias is also partly non-associative, some debiasing experiments dissociate the effect of associative processing on implicit bias from the effect(s) of other kinds of processing (e.g., Calanchini, Gonsalkorale, Sherman, & Klauer, 2013). These process dissociation debiasing experiments find that debiasing is explained by both (a) the degree to which associations are activated and (b) the degree to which participants reflect on appropriate responses. The fact that reflection can change implicitly biased behavior is consistent with the strongest evidence under consideration. If reflection were necessarily non-associative, then the negative intervention principle would allow us to infer from this evidence that implicit bias is not predicated on only associative processes. However, many have realized that reflection is not necessarily non-associative.

Nonetheless, the fact that reflection can help reduce implicitly biased behavior supports a sort of reflectivism about implicit bias. Reflectivism is just the idea that reflection is an important part of improving our judgments and behavior (Doris, 2015; Ferrin, 2017). And reflection seems to be involved in counterconditioning implicit bias. This is not to say that there is strong evidence for an infallibilist reflectivism, according to which reflection fully or permanently ameliorates implicit bias. Rather, the strong evidence suggests only a kind of "sensible reflectivism" according to which reflection can—but does not necessarily—ameliorate implicit bias, albeit only briefly and incompletely (Schwenkler, 2018). If that is right, then we can infer, via the affirmative manipulation principle that implicit bias can be reflective.

5.4.2 Interactionism & Minimalism About Implicit Bias

This paper has focused on implicit racial bias and on three categories of evidence from experimental attempts to reduce such biases. Given how many debiasing experiments have been conducted and how thoroughly one should analyze these experiments, one paper cannot sufficiently review all racial debiasing experiments—let alone all debiasing experiments. So, there may be strong evidence, now or in the future, of non-associative interventions on implicitly biased behavior. If or when such evidence exists, then associationism about implicit bias would be false: i.e., implicitly biased behavior would not be predicated on only associative processing.

Of course, falsifying associationism about implicit bias would not support non-associationism about implicit bias, given the strong evidence already considered. That is, if we add non-associative interventions on implicitly biased behavior to our body of strong evidence, then our total evidence would suggest that implicit bias can be changed via associative processes, given the strong evidence considered herein, as well as non-associative processes, given the additional evidence. According to the affirmative manipulation principle, that total evidence entails that implicit bias can be predicated on either associative or non-associative processes. Such a disjunctive conclusion brings us to a fork in the road.

Interactionism. Down one side of the fork, there are interactionist views of implicit bias. Interactionist views of implicit bias accept that implicit bias is predicated on associative and non-associative processes. However, interactionist views also aim to describe precisely how these processes interact to produce the observed dynamics of implicitly biased behavior (e.g., by testing for manipulation patterns matching c, d, e, and f in Figure 10). In other words, the goal of an interactionist view is its attempt to construct and test cognitive models of implicit bias. There

are a variety of interactionist views that seem to accomplish this goal (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Gawronski & Bodenhausen, 2014; Perugini, 2005).

Minimalism. Interactionist views of implicit bias include more cognitive details than are necessary for some philosophers' claims about implicit bias. So, some philosophers can go down the other side of the fork toward minimalist views of implicit bias. Minimalist views accept the complexity of implicit bias. They acknowledge that implicit bias can seem associative in some circumstances and seem non-associative in other circumstances. Minimalist views of implicit bias also acknowledge that implicit bias seems to involve more reflection in some circumstances and less reflection in other circumstances. Crucially, however, minimalist views of implicit bias do not aim to provide a falsifiable account of whether and how implicit bias is predicated on certain types of cognitive processing. Rather, the goal of minimalism about implicit bias is to account for our normative intuitions about cases of implicit bias without relying on any particular cognitive model of implicit bias. There are various discussions of implicit bias that might be able to accomplish these goals (e.g., Levy, 2016; Smith, 2018; Sullivan-Bissett, 2015).

Of course, the goals of minimalist views and interactionist views are not mutually exclusive. After all, while some normative intuitions about implicit bias need not commit to any particular cognitive model of implicit bias, some cognitive models of implicit bias, if justified, will justify some normative intuitions about implicit bias more than others (e.g., Huebner, 2016; Toribio, 2018a). So, there are advantages to basing normative claims about implicit bias on the most promising interactionist views of implicit bias. And some philosophers seem to realize as much (e.g., Berger, 2018; Holroyd & Sweetman, 2016; Madva, 2017; Levy, 2015).

5.5 Conclusion

On the one hand, the conclusion of this investigation is somewhat progressive. The stronger debiasing evidence under consideration did not support the received, unreflective, associationist views of implicit bias. On the contrary, there was strong evidence that implicitly biased behavior can also be changed via more reflective processing, supporting the more capacious associationist views of implicit bias. On the other hand, the conclusion of this investigation is somewhat conservative. While the received view of implicit bias was only partially supported by strong debiasing experiments, the road to far-right, non-associationist views of implicit bias remained blocked by strong evidence of associative debiasing manipulations.

Nonetheless, future or overlooked evidence might clearly show that non-associative manipulations change implicitly biased behavior. If that is the case, then more centrist, interactionist views of implicit bias can be inferred. Otherwise, two views of implicit bias can be inferred from debiasing experiments: first, associationist views of implicit bias and second, minimalist views if implicit bias.

Of course, there are limitations to the existing investigation. First, views about implicit bias are based on more than just debiasing experiments. So, the conclusions about implicit bias that were inferred from debiasing experiments herein are not all-things-considered conclusions. As such, views of implicit bias may be further supported or further undermined by considerations besides debiasing experiments. Second, the debate between associationism or non-associationism about implicit bias is just a specific instance of the more general debate between associationism or non-associationism about mind. So, this investigation cannot settle that general debate. However, this investigation can recommend that debaters avoid the science

columnists' any-only mix-up by relying on the negative intervention principle rather than the negative manipulation principle.

APPENDIX A

IRB APPROVAL & CONSENT FORM

Office of the Vice President for Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 09/12/2018
From: Thomas L. Jacobson, Chair
Re: Use of Human Subjects in Research
Thinking Styles & Philosophy

To: Nick Byrd < @my.fsu.edu>
Address: Nick Byrd
Dept.: PHILOSOPHY DEPARTMENT

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Expedited per 45 CFR § 46.110(7) and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 09/11/2019 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

HSC No. 2018.25325

Title of Project: Thinking Styles and Philosophy
Investigators: Nick Byrd at Florida State University

This research examines how thinking styles relate to decisions in difficult situations. Participation involves reading about situations people may experience and making decisions about what should be done, as well as answering some questions about your personal thinking style. The survey will take about 20 minutes. Participants must be between 18-65 years old in order to participate.

Participation is entirely voluntary. You are free to end participation at any time, without any penalty. You will learn about the purpose of the study at the end.

All responses will remain confidential to the extent allowed by law. You will not be asked to provide identifying information, and we will report only group findings. All data will all be stored on password-protected computers at Florida State University.

There is a possibility of a minimal level of risk involved in this study. Participants might experience minor discomfort when making decisions about difficult situations. Please inform the experimenter immediately if you experience any discomfort at @fsu.edu. You are welcome to contact Nick Byrd to discuss any concerns you may have.

There are also benefits to participating in this research project: beyond valuable insight into the psychological processes involved in decision-making, you may find the experience interesting and informative.

You have the right to ask and have answered any questions concerning the study. You may contact Nick Byrd at @fsu.edu or - - for answers to questions about this research, or my rights, or results. If you have questions about your rights as a participant in this research, or feel you have been placed at risk, you may contact Florida State University's Human Subjects Office (Phone: - - , Email: @fsu.edu).

☐ I have read the Letter of Information, have had the nature of the study explained to me, and I agree to participate. All questions have been answered to my satisfaction.

Name: _____

Signature: _____

FSU Human Subjects Committee approved on 09/12/2018. Void after 09/11/2019. HSC No. 2018.25325

REFERENCES

Aarnio, K., & Lindeman, M. (2005). Paranormal beliefs, education, and thinking styles. Personality and Individual Differences, 39(7), 1227–1236.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. Psychonomic Bulletin & Review, 19(6), 1187–1192. DOI: 10.3758/s13423-012-0305-z

Alper, S., & Yilmaz, O. (2019). How is the Big Five related to moral and political convictions: The moderating role of the WEIRDness of the culture. Personality and Individual Differences, 145, 32–38. DOI: 10.1016/j.paid.2019.03.018

Anderson, J. R., & Bower, G. H. (1980). Human Associative Memory. New Jersey: Lawrence Erlbaum Associates, Inc.

