ON SECOND THOUGHT: REFLECTIONS ON THE REFLECTION DEFENSE

Markus Kneer, David Colaço, Joshua Alexander & Edouard Machery

Forthcoming in T. Lombrozo, J. Knobe & S. Nichols, *Oxford Studies of Experimental Philosophy*.

Please cite the final version.

Abstract

This article sheds light on a response to experimental philosophy that has not yet received enough attention: *the reflection defense*. According to proponents of this defense, judgments about philosophical cases are relevant only when they are the product of careful, nuanced, and conceptually rigorous reflection. We argue that the reflection defense is misguided: We present five studies (N>1800) showing that people make the same judgments when they are primed to engage in careful reflection as they do in the conditions standardly used by experimental philosophers.

1. The Restrictionist Challenge

This much should be uncontroversial: the method of cases plays an important role in contemporary philosophy. While there is disagreement about how best to interpret this method (Williamson, 2007; Malgrem, 2011; Cappelen, 2012; Alexander, 2012; Deutsch, 2015; Nado, 2016; Colaço and Machery, 2017; Machery, 2017; Strevens, 2019), there is little doubt that philosophers often proceed by considering actual or hypothetical situations, and use intuitions about such situations to assess philosophical theories. Despite its central role in philosophical practice, the method of cases has recently come under pressure: A series of experimental studies suggests that

judgments regarding classic philosophical thought experiments (aka "cases") are sensitive to factors such as culture, gender, affect, framing and presentation order, factors, that is, that are not standardly thought to be of philosophical relevance (for review and discussion, see Alexander, 2012; Machery, 2017; Stich and Machery, forthcoming).

Critics of experimental philosophy have responded to this challenge in various ways (for discussion, see, e.g., Alexander, 2012, 2016; Alexander and Weinberg 2007; Weinberg, Gonnerman, Buckner, and Alexander 2010; Cappelen, 2012; Machery 2011, 2012, 2017; Schwitzgebel and Cushman 2012, 2015; Deutsch, 2015; Mizrahi, 2015). Our goal in this article is to shed light on one response that has not yet received enough attention: *the reflection defense* (for previous discussion, see Weinberg, Alexander, Gonnerman, and Reuter, 2012). The reflection defense targets features of the deliberative process invoked in experimental studies of ordinary judgments about philosophical cases: According to proponents of this defense, judgments about philosophical cases are relevant only when they are the product of careful, nuanced, and conceptually rigorous reflection, while, they hold, the judgments elicited in experimental studies are swift shots from the hip that lack the necessary deliberative care; as such, they are easily distorted by irrelevant factors. Proponents of the reflection defense conclude that, since these kinds of judgments are unfit to serve as input for responsible philosophical inquiry, experimental studies that reveal their vagaries can be safely ignored.

We suspect that the reflection defense is misguided, and this article is an attempt to defend this suspicion. The reflection defense assumes that reflection (i) *influences* how people think about philosophical cases and (ii) brings their judgments more into *alignment* with philosophical orthodoxy (where it exists). We call this the

"Influence & Alignment Assumption". To illustrate the point, take Gettier cases, invoked, for instance, in Kauppinnen's exposition of the reflection defense: The assumption is that increased reflection not only occasions a *different* rate of knowledge ascriptions in Gettier cases than the standardly high rates of folk ascriptions, but – in line with textbook epistemology – a *lower* rate of knowledge ascriptions. The idea is thus that increased reflection influences and – from the point of view of philosophical orthodoxy – *improves* the responses to the cases at hand.

In order to examine the Influence and Alignment Assumption, we present studies that explore how the folk think about four philosophical cases, or pairs of cases, that have generated a great deal of attention from both traditional and experimental philosophers across various areas of philosophy: cases used to challenge the idea that knowledge is justified true belief, the idea that reference is fixed by description, the idea that knowledge depends only on epistemic considerations, and the idea that knowledge entails belief. For each of these cases, we attempted to manipulate reflective care using four common tools from social psychology and behavioral economics: a standard delay manipulation, a standard incentivization procedure, a standard manipulation for increased accountability, and a standard prime for analytic thinking. We also examined whether people who are primed to give more reflective responses actually respond differently to philosophical cases than people who are not so primed. Finally, we explored correlations between how people responded to the cases at hand and individual differences in preference for slow, careful deliberation using the Rational-Experiential Inventory (Epstein, Pacini, Denes-Raj, and Heier, 1996). Nothing mattered. People seem to make the same judgments when they are primed to engage in careful reflection as they do in the conditions standardly used by experimental philosophers. The reflection defense thus

seems unwarranted to presume that reflection relevantly changes how people think about philosophical cases.