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. Journal of the Economic Science Association, 3(1), 1–11. DOI: 10.1007/s40881-017-0035-0

Bächtiger, A., Dryzek, J. S., Mansbridge, J., & Warren, M. E. (2018). The Oxford Handbook of Deliberative Democracy. Oxford University Press.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. Cognition, 158, 90–109. DOI: 10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. Thinking & Reasoning, 1–43. DOI: 10.1080/13546783.2018.1507949

Banaji, M. R., & Hardin, C. D. (1996). Automatic Stereotyping. Psychological Science, 7(3), 136–141. DOI: 10.1111/j.1467-9280.1996.tb00346.x

Bargh, J. A. (1992). The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects. The American Journal of Psychology, 105(2), 181–199. DOI: 10.2307/1423027

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44(3), 211–233. DOI: 10.1016/0001-6918(80)90046-3

Baron, J. (1994). Nonconsequentialist decisions. Behavioral and Brain Sciences, 17(1), 1–10. DOI: 10.1017/S0140525X0003301X

Baron, J. (2018). Actively Open-minded Thinking Scale. Retrieved March 15, 2019, from The Society for Judgment and Decision Making website: sjdm.org/dmidi/Actively_Open-Minded_Thinking_Beliefs.html

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? Journal of Applied Research in Memory and Cognition, 4(3), 265–284. DOI: 10.1016/j.jarmac.2014.09.003

Benjafield, J. (1969). Logical and empirical thinking in a problem solving task. Psychonomic Science, 14(6), 285–286. DOI: 10.3758/BF03329126

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6. DOI: 10.1038/s41562-017-0189-z

Berger, J. (2018). Implicit attitudes and awareness. Synthese, 1–22. DOI: 10.1007/s11229-018-1754-3

Berko, J. (1958). The Child's Learning of English Morphology. WORD, 14(2–3), 150–177. DOI: 10.1080/00437956.1958.11659661

Białek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. Behavior Research Methods, 50(5), 1953–1959. DOI: 10.3758/s13428-017-0963-x

Bishop, M. A., & Trout, J. D. (2004). Epistemology and the Psychology of Human Judgment. New York: Oxford University Press.

Bishop, M. A., & Trout, J. D. (2008). Strategic Reliabilism: A Naturalistic Approach to Epistemology. Philosophy Compass, 3(5), 1049–1065. DOI: 10.1111/j.1747-9991.2008.00161.x

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. Journal of Personality and Social Psychology, 81(5), 828–841. DOI: 10.1037/0022-3514.81.5.828

Blake, A. (2016, September 26). The first Trump-Clinton presidential debate transcript, annotated. Washington Post. Retrieved from washingtonpost.com/news/the-fix/wp/2016/09/26/the-first-trump-clinton-presidential-debate-transcript-annotated/

Blanton, H., Jaccard, J., & Burrows, C. N. (2015). Implications of the Implicit Association Test D-Transformation for Psychological Assessment. Assessment, 22(4), 429–440. DOI: 10.1177/1073191114551382

Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227–247. DOI: 10.1017/S0140525X00038188

Bonnefon, J.-F. (2016). The Pros and Cons of Identifying Critical Thinking with System 2 Processing. Topoi, 1–7. DOI: 10.1007/s11245-016-9375-2

Bourget, D., & Chalmers, D. (2014). What do philosophers believe? Philosophical Studies, 170(3), 465–500. DOI: 10.1007/s11098-013-0259-7

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull, R. S. Wyer, & Jr. (Eds.), A dual process model of impression formation (pp. 1–36). Retrieved from psycnet.apa.org/record/1988-98192-001

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference. In R. H. Fazio, R. E. Petty, & P. Brinol (Eds.), Attitudes: Insights from the new implicit measures (pp. 285–326). Psychology Press.

Brownstein, M. (2018). The Implicit Mind: Cognitive Architecture, the Self, and Ethics. New York, NY: Oxford University Press.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. Synthese, 195(12), 5339–5372. DOI: 10.1007/s11229-018-01949-1

Buckner, C. (2019). Rational Inference: The Lowest Bounds. Philosophy and Phenomenological Research, 98(3), 697–724. DOI: 10.1111/phpr.12455

Bulbrook, M. E. (1932). An Experimental Inquiry into the Existence and Nature of "Insight." The American Journal of Psychology, 44(3), 409–453. DOI: 10.2307/1415348

Byrd, N. (2014). Intuitive and Reflective Responses in Philosophy. University of Colorado.

Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. Synthese, (Online first), 1–29. DOI: 10.1007/s11229-019-02128-6

Byrd, N., & Conway, P. (2019). Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. Cognition, 192, 103995. DOI: 10.1016/j.cognition.2019.06.007

Byrd, N., Gongora, G., Joseph, B., & Sirota, M. (Forthcoming). Tell Us What You Really Think: A Think Aloud Protocol Analysis of the Verbal Cognitive Reflection Test.

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. European Journal of Social Psychology, 43(5), 321–325. DOI: 10.1002/ejsp.1941

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 1. DOI: 10.1038/s41562-018-0399-z

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. Memory & Cognition, 42(3), 434–447. DOI: 10.3758/s13421-013-0367-9

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. Judgment and Decision Making, 5(3), 182–191.

Capraro, V., & Peltola, N. (2018). Lack of deliberation drives honesty among men but not women. ArXiv:1805.08316 [Physics, q-Bio]. Retrieved from arxiv.org/abs/1805.08316

Carley, L. (2018, October 31). Breaking the Bias Habit: An Evidence-Based Intervention in Duke's Biology Department | Duke Graduate School. Retrieved November 6, 2018, from Duke—The Graduate School: Professional Development Blog website: gradschool.duke.edu/professional-development/blog/breaking-bias-habit-evidence-based-intervention-duke-s-biology

Carnap, R. (1950a). Empiricism, semantics, and ontology. Revue Internationale de Philosophie, 4(11), 20–40.

Carnap, R. (1950b). Logical Foundations of Probability. Chicago: University of Chicago Press.

Čavojová, V., Secară, E.-C., Jurkovič, M., & Šrol, J. (2019). Reception and willingness to share pseudo-profound bullshit and their relation to other epistemically suspect beliefs and cognitive ability in Slovakia and Romania. Applied Cognitive Psychology. DOI: 10.1002/acp.3486

Chalmers, D. J. (2014). Intuitions in philosophy: A minimal defense. Philosophical Studies, 171(3), 535–544. DOI: 10.1007/s11098-014-0288-x

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016a). Mind-wandering as spontaneous thought: A dynamic framework. Nature Reviews Neuroscience, 17(11), 718–731. DOI: 10.1038/nrn.2016.113

Clarke, S. (2013). Intuitions as Evidence, Philosophical Expertise and the Developmental Challenge. Philosophical Papers, 42(2), 175–207. DOI: 10.1080/05568641.2013.806287

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self. Personality and Social Psychology Bulletin, 26(9), 1151–1164. DOI: 10.1177/01461672002611011

Cohen, S. L., & Bunker, K. A. (1975). Subtle effects of sex role stereotypes on recruiters' hiring decisions. Journal of Applied Psychology, 60(5), 566–572. DOI: 10.1037/0021-9010.60.5.566

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. Judgment and Decision Making, 7(1), 25. Retrieved from search.proquest.com/docview/1011295450/abstract/88072502EA77452DPQ/1

Colaço, D., Kneer, M., Alexander, J., & Machery, E. (2016). On second thought: A refutation of the reflection defense. Presented at the Buffalo Experimental Philosophy Conference.

Comalli, P. E., Wapner, S., & Werner, H. (1962). Interference effects of Stroop Color-Word Test in childhood, adulthood, and aging. The Journal of Genetic Psychology; Provincetown, Mass., Etc., 100(1), 47–53. Retrieved from search.proquest.com/docview/1297158987/citation/55CAC2957F3C44E4PQ/1

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Chapter Three - Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised. In J. M. Olson (Ed.), Advances in Experimental Social Psychology (Vol. 56, pp. 131–199). DOI: 10.1016/bs.aesp.2017.03.001

Congreve, W. (1797). The Mourning Bride: A Tragedy. J. Bell.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance. Journal of Personality and Social Psychology, 89(4), 469–487. DOI: 10.1037/0022-3514.89.4.469

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. Journal of Personality and Social Psychology, 104(2), 216–235. DOI: 10.1037/a0031021

Corneille, O., & Stahl, C. (2018). Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models. Personality and Social Psychology Review, 1088868318763261. DOI: 10.1177/1088868318763261

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., … Zhou, X. (2018). Estimating the Reproducibility of Experimental Philosophy. Review of Philosophy and Psychology, 1–36. DOI: 10.1007/s13164-018-0400-9

Cullen, S. (2010). Survey-Driven Romanticism. Review of Philosophy and Psychology, 1(2), 275–296. DOI: 10.1007/s13164-009-0016-1

Cushman, F. (2019). Rationalization is rational. Behavioral and Brain Sciences, 1–69. DOI: 10.1017/S0140525X19001730

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test Is Statistically Detectable and Partly Correctable. Basic and Applied Social Psychology, 32(4), 302–314. DOI: 10.1080/01973533.2010.519236

Dacey, M. (2016). Rethinking associations in psychology. Synthese, 193(12), 3763–3786. DOI: 10.1007/s11229-016-1167-0

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. Journal of Personality and Social Psychology, 81(5), 800–814. DOI: 10.1037//0022-3514.81.5.800

Dasgupta, Nilanjana, & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. Journal of Experimental Social Psychology, 40(5), 642–658. DOI: 10.1016/j.jesp.2004.02.003

Davies, R., Ives, J., & Dunn, M. (2015). A systematic review of empirical bioethics methodologies. BMC Medical Ethics, 16(1), 15. DOI: 10.1186/s12910-015-0010-3

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. Personality and Social Psychology Review, 15(4), 332–351.

De Cruz, H. (2014a). The Enduring Appeal of Natural Theological Arguments. Philosophy Compass, 9(2), 145–153.

De Cruz, H. (2014b). Where Philosophical Intuitions Come From. Australasian Journal of Philosophy, 0(0), 1–17. DOI: 10.1080/00048402.2014.967792

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. Learning and Motivation, 37(2), 176–187. DOI: 10.1016/j.lmot.2005.12.002

De Houwer, J. (2018). Propositional Models of Evaluative Conditioning. Social Psychological Bulletin, 13(3), e28046. DOI: 10.5964/spb.v13i3.28046

De Keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2019). Investigating the Robustness of the Illusory Truth Effect Across Individual Differences in Cognitive Ability, Need for Cognitive Closure, and Cognitive Style. Personality and Social Psychology Bulletin, 0146167219853844. DOI: 10.1177/0146167219853844

De Neys, W. (Ed.). (2017). Dual Process Theory 2.0 (1 edition). Routledge.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. Cognition, 113(1), 45–61. DOI: 10.1016/j.cognition.2009.07.009

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. Current Directions in Psychological Science, 0963721419855658. DOI: 10.1177/0963721419855658

de Thévenot, J. (1687). The travels of Monsieur de Thevenot into the Levant.

Del Pinal, G. D., & Spaulding, S. (2018). Conceptual centrality and implicit bias. Mind & Language, 33(1), 95–111. DOI: 10.1111/mila.12166

Dennett, D. C. (1969). Content and Consciousness. Routledge & Kegan Paul PLC.

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. Judgment and Decision Making, 10(4), 314–331.

Deutsch, R., Gawronski, B., & Hofmann, W. (2017). Reflection and Impulse: A Framework for Basic Research and Applied Science. In R. Deutsch, B. Gawronski, & W. Hofmann (Eds.), Reflective and Impulsive Determinants of Human Behavior. New York, NY: Psychology Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology, 56(1), 5–18. DOI: 10.1037/0022-3514.56.1.5

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. Journal of Experimental Social Psychology, 48(6), 1267–1278. DOI: 10.1016/j.jesp.2012.06.003

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., … Zinger, J. F. (2018). At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. Perspectives on Psychological Science, 1745691617746796. DOI: 10.1177/1745691617746796

Doris, J. M. (2015). Talking to Our Selves: Reflection, Ignorance, and Agency. OUP Oxford.

Dreyfus, H. (1986). Mind Over Machine. Oxford: Blackwell Publishers.

Duncker, K. (1926). A Qualitative (Experimental and Theoretical) Study of Productive Thinking (Solving of Comprehensible Problems). Pedagogical Seminary and Journal of Genetic Psychology; Provincetown, Mass., Etc., 33, 642–708. Retrieved from search.proquest.com/docview/1297221603/citation/381F4FDDC3564CB9PQ/1

Dutilh Novaes, C. (2018). Carnapian explication and ameliorative analysis: A systematic comparison. Synthese, 1–24. DOI: 10.1007/s11229-018-1732-9

Earp, B. D., Demaree-Cotton, J., Dunn, M., Dranseika, V., Everett, J. A. C., Feltz, A., … Tobia, K. (in press). Experimental Philosophical Bioethics. AJOB Empirical Bioethics.

Easton, C. (2018). Women and 'the philosophical personality': Evaluating whether gender differences in the Cognitive Reflection Test have significance for explaining the gender gap in Philosophy. Synthese. DOI: 10.1007/s11229-018-01986-w

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. American Psychologist, 49(8), 709.

Ericsson, A. (2003). Valid and Non-Reactive Verbalization of Thoughts During Performance of Tasks Towards a Solution to the Central Problems of Introspection as a Source of Scientific Data. Journal of Consciousness Studies, 10(9–10), 9–10.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. The Cambridge Handbook of Expertise and Expert Performance, 223–242.

Ericsson, K. A. (2018). Capturing Expert Thought with Protocol Analysis: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), The Cambridge Handbook of Expertise and Expert Performance (2nd ed., pp. 192–212). DOI: 10.1017/9781316480748.012

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87(3), 215–251. DOI: 10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1984). Protocol Analysis: Verbal Reports as Data. MIT Press.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. Mind, Culture, and Activity, 5(3), 178–186. DOI: 10.1207/s15327884mca0503_3

Evans, J., & Frankish, K. (Eds.). (2009). In two minds: Dual processes and beyond. New York, NY, US: Oxford University Press.

Evans, J. S. B. T. (2009). How Many Dual Process Theories Do We Need: One, Two or Many? In J. S. B. T. Evans & K. Frankish (Eds.), In Two Minds: Dual Processes and Beyond (pp. 31–54). Oxford: Oxford University Press.

Evans, J. S. B. T. (2013). Reasoning, Rationality and Dual Processes: Selected works of Jonathan St B.T. Evans. Psychology Press.

Evans, J. S. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. Thinking & Reasoning, 0(0), 1–33. DOI: 10.1080/13546783.2019.1623071

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. Thinking & Reasoning, 13(4), 321–339. DOI: 10.1080/13546780601008825

Evans, J., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition Advancing the Debate. Perspectives on Psychological Science, 8(3), 223–241. DOI: 10.1177/1745691612460685

Everett, J. A., Clark, C. J., Luguri, J., Earp, B., Ditto, P., & Shariff, A. (2018). Political differences in free will belief are driven by differences in moralization. Manuscript under Review.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. DOI: 10.3758/BF03193146

Feltz, A., & Cokely, E. T. (2008). The Fragmented Folk: More Evidence of Stable Individual Differences in Moral Judgments and Folk Intuitions. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society. (pp. 1771–1776). Cognitive Science Society.

Feltz, Adam, & Cokely, E. (2019). Extraversion and compatibilist intuitions: A ten-year retrospective and meta-analyses. Philosophical Psychology, 32(3), 388–403. DOI: 10.1080/09515089.2019.1572692

Feltz, Adam, & Cokely, E. T. (2012). The Philosophical Personality Argument. Philosophical Studies, 161(2), 227–246. DOI: 10.1007/s11098-011-9731-4

Ferreira, M. B., Mata, A., Donkin, C., Sherman, S. J., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. Memory & Cognition, 44(7), 1050–1063. DOI: 10.3758/s13421-016-0618-7

Ferrin, A. (2017). Good Moral Judgment and Decision-Making Without Deliberation. The Southern Journal of Philosophy, 55(1), 68–95. DOI: 10.1111/sjp.12210

Figdor, C. (2018). Pieces of Mind: The Proper Domain of Psychological Predicates. Oxford University Press.

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. Psychology and Aging, 25(2), 271–288. DOI: 10.1037/a0019106

Fischer, E., & Engelhardt, P. E. (forthcoming). Lingering stereotypes: Salience bias in philosophical argument. Mind & Language. DOI: 10.1111/mila.12249

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5–15.

Foot, P. (1995). Moral dilemmas revisited. In W. Sinnott-Armstrong (Ed.), Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus (pp. 117–128). Cambridge: Cambridge University Press.

Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2018). A Meta-Analysis of Procedures to Change Implicit Measures. Retrieved from psyarxiv.com/dv8tu/

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. Psychological Bulletin, 137(2), 316–344. DOI: 10.1037/a0021663

Frank, F. D., & Drucker, J. (1977). The influence of evaluatee's sex on evaluations of a response on a managerial selection instrument. Sex Roles, 3(1), 59–64. DOI: 10.1007/BF00289690

Frankish, K. (2004). Mind and Supermind. Cambridge University Press.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. Philosophy Compass, 5(10), 914–926. DOI: 10.1111/j.1747-9991.2010.00330.x

Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19(4), 25–42. DOI: 10.1257/089533005775196732

Fridland, E. (2016). Skill and motor control: Intelligence all the way down. Philosophical Studies, 174(6), 1539–1560. DOI: 10.1007/s11098-016-0771-7

Fridland, E. (2017). Automatically minded. Synthese, 194(11), 4337–4363. DOI: 10.1007/s11229-014-0617-9

Fridland, E. (2019). Longer, smaller, faster, stronger: On skills and intelligence. Philosophical Psychology, 32(5), 760–784. DOI: 10.1080/09515089.2019.1607275

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender Differences in Responses to Moral Dilemmas A Process Dissociation Analysis. Personality and Social Psychology Bulletin, 41(5), 696–713. DOI: 10.1177/0146167215575731

Gaertner, S. L., & Dovidio, J. F. (1986). The Aversive Form Of Racism. In J. F. Dovidio & S. L. Gaertner (Eds.), Prejudice, Discrimination, and Racism (pp. 61–89). San Diego: Academic Press.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics. Social Psychology Quarterly, 46(1), 23–30. DOI: 10.2307/3033657

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. Journal of Personality and Social Psychology, 78(4), 708–724. DOI: 10.1037/0022-3514.78.4.708

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. Perspectives on Psychological Science, 14(4), 574–595. DOI: 10.1177/1745691619826015

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. Journal of Personality and Social Psychology, 113(3), 343–376. DOI: 10.1037/pspa0000086

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. Psychological Bulletin, 132(5), 692.

Gawronski, B., & Bodenhausen, G. V. (2014). The associative—propositional evaluation model: Operating principles and operating conditions of evaluation. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), Dual-process theories of the social mind (pp. 188–203). New York, NY, US: Guilford Press.

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. Journal of Experimental Social Psychology, 43(2), 221–232. DOI: 10.1016/j.jesp.2006.04.001

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. Handbook of Research Methods in Social and Personality Psychology, 2.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. Personality and Social Psychology Bulletin, 43(3), 300–312. DOI: 10.1177/0146167216684131

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. Journal of Experimental Social Psychology, 41(6), 618–626. DOI: 10.1016/j.jesp.2004.10.005

Gendler, T. S. (2008a). Alief and belief. Journal of Philosophy, 105(10), 634.

Gendler, T. S. (2008b). Alief in action (and reaction). Mind & Language, 23(5), 552–585.

Gershman, S. J. (2018). How to never be wrong. Psychonomic Bulletin & Review, 1–16. DOI: 10.3758/s13423-018-1488-8

Gervais, W. M., van Elk, M., Xygalatas, D., McKay, R. T., Aveyard, M., Buchtel, E. E., … Bulbulia, J. (2018). Analytic atheism: A cross-culturally weak and fickle phenomenon? Judgment and Decision Making, 13(3), 268–274. Retrieved from econpapers.repec.org/article/jdmjournl/v_3a13_3ay_3a2018_3ai_3a3_3ap_3a268-274.htm

Gette, C. R., & Kryjevskaia, M. (2019). Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses. Physical Review Physics Education Research, 15(1), 010118. DOI: 10.1103/PhysRevPhysEducRes.15.010118

Gianotti, L. R., Mohr, C., Pizzagalli, D., Lehmann, D., & Brugger, P. (2001). Associative processing and paranormal belief. Psychiatry and Clinical Neurosciences, 55(6), 595–603.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102, 74–78. DOI: 10.1016/j.paid.2016.06.069

Goethals, G. R., & Reckman, R. F. (1973). The perception of consistency in attitudes. Journal of Experimental Social Psychology, 9(6), 491–501. DOI: 10.1016/0022-1031(73)90030-9

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. American Psychologist, 54(7), 493–503. DOI: 10.1037/0003-066X.54.7.493

Goodman, N. (1983). Fact, Fiction, and Forecast. Harvard University Press.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37(6), 504–528. DOI: 10.1016/S0092-6566(03)00046-1

Graham, J., & Haidt, J. (2010). Beyond Beliefs: Religions Bind Individuals Into Moral Communities. Personality and Social Psychology Review, 14(1), 140–150. DOI: 10.1177/1088868309353415

Greene, J. D. (2013). Moral Tribes: Emotion, Reason, and the Gap Between Us and Them. Penguin.

Greenwald, A. G., Andrew, T., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology, 97(1), 17–41. DOI: 10.1037/a0015575

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27. DOI: 10.1037/0033-295X.102.1.4

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. Journal of Personality and Social Psychology, 108(4), 553–561. DOI: 10.1037/pspa0000016

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480. DOI: 10.1037/0022-3514.74.6.1464

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? Psychophysiology, 33(2), 175–183. DOI: 10.1111/j.1469-8986.1996.tb02121.x

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. Journal of Experimental Psychology: General, 143(3), 1369–1392. DOI: 10.1037/a0035028

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? Journal of Personality and Social Psychology, 65(4), 613–628. DOI: 10.1037/0022-3514.65.4.613

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. Journal of Experimental Social Psychology, 42(3), 259–272. DOI: 10.1016/j.jesp.2005.04.006

Hannikainen, I., & Cova, F. (In prep.). Trait reflectivity and consequentialist moral principles: Meta-analytic evidence.

Hannikainen, I. R., Machery, E., Rose, D., Stich, S., Olivola, C. Y., Sousa, P., … Zhu, J. (2019). For Whom Does Determinism Undermine Moral Responsibility? Surveying the Conditions for Free Will Across Cultures. Frontiers in Psychology, 10. DOI: 10.3389/fpsyg.2019.02428

Hannon, M. (2020). Empathetic Understanding and Deliberative Democracy. Philosophy and Phenomenological Research. DOI: 10.1111/phpr.12624

Harington, S. J. (1804). Nugæ Antiquæ: Being a Miscellaneous Collection of Original Papers, in Prose and Verse; Written During the Reigns of Henry VIII. Edward VI. Queen Mary, Elizabeth, and King James. Vernor and Hood.

Harrington, J. A., & Blankenship, V. (2002). Ruminative thoughts and their relation to depression and anxiety. Journal of Applied Social Psychology, 32(3), 465–485.

Harris, M. A., Brett, C. E., Johnson, W., & Deary, I. J. (2016). Personality Stability From Age 14 to Age 77 Years. Psychology and Aging, 31(8), 862–874. DOI: 10.1037/pag0000133

Haslanger, S. (2012). Resisting Reality: Social Construction and Social Critique. OUP USA.

Heider, J. D., & Skowronski, J. J. (2007). Improving the Predictive Validity of the Implicit Association Test. North American Journal of Psychology, 9(1), 53. Retrieved from questia.com/library/journal/1G1-164638530/improving-the-predictive-validity-of-the-implicit

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2013). The development of reasoning about beliefs: Fact, preference, and ideology. Journal of Experimental Social Psychology, 49(3), 559–565. DOI: 10.1016/j.jesp.2012.09.005

Helton, G. (2017, March 23). Personal Communication at 109th Annual Meeting of the Southern Society for Psychology and Philosophy.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–83. DOI: 10.1017/S0140525X0999152X

Hertzog, C., Smith, R. M., & Ariel, R. (2018). Does the Cognitive Reflection Test actually capture heuristic versus analytic reasoning styles in older adults? Experimental Aging Research, 44(1), 18–34. DOI: 10.1080/0361073X.2017.1398508

Holroyd, J., & Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In M. Brownstein & J. Saul (Eds.), Implicit Bias and Philosophy. Oxford University Press.

Holtzman, G. (2013). Do Personality Effects Mean Philosophy is Intrinsically Subjective? Journal of Consciousness Studies, 20(5–6), 5–6.

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. Decision. DOI: 10.1037/dec0000107

Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. Psychology & Marketing, 27(10), 938–963. DOI: 10.1002/mar.20367

Horgan, T., & Nichols, S. (2015). The zero point and I. In Pre-reflective Consciousness (pp. 155–187). Routledge.