We proceed as follows: In Section 2 we discuss the reflection defense in more detail, and describe our strategy for addressing it in Section 3. After setting the stage for our empirical research, we present five experimental studies and their results in Sections 4 through 8, and conclude in Section 9 by explaining what we take these results to mean for the reflection defense.

2. The Reflection Defense

Let's begin with a few particularly clear examples of the reflection defense, starting with Ludwig's influential formulation (2007, 149):

> We should not expect that in every case in which we are called on to make a judgment we are at the outset equipped to make correct judgments without much reflection. Our concepts generally have places in a family of related concepts, and these families of concepts will have places in larger families of concepts. How to think correctly about some cases we are presented can be a matter that requires considerable reflection. When a concept, like that of justification, is interconnected without our thinking in a wide variety of domains, it becomes an extremely complex matter to map out the conceptual connections and at the same time sidestep all the confusing factors.

Kauppinen largely concurs (2007, 97):

> When philosophers claim that according to our intuitions, Gettier cases are not knowledge, they are not presenting a hypothesis about gut reactions to counterfactual scenarios but, more narrowly, staking a claim of how competent and careful users of the ordinary concept of knowledge would pre-

> theoretically classify the case in suitable conditions. The claim, then, is not about what I will call *surface intuitions* but about *robust intuitions*.

Liao presents "the argument from robust intuitions" (without embracing it) as follows (2008, 256):

> [S]ome might think that one should distinguish between surface intuitions, which are "first-off" intuitions that may be little better than mere guesses; and robust intuitions, which are intuitions that a competent speaker might have under sufficiently ideal conditions such as when they are not biased.

Horvath presents the reflection defense (without embracing it) as follows (2010, 453):

> [T]he existing studies only aim at spontaneous responses to hypothetical cases (…). The opposing claim (…) is that what we actually rely on in philosophy are reflective intuitions, which are, it is suggested, of a much better epistemic quality than the typically spontaneous—and unreflective—intuitive responses of the folk (…). But if the "intuitions"' (…) really have to be understood as "reflective intuitions," then the available experimental studies do not contribute much to its support, or so the objection goes.

Finally, Nado (2015) also discusses the reflection defense, connecting it to the place of expertise in philosophical methodology (see also Swain, Alexander, and Weinberg 2008, section 3; Bengson, 2013; Gerken and Beebe, 2015).

The basic idea contained in these passages is rather straightforward: philosophers who use the method of cases are only interested in judgments generated by careful reflection about the cases themselves and the concepts we deploy in response to these cases, and whatever it is that experimental philosophers have been studying, they have not been studying those kinds of things. Thus, experimental studies revealing that unreflective judgments are susceptible to a host of irrelevant

factors do nothing to disqualify reflective judgments from playing a role in philosophical argumentation.

In more detail, the reflection defense begins with a necessary condition for the philosophical relevance of judgments about thought experiments: These judgments are philosophically relevant only when they result from careful reflection (Premise 1). It then makes a claim about experimental-philosophy studies: These studies do not examine judgments that result from careful reflection (Premise 2). It concludes that experimental philosophy findings are not philosophically relevant.

The two premises of the reflection defense call for clarification. First, and least important, Premise 2 can be formulated in several different ways. The weakest formulation would merely assert that experimental philosophers have not clearly demonstrated that their studies examine the right kind of judgment; for all experimental philosophers have shown, their studies could bear on the vagaries of unreflective judgments. A stronger formulation would assert that extant experimental philosophy studies fail to examine the right kind of judgment, while leaving open the possibility that improved studies would get at the right kind of judgment. The strongest formulation would assert that experimental-philosophy studies are necessarily unable to examine the judgments that result from careful reflection, perhaps because of what careful reflection involves. Kauppinen comes close to embracing the strongest reading, asserting that experimental-philosophy studies are necessarily unable to elicit reflective judgments (2007, 106):

> Testing for ideal conditions and careful consideration does not seem to be possible without engaging in dialogue with the test subjects, and that, again, violates the spirit and letter of experimentalist quasi-observation. (…) We can imagine a researcher going through a test subject's answers together with her,

> asking for the reasons why she answered one way rather than another (…). But this is no longer merely 'probing' the test subjects. It is not doing experimental philosophy in the new and distinct sense, but rather a return to the good old Socratic method.

The content and plausibility of Premises 1 and 2 also depend on how the distinction between robust and surface judgments, or, as we will say in the remainder of this article, reflective and unreflective judgments, is characterized. It is useful to tease apart thin and thick characterizations of this distinction.[1] One end of this continuum is anchored by what we will call the "thin characterization of reflective judgment." A judgment is thinly reflective just in case it results from a deliberation process involving attention, focus, cognitive effort, and so on—the type of domain-general psychological resources that careful and attentive thinking requires—and unreflective otherwise. We suspect that the thin characterization of reflection is similar to both lay and psychological conceptions of reflection (e.g., Paxton, Ungar, and Greene, 2011). Horvath (2010) and Nado (2015) also seem to understand reflection thinly.

Thicker conceptions of reflective judgments add requirements to the thin conception of reflective judgments. For Kauppinen, for example, reflective judgments are the products of the kind of dialogical activity central to the Socratic method (2007, 109):

[1] Weinberg and Alexander (2014) provide an overview of the different conceptions of "intuition" used in current metaphilosophical debates. Both the thin and thicker characterizations of reflective judgment discussed below count as thick conceptions on their way of carving up the landscape.

> [T]here is no way for a philosopher to ascertain how people would respond in such a situation without (…) entering into dialogue with them, varying examples, teasing out implications, presenting alternative interpretations to choose from to separate the semantic and the pragmatic, and so on. I will call this approach the Dialogue Model of the epistemology of folk concepts.

Ludwig is also interested only in thick reflective judgments, and on his view reflective judgments are based solely on conceptual competence (2007, 135):

> Conducting and being the subject of a thought experiment is a reflective exercise. It requires that both the experimenter and the subject understand what its point is. As it is a reflective exercise, it also presupposes that the subject of the thought experiment is able to distinguish between judgments solely based on competence (or recognition of the limits of competence) in deploying concepts in response to the described scenario.

Ludwig clarifies what he means by "conceptual competence" in a footnote (2007, 136):

> Failing to draw a distinction between unreflective judgments based on empirical beliefs and judgments based solely on competence in the deployment of concepts not uncommonly leads to a failure to appreciate the special epistemic status of the latter, the special role that first person investigation of them plays in the acquisition of a priori knowledge, and the stability of the judgments which are reached on this basis.

That is, on his view, reflective judgments are epistemically analytic (that is, entertaining the propositions they express can be sufficient for their justification).

These are just some ways that we might think about reflective judgment (for another proposal, see Hannon, 2018); what's important for our purposes is just that

the reflection defense will take different forms depending on which characterization of reflection is involved. It is beyond the scope of a single article to address all the possible variants in depth, and so we will focus here only on a version of the reflection defense that appeals to the thin conception of reflection presented above. While this means that we will be leaving thicker versions of the defense to the side for now, we think that there are good reasons for focusing on a thin version of the reflection defense. First, this version is the most easily tractable by means of experimental tools—the tools we intend to deploy in what follows. There is a wealth of tools in psychology and behavioral economics to single out judgments that result from reflective deliberation thinly understood, and we can use these to assess the reflection defense. Second, versions of the reflection defense that appeal to thick characterizations of reflection face problems of their own.[2] First, thicker versions of the reflection defense face what we will call a "descriptive-inadequacy" problem: the thicker the notion of reflection appealed to, the less likely it is that philosophers' judgments in usual philosophical debates result from a reflective deliberation so understood. To illustrate, consider Ludwig's claim that answers to philosophical cases must be "judgments based solely on competence." Although we will not argue for this claim here, we doubt that the judgments elicited by cases in philosophy are typically

[2] Weinberg and Alexander (2014) also propose a set of conditions that must be met by anyone attempting to argue that experimental philosophers simply have not been studying the right kind of judgments or "intuitions." Among those conditions is one that they call the "current practice condition"; failure to meet their current practice condition is very similar to what we will call the "descriptive-inadequacy" problem below.

of this kind; many of them do not seem to express analytic propositions at all (Williamson, 2007; Cappelen, 2012; Machery, 2017). Second, thicker versions of the reflection defense face what we will call a "stipulation" problem: Characterizations of reflective judgments should not make it the case *by stipulation* that experimental philosophers' findings happen to bear only on unreflective judgments. Stipulative victories are no victories at all, and it should be an empirical question whether reflective judgments suffer from the vagaries evidenced by fifteen years of experimental philosophy. To illustrate, when Ludwig proposes that reflective judgments are solely based on conceptual competence, he makes it the case by sheer stipulation that a large part of experimental philosophy, which examines the influence of pragmatic factors on judgments about thought experiments, happens to be studying unreflective judgments. A more satisfying strategy, we propose, would specify "reflection" so as to allow for the empirical study of whether reflective judgments are immune to the influence of pragmatic considerations.