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. Nature Climate Change, 1. DOI: 10.1038/s41558-018-0157-2

Huebner, B. (2016). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. In M. Brownstein & J. Saul (Eds.), Implicit Bias and Philosophy, Vol 1. Oxford University Press.

Huemer, M. (2006). Phenomenal conservatism and the internalist intuition. American Philosophical Quarterly, 147–158.

Huemer, M. (2007). Compassionate phenomenal conservatism. Philosophy and Phenomenological Research, 74(1), 30–55.

Huffman, S. (2018). In response to recent reports about the integrity of Reddit, I'd like to share our thinking. • r/announcements. Retrieved April 11, 2018, from Reddit website: reddit.com/r/announcements/comments/827zqc/in_response_to_recent_reports_about_the_integrity/

Hume, D. (1978). A Treatise of Human Nature (2nd edition; L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford; New York: Oxford University Press.

Hursthouse, R. (1999). On Virtue Ethics. Oxford University Press.

Hütter, M., & Sweldens, S. (2018). Dissociating Controllable and Uncontrollable Effects of Affective Stimuli on Attitudes and Consumption. Journal of Consumer Research, 45(2), 320–349. DOI: 10.1093/jcr/ucx124

Institute of Education Sciences. (2017). What Works Clearinghouse: Procedures Handbook (Version 4.0).

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. Journal of Memory and Language, 30(5), 513–541. DOI: 10.1016/0749-596X(91)90025-F

James, G. P. R. (1853). The Brigand, Or, Corse de Leon: A Romance. Simms and M'Intyre.

Janis, I. L., & Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. Journal of Experimental Psychology, 33(1), 73–77. DOI: 10.1037/h0060675

Jiménez, N., Rodriguez-Lara, I., Tyran, J.-R., & Wengstrom, E. (2017). Thinking fast, thinking badly. Economics Letters. Retrieved from eprints.mdx.ac.uk/22754/

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. Science, 310(5745), 116–119. DOI: 10.1126/science.1111709

Johnson-Laird, P., & Bara, B. G. (1984). Syllogistic inference. Cognition, 16(1), 1–61. DOI: 10.1016/0010-0277(84)90035-0

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. Cognition, 193, 103950. DOI: 10.1016/j.cognition.2019.04.019

Joshua, K. (2019). Philosophical Intuitions Are Surprisingly Robust Across Demographic Differences. Epistemology & Philosophy of Science, 56(2). Retrieved from

cyberleninka.ru/article/n/philosophical-intuitions-are-surprisingly-robust-across-demographic-differences

Jost, J. T. (2018). The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology. Current Directions in Psychological Science, 0963721418797309. DOI: 10.1177/0963721418797309

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior, 29, 39–69. DOI: 10.1016/j.riob.2009.10.001

Joy-Gaba, J. A., & Nosek, B. A. (2010). The Surprisingly Limited Malleability of Implicit Racial Evaluations. Social Psychology, 41(3), 137–146. DOI: 10.1027/1864-9335/a000020

Kahan, D. M. (2013). Ideology, Motivated Reasoning, and Cognitive Reflection: An Experimental Study. Judgment and Decision Making, 8, 407–424. Retrieved from sjdm.org/journal/13/13313/jdm13313.pdf

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science Curiosity and Political Information Processing. Political Psychology, 38, 179–199. DOI: 10.1111/pops.12396

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. Behavioural Public Policy, 54–86. DOI: 10.1017/bpp.2016.2

Kahneman, D. (2011). Thinking, Fast and Slow. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237–251. DOI: 10.1037/h0034747

Kahneman, D., & Tversky, A. (1974). Subjective Probability: A Judgment of Representativeness. In C.-A. S. S. V. Holstein (Ed.), The Concept of Probability in Psychological Experiments (pp. 25–48). Retrieved from link.springer.com/chapter/10.1007/978-94-010-2288-0_3

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. Journal of Personality and Social Psychology, 78(5), 871–888.

Kawakami, Kerry, Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. Journal of Personality and Social Psychology, 92(6), 957–971. DOI: 10.1037/0022-3514.92.6.957

Keller, H. (1903). The Story of My Life and Selected Letters. Retrieved from books.google.com/books/about/Helen_Keller.html?id=Y6-TDgAAQBAJ

Keller, J. (2005). In Genes We Trust: The Biological Component of Psychological Essentialism and Its Relationship to Mechanisms of Motivated Social Cognition. Journal of Personality and Social Psychology, 88(4), 686–702. DOI: 10.1037/0022-3514.88.4.686

Kennett, J., & Fine, C. (2009). Will the Real Moral Judgment Please Stand Up? Ethical Theory and Moral Practice, 12(1), 77–96. DOI: 10.1007/s10677-008-9136-4

Kim, M., & Yuan, Y. (2015). No Cross-Cultural Differences In The Gettier Car Case Cast Intuition: A Replication Study Of Weinberg et al. 2001. Episteme, 12(3), 355–361. DOI: 10.1017/epi.2015.17

Klein, G. A. (1998). Sources of power: How people make decisions. MIT press.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science, 1(4), 443–490. DOI: 10.1177/2515245918810225

Knobe, J., & Nichols, S. (2007). An Experimental Philosophy Manifesto. In J. Knobe & S. Nichols (Eds.), Experimental Philosophy (pp. 3–14). Oxford University Press.

Knutsen, J. A., & Presser, S. (2010). Question and Questionnaire Design. In Handbook of Survey Research (2nd ed., pp. 263–313). Emerald.

Koriat, A. (2019). Confidence judgments: The monitoring of object-level and same-level performance. Metacognition and Learning. DOI: 10.1007/s11409-019-09195-7

Kornblith, H. (1998). The role of intuition in philosophical inquiry: An account with no unnatural ingredients. In M. Depaul & W. Ramsey (Eds.), Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry (pp. 129–141). Oxford: Rowman & Littlefield.

Kornblith, H. (2012). On Reflection. OUP Oxford.

Kornblith, H. (2019a). Don’t Think Twice, It’s Alright. Philosophic Exchange, 48(1). Retrieved from digitalcommons.brockport.edu/phil_ex/vol48/iss1/1

Kornblith, H. (2019b). Second Thoughts and the Epistemological Enterprise. Cambridge University Press.

Korsgaard, C. M. (1996). The Sources of Normativity. Cambridge University Press.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. Nature Communications, 6, 7455. DOI: 10.1038/ncomms8455

Krizo, P. (2012). A summer high school computer game programming curriculum and an assessment of its effects on student motivation (Master's Thesis, California State University, Sacramento). Retrieved from hdl.handle.net/10211.9/1481

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. Journal of Experimental Psychology. General, 145(8), 1001–1016. DOI: 10.1037/xge0000179

Lee, S., Kawachi, I., Berkman, L. F., & Grodstein, F. (2003). Education, Other Socioeconomic Indicators, and Cognitive Function. American Journal of Epidemiology, 157(8), 712–720. DOI: 10.1093/aje/kwg042

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. Noûs, 49(4), 800–823. DOI: 10.1111/nous.12074

Levy, N. (2016). 'My Name is Joe and I'm an Alcoholic': Addiction, Self-Knowledge and the Dangers of Rationalism. Mind and Language, 31(3), 265–276.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. Journal of Behavioral Decision Making, 25(4), 361–381. DOI: 10.1002/bdm.752

Lieberman, D. A. (1979). Behaviorism and the mind: A (limited) call for a return to introspection. American Psychologist, 34(4), 319–333. DOI: 10.1037/0003-066X.34.4.319

Liquin, E., Metz, S. E., & Lombrozo, T. (2018). Explanation and its Limits: Mystery and the Need for Explanation in Science and Religion. Presented at the CogSci.