3. Addressing the Reflection Defense

Our goal in this article is to assess a presupposition of the reflection defense: the Influence and Alignment Assumption, that is, the idea that, when people consider a thought experiment reflectively, they would tend to judge differently than in the conditions standardly employed by experimental philosophers, and their responses would be more in line with what philosophical orthodoxy considers correct. For instance, while many people may judge that, in a fake barn case, the character *knows* that she is seeing a real barn under standard experimental-philosophy conditions (Colaço, Buckwalter, Stich, and Machery, 2014), they would come to the opposite conclusion if they considered the fake-barn case reflectively, or so proponents of the

reflection defense assume.

To determine whether the judgments made in response to a case result from the process of careful reflection (thinly understood), we looked at two distinct types of properties: *dispositional* qualities pertaining to the subject responding to thought experiments and *circumstances* pertaining to the process of deliberation itself; careful reflection can either be fostered by an inherent inclination to engage in careful analytic thinking or else by appropriate conditions of deliberation. So, our strategy was to examine whether the judgments of people disposed to make reflective judgments or the judgments of people primed to engage in deliberation differ from the judgments made under conditions standardly employed by experimental philosophers.

One way to measure people's disposition to reflection is the *Need for Cognition* (NFC) test (Cacioppo and Petty, 1982; Cacioppo, Petty, Kao, and Rodriguez, 1986). Some individuals are naturally drawn to complex analytic-thinking tasks and might thus manifest the necessary care and reflection required for reflective judgments. An alternative measure that targets much the same dispositional quality is the *Cognitive Reflection Task* (Frederick, 2005; Toplak, West, and Stanovich, 2011) or CRT for short. Previous empirical studies using the NFC and CRT (Weinberg et al. 2012; Gerken and Beebe, 2014) found little support for the reflection defense; neither high NFC nor high CRT scores correlated with decreased sensitivity to distortive factors such as contextual priming, print font, or presentation order.[3]

[3] Pinillos Smith, Nair, Marchetto, and Mun (2011) use the CRT as a proxy to measure "general intelligence," and find that "those who display higher general intelligence are less likely to exhibit the Knobe Effect" (124). However, as long as one does not

In our experiments, we employed a third standard psychological questionnaire to measure people's disposition to reflection, namely the *Rational-Experiential Inventory* or REI (Epstein et al., 1996; Pacini and Epstein, 1999). Subjects on the "rational" end of the spectrum typically manifest an increased "ability to think logically and analytically" (1999, 974); those on the "experiential" end of the spectrum manifest a stronger "reliance on and enjoyment of feelings and intuitions in making decisions" (1999, 974). Differently put, "rational" subjects are more prone to analytic cognition, "experiential" subjects to more intuition-driven, cognitively less effortful cognition.

Consistent with the studies cited above, we proposed to operationalize the distinction between reflective and unreflective judgments in terms of the rational/experiential distinction developed by Epstein and colleagues. If people who are reflective as measured by yet a third standard psychological measure (in addition to the Need for Cognition scale and the Cognitive Reflection Test already used by Gonnerman, Reuter, and Weinberg (2011); Weinberg et al. (2012); and Gerken and Beebe (2014)) do not differ from people who are unreflective, then this would be evidence that reflection does not change the judgments people make in response to cases.

Naturally, it could be that the Rational-Experiential Inventory fails to really measure people's tendency to engage in reflective deliberation, even thinly understood, or that these people fail to act on their tendency in our studies. To address

---

defend a *bias* account of the Knobe Effect, according to which people's judgments of intentionality are systematically distorted by outcome valence, the findings of Pinillos and colleagues do not constitute evidence for the reflection defense.

these concerns, we needed to look at other ways of determining whether people are reporting reflective judgments. A second way to distinguish reflective from unreflective judgments draws on the circumstances that lead people to engage in reflection when making judgments about philosophical cases. Kauppinen (2007, 104) highlights the importance of such circumstances: Reflective judgment "can take hard thinking and time, and the attempt could be thwarted by passions or loss of interest", while "there is a general requirement to think through the implications of individual judgements—a hasty judgement… will not count as one's robust intuition about the case".[4] We attempted to encourage careful reflective processes by means of four standard experimental manipulations familiar from social psychology and experimental economics: forced delay, financial incentive, response justification via provision of reasons, and priming of analytic cognition.