Livengood, J., Sytsma, J., Feltz, A., Scheines, R., & Machery, E. (2010). Philosophical temperament. Philosophical Psychology, 23(3), 313–330. DOI: 10.1080/09515089.2010.490941

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. Psychological Review, 91(3), 295–327. DOI: 10.1037/0033-295X.91.3.295

Lycan, W. G. (1975). Occam's Razor. Metaphilosophy, 6(3/4), 223–237. Retrieved from jstor.org/stable/24435155

Lynch, M. P. (2018). Arrogance, Truth, and Public Discourse. Episteme, 15(3), 283–296. DOI: 10.1017/epi.2018.23

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. Cognition, 92(3), B1–B12. DOI: 10.1016/j.cognition.2003.10.003

Madva, A. (2015). Why implicit attitudes are (probably) not beliefs. Synthese, 193(8), 2659–2684. DOI: 10.1007/s11229-015-0874-2

Madva, A. (2017). Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice. Ergo, an Open Access Journal of Philosophy, 4. DOI: 10.3998/ergo.12405314.0004.006

Maier, N. (1931). Reasoning in humans. The solution of a problem and its appearance in consciousness. Journal of Comparative Psychology, 12(2), 181–194. DOI: 10.1037/h0071361

Mallon, R. (2016). Intuitive Diversity and Disagreement. In J. Nado (Ed.), Advances in Experimental Philosophy and Philosophical Methodology (pp. 99–123). London: Bloomsbury Press.

Mandelbaum, E. (2013). Thinking is Believing. Inquiry, 57(1), 55–96. DOI: 10.1080/0020174X.2014.858417

Mandelbaum, E. (2017). Associationist Theories of Thought. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2017). Retrieved from plato.stanford.edu/archives/sum2017/entries/associationist-thought/

Mandelbaum Eric. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. Noûs, 50(3), 629–658. DOI: 10.1111/nous.12089

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. Memory & Cognition, 17(1), 11–17. DOI: 10.3758/BF03199552

Maundrell, H. (1703). A Journey from Aleppo to Jerusalem at Easter, AD 1697.

McCloskey, D. N., & Ziliak, S. T. (1996). The Standard Error of Regressions. Journal of Economic Literature, 34(1), 97–114. Retrieved from jstor.org/stable/2729411

McCoy, M. K. (2018, June 1). Researcher: Despite Good Intentions, Anti-Bias Training Can Actually Backfire. Wisconsin Public Radio. Retrieved from wpr.org/researcher-despite-good-intentions-anti-bias-training-can-actually-backfire

McNeish, D., & Wolf, M. G. (2019). Sum Scores Are Factor Scores. DOI: 10.31234/osf.io/3wy47

McPhetres, J. (2018). What does the cognitive reflection test really measure: A process dissociation investigation. DOI: 10.31219/osf.io/m43gn

Mele, A. (2008). Proximal intentions, intention-reports, and vetoing. Philosophical Psychology, 21(1), 1–14.

Mele, M. L., Federici, S., & Dennis, J. L. (2014). Believing Is Seeing: Fixation Duration Predicts Implicit Negative Attitudes. PLoS ONE, 9(8), e105106. DOI: 10.1371/journal.pone.0105106

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. Trends in Cognitive Sciences, 22(4), 280–293. DOI: 10.1016/j.tics.2018.02.001

Mercier, H., & Sperber, D. (2009). Intuitive and Reflective Inferences. In K. Frankish & J. S. B. T. Evans (Eds.), In Two Minds: Dual Processes and Beyond (pp. 149–170). Oxford University Press.

Mercier, H., & Sperber, D. (2017). The Enigma of Reason. Harvard University Press.

Metz, S. E., Weisberg, D. S., & Weisberg, M. (2018). Non-Scientific Criteria for Belief Sustain Counter-Scientific Beliefs. Cognitive Science, 42(5), 1477–1503. DOI: 10.1111/cogs.12584

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. Judgment and Decision Making, 13(3), 246–259. Retrieved from ideas.repec.org/a/jdm/journl/v13y2018i3p246-259.html

Meyer, D. (2018, May 29). Starbucks Is Closing Today For Its Company-Wide Unconscious Bias Training: Here's What You Need To Know. Fortune. Retrieved from fortune.com/2018/05/29/starbucks-closing-today-unconscious-bias-training/

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. Journal of Personality and Social Psychology, 65(3), 469. DOI: 10.1037/0022-3514.65.3.469

Montgomery, L. M. (1908). Anne of Green Gables. Retrieved from gutenberg.org/ebooks/45

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. Psychological Bulletin, 132(2), 297–326. DOI: 10.1037/0033-2909.132.2.297

Morsanyi, K., Byrne, R. M. J., & Byrne, R. M. J. (2019). Thinking, Reasoning, and Decision Making in Autism. DOI: 10.4324/9781351060912

Nagel, J. (2012). Intuitions and experiments: A defense of the case method in epistemology. Philosophy and Phenomenological Research, 85(3), 495–527.

Nagel, J. (2014). Intuition, Reflection, and the Command of Knowledge. Aristotelian Society Supplementary Volume, 88(1), 219–241. DOI: 10.1111/j.1467-8349.2014.00240.x

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. Midwest Studies in Philosophy, 31(1), 214–242.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. Journal of Experimental Psychology: General, 106(3), 226–254. Retrieved from insights.ovid.com/experimental-psychology-general/jepge/1977/09/000/semantic-priming-retrieval-lexical-memory/2/00004785

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Computer Science Department, Paper 2033.

Nisbett, R. E., & Bellows, N. (1977). Verbal reports about causal influences on social judgments: Private access versus public theories. Journal of Personality and Social Psychology, 35(9), 613–624. DOI: 10.1037/0022-3514.35.9.613

Nisbett, R. E., & Wilson, T. D. (1977a). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231–259. DOI: 10.1037/0033-295X.84.3.231

Nisbett, R. E., & Wilson, T. D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. Journal of Personality and Social Psychology, 35(4), 250–256. DOI: 10.1037/0022-3514.35.4.250

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. Perspectives on Psychological Science, 3(5), 400–424.

Norris, J. (1704). An essay towards the theory of the ideal or intelligible world. Design'd for two parts. Retrieved from books.google.com/books/about/An_essay_towards_the_theory_of_the_ideal.html?id=AgdQAAAAYAAJ

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. Social Cognition, 19(6), 625–666. DOI: 10.1521/soco.19.6.625.20886

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. Neuropsychologia, 91, 499–508. DOI: 10.1016/j.neuropsychologia.2016.09.010

Olson, M. A., & Fazio, R. H. (2006). Reducing Automatically Activated Racial Prejudice Through Implicit Evaluative Conditioning. Personality and Social Psychology Bulletin, 32(4), 421–433. DOI: 10.1177/0146167205284004

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716–aac4716. DOI: 10.1126/science.aac4716

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. Journal of Personality and Social Psychology, 108(4), 562–571. DOI: 10.1037/pspa0000023

Oyserman, D., & Dawson, A. (2019). Your fake news, our facts: Identity-based motivation shapes what we believe, share, and accept. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), The psychology of fake news: Accepting, sharing, and correcting misinformation. London, UK: Psychology Press.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. Journal of Personality and Social Psychology, 76(6), 972–987. DOI: 10.1037/0022-3514.76.6.972

Paley, W. (1785). The Principles of Moral and Political Philosophy. Retrieved from oll.libertyfund.org/titles/paley-the-principles-of-moral-and-political-philosophy

Paolini, S., Hewstone, M., Cairns, E., & Voci, A. (2004). Effects of direct and indirect cross-group friendships on judgments of Catholics and Protestants in Northern Ireland: The mediating role of an anxiety-reduction mechanism. Personality and Social Psychology Bulletin, 30(6), 770–786.

Patel, N. (2017). The Cognitive Reflection Test: A measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs (Thesis, University of Missouri--Columbia). Retrieved from mospace.umsystem.edu/xmlui/handle/10355/62365

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and Reasoning in Moral Judgment. Cognitive Science, 36(1), 163–177. DOI: 10.1111/j.1551-6709.2011.01210.x

Payne, B. K., & Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. European Review of Social Psychology, 20(1), 272–314.

Payne, K., Niemi, L., & Doris, J. (2018, March 27). How to Think about "Implicit Bias." Scientific American. Retrieved from scientificamerican.com/article/how-to-think-about-implicit-bias/

Peacocke, C. (2014). The Mirror of the World: Subjects, Consciousness, and Self-Consciousness. Oxford University Press.

Pennycook, G. (Ed.). (2018). The New Reflectionism in Cognitive Psychology: Why Reason Matters (1 edition). Abingdon, Oxon ; New York, NY: Routledge.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014a). Cognitive style and religiosity: The role of conflict detection. Memory & Cognition, 42(1), 1–10. DOI: 10.3758/s13421-013-0340-7

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014b). The role of analytic thinking in moral judgements and values. Thinking & Reasoning, 20(2), 188–214. DOI: 10.1080/13546783.2013.865000

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. Judgment and Decision Making, 10(6), 549–563. Retrieved from search.proquest.com/docview/1739211836/abstract

Pennycook, G., Cheyne, J. A., Koehler, D., & Fugelsang, J. A. (in prep.). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. DOI: 10.31234/osf.io/a7k96

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2013). Belief bias during reasoning among religious believers and skeptics. Psychonomic Bulletin & Review, 20(4), 806–811. DOI: 10.3758/s13423-013-0394-3

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? Behavior Research Methods, 1–8. DOI: 10.3758/s13428-015-0576-1

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. Cognition, 123(3), 335–346. DOI: 10.1016/j.cognition.2012.03.003

Pennycook, G., De Neys, W., Evans, J. St. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The Mythical Dual-Process Typology. Trends in Cognitive Sciences. DOI: 10.1016/j.tics.2018.04.008

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? Cognition, 124(1), 101–106. DOI: 10.1016/j.cognition.2012.04.004

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive Psychology, 80, 34–72. DOI: 10.1016/j.cogpsych.2015.05.001

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7), 2521–2526. DOI: 10.1073/pnas.1806781116

Pennycook, G., & Rand, D. G. (2019b). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of Personality. DOI: 10.1111/jopy.12476

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2016). Atheists and Agnostics Are More Reflective than Religious Believers: Four Empirical Studies and a Meta-Analysis. PLOS ONE, 11(4), e0153039. DOI: 10.1371/journal.pone.0153039

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. Psychonomic Bulletin & Review, 19(3), 528–534. DOI: 10.3758/s13423-012-0249-3

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2), 544–554. DOI: 10.1037/a0034887

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. British Journal of Social Psychology, 44(1), 29–45. DOI: 10.1348/014466604X23491

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. Handbook of Implicit Social Cognition: Measurement, Theory, and Applications, 10, 255–278.

Peters, U. (2019). Implicit bias, ideological bias, and epistemic risks in philosophy. Mind & Language, 34(3), 393–419. DOI: 10.1111/mila.12194

Petitmengin, C., Remillieux, A., Cahour, B., & Carter-Thomas, S. (2013). A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. Consciousness and Cognition, 22(2), 654–669. DOI: 10.1016/j.concog.2013.02.004

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology, 90(5), 751. DOI: 10.1037/0022-3514.90.5.751

Petty, R. E., Fazio, R. H., & Briñol, P. (2009). The new implicit measures: An overview. Attitudes: Insights from the New Implicit Measures, 3–18.

Pew Research Center. (2017, October 5). Views on global warming and environmental regulation, personal environmentalism. Retrieved October 20, 2019, from people-press.org/2017/10/05/7-global-warming-and-environmental-regulation-personal-environmentalism/

Pinillos, N. Á., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. Mind & Language, 26(1), 115–139.

Plantinga, A. (1967). God and Other Minds: A Study of the Rational Justification of Belief in God. Cornell University Press.

Posner, M. I., Snyder, C. R., & Solso, R. (1975). Attention and Cognitive Control. In Information Processing and Cognition: Loyola Symposium. Wiley Online Library.

Price-Blackshear, M. A., Sheldon, K. M., Corcoran, M. J., & Bettencourt, B. A. (2019). Individuating information influences partisan judgments. Journal of Applied Social Psychology. DOI: 10.1111/jasp.12595

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). Journal of Behavioral Decision Making, 29(5), 453–469. DOI: 10.1002/bdm.1883

Quilty-Dunn, J., & Mandelbaum, E. (2017). Against dispositionalism: Belief in cognitive science. Philosophical Studies, 1–20. DOI: 10.1007/s11098-017-0962-x

Quine, W. V. (1951). Two Dogmas of Empiricism. The Philosophical Review, 60(1), 20–43. DOI: 10.2307/2181906

Rand, D. G. (2019). Personal Correspondence.

Rawls, J. (1971). A Theory of Justice (Revised edition). Cambridge, Mass: Belknap Press.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How Numeracy Influences Risk Comprehension and Medical Decision Making. Psychological Bulletin, 135(6), 943–973. (American Psychological Association. Journals Department, 750 First Street NE, Washington, DC 20002-4242. Tel: 800-374-2721; Tel: 202-336-5510; Fax: 202-336-5502; e-mail: order@apa.org; Web site: apa.org/publications).

Reynolds, C. J., Byrd, N., & Conway, P. (forthcoming). Trait reflectiveness and responses to moral dilemmas: A meta-analytic process dissociation approach.

Richeson, J. A., & Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. Journal of Experimental Social Psychology, 40(3), 417–423. DOI: 10.1016/j.jesp.2003.09.002

Robbins, E., Shepard, J., & Rochat, P. (2017). Variations in judgments of intentional action and moral evaluation across eight cultures. Cognition, 164, 22–30. DOI: 10.1016/j.cognition.2017.02.012

Ronayne, D., Sgroi, D., & Tuckwell, A. (2020). Evaluating the Sunk Cost Effect (No. 1269; The Warwick Economics Research Paper Series (TWERPS)). University of Warwick, Department of Economics. ideas.repec.org/p/wrk/warwec/1269.html

Rosenfeld, H. M., & Baer, D. M. (1969). Unnoticed verbal conditioning of an aware experimenter by a more aware subject: The double-agent effect. Psychological Review, 76(4), 425–432. DOI: 10.1037/h0027451

Ross, R. M., Pennycook, G., McKay, R., Gervais, W. M., Langdon, R., & Coltheart, M. (2016). Analytic cognitive style, not delusional ideation, predicts data gathering in a large beads

task study. Cognitive Neuropsychiatry, 0(0), 1–15. DOI: 10.1080/13546805.2016.1192025

Rouder, J., Kumar, A., & Haaf, J. M. (2018). Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. Manuscript under Review. DOI: 10.31234/osf.io/3cjr5

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. Memory & Cognition, 17(6), 759–769. DOI: 10.3758/BF03202637

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. Journal of Personality and Social Psychology, 91(6), 995–1008. DOI: 10.1037/0022-3514.91.6.995

Saroglou, V. (2002). Religion and the five factors of personality: A meta-analytic review. Personality and Individual Differences, 32(1), 15–25. DOI: 10.1016/S0191-8869(00)00233-6

Saul, J. (2013a). Implicit bias, stereotype threat and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), Women in philosophy: What needs to change (pp. 39–60). Oxford University Press.

Saul, J. (2013b). Scepticism and Implicit Bias. Disputatio, 5(37), 243–263.

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. Nature Human Behaviour, 1–3. DOI: 10.1038/s41562-019-0711-6

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining Power and Sample Size for Simple and Complex Mediation Models. Social Psychological and Personality Science, 8(4), 379–386. DOI: 10.1177/1948550617715068

Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. Consciousness and Cognition, 20(4), 1722–1731. DOI: 10.1016/j.concog.2011.04.007

Schwenkler, J. (2018). Self-Knowledge and Its Limits. Journal of Moral Philosophy, 15(1), 85–95. DOI: 10.1163/17455243-01501005

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. Noûs, 36(2), 249–275. DOI: 10.1111/1468-0068.00370

Schwitzgebel, E. (2010). Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief. Pacific Philosophical Quarterly, 91(4), 531–553.

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. Journal of Personality and Social Psychology, 80(4), 645–654. DOI: 10.1037/0022-3514.80.4.645

Seyedsayamdost, H. (2015). On Normativity and Epistemic Intuitions: Failure of Replication. Episteme, 12(1), 95–116. DOI: 10.1017/epi.2014.27

Shea, N. (2013). Naturalising Representational Content. Philosophy Compass, 8(5), 496–509. DOI: 10.1111/phc3.12033

Shea, N. (2018). Representation in Cognitive Science. Oxford University Press.

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: The case for 'Type Zero' cognition. Neuroscience of Consciousness, 2016(1). DOI: 10.1093/nc/niw005

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. Journal of Experimental Psychology: General, 141(3), 423.

Shtulman, A., & Mccallum, K. (2014). Cognitive reflection predicts science understanding. Proceedings of the 36th Annual Conference of the Cognitive Science Society, 2937–2942.