In the *forced-delay* condition, participants were encouraged to read the vignette slowly, carefully, and to think about possible variations of the scenario. They could only proceed to a screen registering their answer after a certain delay, which varied from 40 to 60 seconds depending on the word count of the vignettes. Delay manipulations are frequently used in social-psychological research; the speed-accuracy trade-off is one of the most well-studied and pervasive effects in human judgment, perception, and decision making. Slower responses tend to correlate positively with improved accuracy and are less susceptible to biases or other

[4] Discussion of the circumstances that lead to reflection are also present in Dewey's five steps of reflective thought: He notes that one must take time to deliberate on a case, rather than prematurely accepting the conclusion at which one arrives (1910, 73-74).

distorting factors (Garrett, 1922; Hick, 1952; Ollman, 1966; Schouten and Bekker, 1967; Pachella, 1973; Wickelgren, 1977; Ratcliff and Rouder, 1998; Forstmann et al., 2008). Forced delay has been applied in various kinds of experiments. Rand, Greene, and Nowak (2012), for example, compare people's level of altruistic behavior in a one-shot public good games with and without a time delay, stating that in the former condition "decisions are expected to be driven more by reflection" (2012, 428). Rand and colleagues find that people become less altruistic in the latter condition, and conclude that "intuition supports cooperation in social dilemmas, and that reflection can undermine these cooperative impulses" (2012, 427). In their fourth experiment, Pizarro, Uhlmann, and Bloom (2003, 657) do not impose a delay on participants' answers, but they asked participants in the rational-instructions condition to "make these judgments from (…) a deliberative perspective (i.e., "my most rational, objective judgment is that…"), which is similar to the instructions we used. Pizarro and colleagues found that the moral assessment of causally deviant acts changes when people are asked to judge from this "rational, objective" perspective.

In the *financial-incentive* condition, participants were promised double compensation in case they got the answer "right," which was intended to encourage careful reflection. All participants in this condition received extra compensation independently of the answer chosen. Hertwig and Ortmann (2001) survey studies in experimental economics that invoke financial incentives and conclude that in certain areas—"in particular, research on judgement and decision making" (395)—such incentives lead to "convergence of the data toward the performance criterion and reduction of the data's variance." The authors recommend that financial incentives, which are common practice in experimental economics, be used more widely in psychological studies so as to obtain more reliable and robust data. Camerer and

Hogarth's (1999) literature review also shows that while incentivizing participants financially does not always improve the rational standing of their decision and judgment, it actually improves it in tasks similar to making a judgment in response to a thought experiment.

In the *reasons* condition, the vignette and questions were preceded by a screen which instructed participants that they would have to provide detailed explanations of their answers. The aim of this manipulation consisted in fostering an increased sensitivity to rational justification of the chosen response. A large literature suggests that in many cases increasing accountability by asking participants to justify their judgment or decision improves the rational standing of these. For instance, Koriat, Lichtenstein, and Fischhoff (1980) find that requiring participants "to list reasons for and against each of the alternatives prior to choosing an answer" reduces the overconfidence bias (see also, e.g., the reduction of the sunk-cost fallacy in Simonson and Nye (1992)). In their important literature review, Lerner and Tetlock (1999) conclude that "[w]hen participants expect to justify their judgments […] [they tend to] (a) survey a wider range of conceivably relevant cues; (b) pay greater attention to the cues they use; (c) anticipate counter arguments, weigh their merits relatively impartially, and factor those that pass some threshold of plausibility into their overall opinion or assessment of the situation; and (d) gain greater awareness of their cognitive processes by regularly monitoring the cues that are allowed to influence judgment and choice'" (1999, 263).[5]

[5] Increasing accountability, e.g. by reason giving, can also *aggravate,* rather than *attenuate*, certain biases in judgment and decision making (Lerner and Tetlock, 1999).

A final condition made use of *analytic priming*: Before receiving the vignettes and questions, participants had to solve a simple mathematical puzzle—a standard procedure to trigger analytic cognition. To our knowledge, the puzzle we used has not been employed in the social-psychological literature, but the procedure of triggering analytic cognition by means of a mathematical problem is standard practice. Paxton et al. (2011) and Pinillos et al. (2011) use the Cognitive Reflection Test, which consists of three simple mathematical puzzles with counterintuitive answers, to prime reflection and reasoning in their participants before giving them some trolley-style moral dilemmas. In a study regarding different explanations of the contrast-sensitivity of knowledge ascriptions, Gerken and Beebe (2014) also employ the Cognitive Reflection Test. On their view, the contrast effect of knowledge ascription is due to a bias in focus on selective bits of evidence. High CRT scores, they hypothesized in ways consistent with the reflection defense, should correlate with lesser susceptibility to the bias, but they failed to find any such correlation.

All four manipulations were independent ways to elicit reflection during the process of deliberation. The control condition, in which vignette and questions were presented without further ado, was intended to be similar to the characteristics of past empirical research of experimental philosophers, which have allegedly failed to elicit reflective judgments.

Finally, reaction time was measured for all five conditions to explore whether people who answered more slowly, presumably because they reflected more before reporting a judgment, answered differently from those who answered more quickly

---

It would be interesting to see how advocates of the reflection defense respond to such findings.

(and probably unreflectively).

Our choice of cases was guided by the following four considerations. First, they should have received widespread attention. Second, there should be relatively little controversy among professional philosophers about what the "correct" response is. As the advocates of the reflection defense make plain, not just any change in judgment occasioned is welcome: they expect reflection to foster increased alignment with the responses favoured by professional philosophers, at least if there is a consensus. Third, in order to assess whether encouraging reflection leads people to give responses aligned to philosophers', the cases must have elicited some disagreement among lay people (in light of past research). Finally, the cases must be drawn from several areas of philosophy. Overall, we chose four scenarios comprising of influential classics and more recent cases. Since all of them are rather well-known, we will confine ourselves to brief summaries here (all vignettes and questions are stated in full in the appendix).

The first vignette was a Gettier case, an adaptation of Russell's (1948) well-known *Clock* scenario: Wanda reads the time off a clock at the train station. This clock has been broken for days, yet happens to display the correct time when Wanda looks at it. Philosophers by and large agree that Wanda does not know what time it is (Sartwell, 1992 is an exception), and Machery et al. (2018) show that lay people are divided about this case: A surprisingly large proportion ascribe knowledge in this case.

A second vignette focused on the thesis that knowledge entails belief. Myers-Schulz and Schwitzgebel (2013) have reported astonishing evidence according to which people are sometimes willing to ascribe knowledge without ascribing belief (see also Murray, Sytsma, and Livengood, 2013). We used Myers-Schulz and

Schwitzgebel's scenario, which is an adaptation of Radford's (1966) famous *Queen Elizabeth* example. Kate has studied hard for her history exam; when she faces a question about the year of Queen Elizabeth's death, she blanks, despite the fact that she has prepared the answer and recited it to a friend. Eventually, Kate settles on a precise year without much conviction—1603—which is the correct response. In a between-subjects design, participants receiving the first condition were asked whether Kate *believed* Elizabeth died in 1603; in the second condition, they were asked whether Kate *knew* Elizabeth died in 1603. Most philosophers hold that knowledge entails belief (but see Radford, 1966; Williams, 1973), but Myers-Schulz and Schwitzgebel (2013) as well as Murray et al. (2013) suggest that many lay people are willing in some circumstances to ascribe knowledge while denying belief.

The third experiment explored the epistemic side-effect effect or "ESEE." Beebe and Buckwalter (2010) report that knowledge ascriptions regarding side effects are sensitive to the latter's general desirability (see also Beebe and Jensen (2012), Dalbauer and Hergovich (2013), Beebe and Shea (2013) Buckwalter (2014), Turri (2014), Beebe (2015) and Kneer (2018)). Beebe (2013) has produced similar data for belief ascriptions.

We used a scenario from Beebe and Jensen (2012) inspired by Knobe's (2003) influential case: The CEO of a movie-studio is approached by his vice-president who suggests implementing a new policy. The new policy would increase profits and make the movies better or worse from an artistic standpoint. The CEO replies that he does not care about the artistic qualities of the movies; the policy is implemented and the vice-president's predictions borne out. The question asked whether the CEO knew or believed the new policy would make the films better or worse from an artistic standpoint. To our knowledge, few philosophers, if any, think that the proper

application of the concepts of knowledge and belief is sensitive to desirability; by contrast, the extensive body of research on the ESEE suggests that for many lay people the ascription of knowledge and belief is sensitive to this factor.

The fourth and final vignette was an adaptation of Kripke's (1972) *Gödel* case, drawn from Machery, Mallon, Nichols, and Stich (2004). John has learned in school that a man called "Gödel" proved the incompleteness theorem, but it turns out that the proof was in fact accomplished by Gödel's friend, Schmidt. The question asked whether the name "Gödel" refers to the man who proved the incompleteness theorem or the man who got hold of the manuscript and claimed credit for it. Nearly all philosophers share Kripke's judgment that "Gödel" refers to the man originally called in the scenario "Gödel," but extensive research suggests that many Americans report the opposite judgment (Machery et al., 2004, 2010, 2015).

4. Experiment 1

4.1 Participants and Materials

Participants were recruited on Amazon Mechanical Turk in exchange for a small compensation. Data sets from participants failing a general attention test or a vignette-specific comprehension test were discarded. The final sample consisted of 179 respondents (Male: 44.7%; mean age: 35.1; age SD: 11.9; age range: 18-69).

The first experiment used the Clock vignette; participants were randomly assigned to one of the five conditions described above: Control, Delay (40 seconds), Incentive, Reasons, Priming. Response times were collected for all five conditions. Having responded to the target questions, all participants completed a 10-item version of Epstein's Rational-Experiential Inventory and a demographic questionnaire.