Sidgwick, H. (1874). The Methods Of Ethics (7th ed.). Chicago: Hackett Publishing.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after P-Hacking. Meeting of the Society for Personality and Social Psychology, 38. Retrieved from papers.ssrn.com/abstract=2205186

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-Positive Citations. Perspectives on Psychological Science, 13(2), 255–259. DOI: 10.1177/1745691617698146

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test. Behavior Research Methods, 1–12. DOI: 10.3758/s13428-018-1029-4

Sirota, M., Kostovičová, L., Juanchich, M., Dewberry, C., & Marshall, A. C. (2018). Measuring Cognitive Reflection without Maths: Developing and Validating the Verbal Cognitive Reflection Test. DOI: 10.31234/osf.io/pfe79

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3–22. DOI: 10.1037/0033-2909.119.1.3

Smith, A. (2018). Implicit Bias, Moral Agency, and Moral Responsibility. In G. Rosen, A. Byrne, J. Cohen, S. V. Harman, Elizabeth, & S. V. Shiffrin (Eds.), The Norton Introduction to Philosophy (2nd ed., pp. 772–781). New York: W. W. Norton & Company.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. Psychological Review, 85(4), 355–362. DOI: 10.1037/0033-295X.85.4.355

Sosa, E. (1991). Knowledge in Perspective: Selected Essays in Epistemology. Cambridge University Press.

Sperber, D. (1997). Intuitive and Reflective Beliefs. Mind & Language, 12(1), 67–83. DOI: 10.1111/j.1468-0017.1997.tb00062.x

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. Judgment and Decision Making, 13(3), 260–267. Retrieved from ideas.repec.org/a/jdm/journl/v13y2018i3p260-267.html

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), In Two Minds: Dual Processes and Beyond (pp. 55–88).

Stich, S. P. (1990). The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation. Cambridge, MA, US: The MIT Press.

Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: Familiarity. PeerJ, 4, e2395. DOI: 10.7717/peerj.2395

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. Personality and Social Psychology Review, 8(3), 220–247. DOI: 10.1207/s15327957pspr0803_1

Strack, F., & Deutsch, R. (2014). The reflective-impulsive model. Dual-Process Theories of the Social Mind, 92–104.

Strack, F., Werth', L., & Deutsch, R. (2006). Reflective and Impulsive Determinants of Consumer Behavior. Journal of Consumer Psychology, 16(3), 205–216. DOI: 10.1207/s15327663jcp1603_2

Strohminger, N. (2018). Identity Is Essentially Moral. In K. Gray & J. Graham (Eds.), Atlas of Moral Psychology (p. 141). Guilford Publications.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643–662. DOI: 10.1037/h0054651

Stupple, E. J. N., Pitchford, M., Ball, L. J., Hunt, T. E., & Steel, R. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. PLOS ONE, 12(11), e0186404. DOI: 10.1371/journal.pone.0186404

Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. Consciousness and Cognition, 33, 548–560. DOI: 10.1016/j.concog.2014.10.006

Sun, R. (2016). Implicit and Explicit Processes: Their Relation, Interaction, and Competition. In L. Macchi, M. Bagassi, & R. Viale (Eds.), Cognitive Unconscious and Human Rationality (pp. 257–274). Cambridge, MA: The MIT Press.

Sytsma, J., & Ozdemir, E. (2019). No Problem: Evidence that the Concept of Phenomenal Consciousness is Not Widespread. Journal of Consciousness Studies, 26(9–10), 241–256.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. Thinking & Reasoning, 23(3), 207–234. DOI: 10.1080/13546783.2017.1292954

Talaifar, S., & Swann, W. B. (n.d.). Deep Alignment with Country Shrinks the Moral Gap Between Conservatives and Liberals. Political Psychology, 0(0). DOI: 10.1111/pops.12534

Tallant, J. (2013). Intuitions in physics. Synthese, 190(15), 2959–2980. DOI: 10.1007/s11229-012-0113-z

Taylor, C. (1976). Responsibility for Self. In A. O. Rorty (Ed.), The Identities of Persons (pp. 281–99). University of California Press.

Thompson, T. P. (1832). Exercises, Political and Others. Retrieved from books.google.com/books/about/Exercises_Political_and_Others.html?id=AGsvAQAAMAAJ

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. Cognitive Psychology, 63(3), 107–140. DOI: 10.1016/j.cogpsych.2011.06.001

Thomson, J. J. (1986). Killing, letting die, and the trolley problem. In W. Parent (Ed.), Rights, Restitution, and Risk: Essays in Moral Theory. Harvard University Press.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making, 11(1), 99–113. Retrieved from psycnet.apa.org/record/2016-09222-009

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. Behavior Research Methods, 43(3), 692–700. DOI: 10.3758/s13428-011-0076-x

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & Cognition, 39(7), 1275. DOI: 10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning, 20(2), 147–168. DOI: 10.1080/13546783.2013.844729

Toribio, J. (2018a). Accessibility, implicit bias, and epistemic justification. Synthese, 1–19. DOI: 10.1007/s11229-018-1795-7

Toribio, J. (2018b). Implicit Bias: From Social Structure to Representational Format. Theoria: An International Journal for Theory, History and Foundations of Science, 33(1), 41–60. Retrieved from jstor.org/stable/26355568

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2014). The grim reasoner: Analytical reasoning under mortality salience. Thinking & Reasoning, 20(3), 333–351. DOI: 10.1080/13546783.2013.823888

Tyler, J. M., & McCullough, J. D. (2009). Violating Prescriptive Stereotypes on Job Resumes: A Self-Presentational Perspective. Management Communication Quarterly, 23(2), 272–287. DOI: 10.1177/0893318909341412

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief. Trends in Cognitive Sciences, 22(3), 213–224. DOI: 10.1016/j.tics.2018.01.004

Van Dessel, P., De Houwer, J., Roets, A., & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. Journal of Personality and Social Psychology, 110(1), e1–e15. DOI: 10.1037/pspa0000039

Velleman, J. D. (1989). Practical Reflection. Princeton University Press.

Velleman, J. D. (2000). The Possibility of Practical Reason. Clarendon Press.

Wallace, R. J. (2006). Normativity and the Will: Selected Papers on Moral Psychology and Practical Reason. Clarendon Press.

Wallace, R. J. (2018). Practical Reason. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2018). Retrieved from plato.stanford.edu/archives/spr2018/entries/practical-reason/

Watson, J. B. (1920). Is Thinking Merely Action of Language Mechanisms1? (V.). British Journal of Psychology. General Section, 11(1), 87–104. DOI: 10.1111/j.2044-8295.1920.tb00010.x

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? Philosophical Psychology, 23(3), 331–355. DOI: 10.1080/09515089.2010.490944

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. Philosophical Topics, 29(1–2), 429–460. DOI: 10.5840/philtopics2001291/217

Welsh, M. B., & Begg, S. H. (2017, July). The Cognitive Reflection Test: Familiarity and predictive power in professionals. Presented at the Annual Meeting of the Cognitive Science Society, London. Retrieved from mindmodeling.org/cogsci2017/papers/0659/paper0659.pdf

Wenar, L. (2008). Property rights and the resource curse. Philosophy & Public Affairs, 36(1), 2–32.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bem. Psychological Review, 87(1), 105–112. DOI: 10.1037/0033-295X.87.1.105

Wilson, T., & Nisbett, R. E. (1978). The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior. Social Psychology, 41(2), 118–131. DOI: 10.2307/3033572

Winskel, H., & Bhatt, D. (2019). The role of culture and language in moral decision-making. Culture and Brain. DOI: 10.1007/s40167-019-00085-y

Woodward, J. (2016). Causation and Manipulability. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2016). Retrieved from plato.stanford.edu/archives/win2016/entries/causation-mani/

Yaden, D. B. (2019). Psychology of Philosophy. Under review.

Yilmaz, O., Adil Saribay, S., & Iyer, R. (2019). Are neo-liberals more intuitive? Undetected libertarians confound the relation between analytic cognitive style and economic conservatism. Current Psychology. DOI: 10.1007/s12144-019-0130-x

Yilmaz, O., & Alper, S. (2019). The link between intuitive thinking and social conservatism is stronger in WEIRD societies. Judgment and Decision Making, 14(2). Retrieved from journal.sjdm.org/18/181212/jdm181212.html

Yılmaz, O., & Sarıbay, S. A. (2016). An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension. Judgment and Decision Making, 11(3), 287–300. Retrieved from journal.sjdm.org/vol11.3.html

Zuckerman, M., Li, C., Lin, S., & Hall, J. A. (2019). The Negative Intelligence–Religiosity Relation: New and Confirming Evidence. Personality and Social Psychology Bulletin, 0146167219879122. DOI: 10.1177/0146167219879122

BIOGRAPHICAL SKETCH

Nick Byrd is a philosopher-scientist who has studied—among other things—reasoning, wellbeing, willpower, and technology. For the latest information, including free preprints and audiopapers of Byrd's articles and chapters, see byrdnick.com/cv.