4.2 Results and Discussion

4.2.1 Main Results

A logistic regression was performed to ascertain the effect of condition on the likelihood that participants judge that the character does not have knowledge. The logistic regression model was not statistically significant, $\chi^2(4) = 4.35$, $p = .36$. The model only explained 3.2% (Nagelkerke R^2) of the variance in participants' answers and correctly classified only 58.7% of the data points. With standard assumptions of $\alpha = .05$ and a moderate effect size ($w = .3$), the power of our χ^2-test is very high (>.91); power remains high (>.7) for smaller effect sizes ($w \geq .23$), but is low for small effect sizes (Faul, Erdfelder, Lang, and Buchner, 2007). Figure 1 presents the proportion of "does not know" answers for the five conditions. Hence, we failed to find any evidence that encouraging careful reflection makes a difference to people's judgments about the clock case.

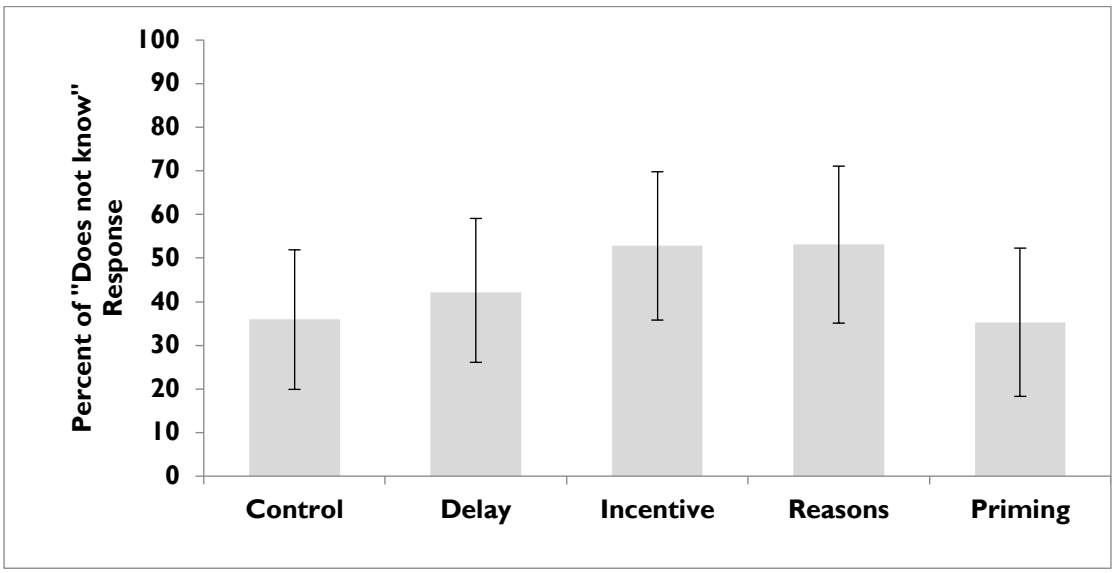

Fig.1: Percentages of participants who deny knowledge in the 5 conditions of Experiment 1 (bars: 95% confidence intervals)

4.2.2 Response Time

We also examined whether people who answer more slowly answer differently, excluding participants in the Delay condition. Averaging across the four other conditions, we did not find any evidence that slower participants answer differently ($r(141)=.02$, $p=.85$).[6] Figure 2 reports the proportion of answers in line with philosophers' consensual judgment for the 50% faster and 50% slower participants in the Clock case ("Does not know") and two other cases with categorical data of the following experiments.

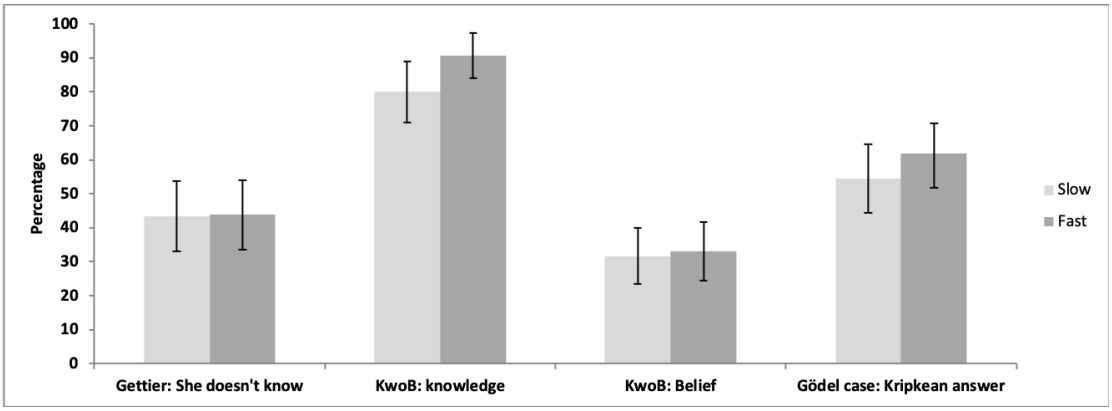

Fig.2: Percentage of participants responding "She does not know" in Experiment 1 (Gettier case), "She knows" in the knowledge condition of Experiment 3, "She believes" in the belief condition of Experiment 3, and giving a Kripkean response in Experiment 5 (bars: 95% confidence intervals)

The results are similar when one looks at each condition (including Delay) separately (Control: $r(39)=-.04$, $p=.80$; Delay: $r(38)=.24$, $p=.15$; Incentive: $r(36)=.16$, $p=.36$; Reasons: $r(32)=-.04$, $p=.83$; Priming: $r(34)=-.05$, $p=.79$). Thus, we failed to find any

[6] The results are similar if one excludes the reaction times two standard deviations below and above the mean RT ($r(141)=-.07$, $p=.55$).

evidence that people who answer more slowly, possibly because they reflect about the case, answer differently.

4.2.3 Analytic Thinking

In addition, we examined whether people who report a preference for analytic thinking answer differently. Averaging across the five conditions, we did not find any evidence that REI scores predict participants' response to the Clock case ($r(179)=-.12$, $p=.12$). Figure 3 reports the proportion of answers in line with philosophers' judgment for the 50% most reflective and 50% least reflective participants for the Clock case, as well as two cases used in the other experiments with categorical data. The results are similar when one looks at each condition separately (Control: $r(39)=-.18$, $p=.27$; Delay: $r(38)=-.31$, $p=.06$; Incentive: $r(36)=.05$, $p=.78$; Reasons: $r(32)=-.16$, $p=.37$; Priming: $r(34)=.03$, $p=.85$). So, there is no evidence that people who have a preference for thinking answer differently from people who don't have such preference.

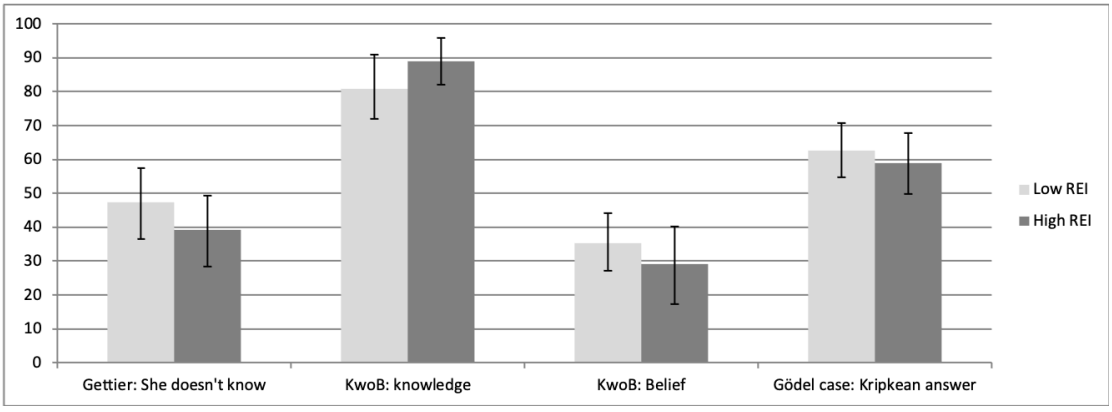

Figure 3: Comparison of more reflective and less reflective participants responding "She does not know" in Experiment 1 (Gettier case), "She knows" in the knowledge condition of Experiment 3, "She believes" in the belief condition of Experiment 3, and giving a Kripkean response in Experiment 5 (bars: 95% confidence intervals)

Note that in contrast to other Gettier cases (Machery et al., 2015), lay people tend not to share philosophers' judgment that the protagonist in a Clock case does not know the relevant proposition. This is in line with previous studies examining the Clock case (Machery et al., 2018).

5. Experiment 2: Follow-up to Experiment 1

While we failed to find any significant result with our manipulations, two of the manipulations of Experiment 1 seemed to lead participants to agree more with philosophers: asking participants to provide reasons for their answer and providing monetary incentives to think things through in detail. To explore these results further, we replicated the Incentive, Reasons and Control conditions of Experiment 1 with a larger sample size.

5.1 Participants and Materials

Participants were recruited on Amazon Mechanical Turk in exchange of a small compensation. Participants who failed the attention check or answered the comprehension question incorrectly were removed. Our final sample consisted of 264 respondents (male: 39.0%; mean age: 40.0; age SD: 13.1; age range: 19-73). Participants were randomly assigned to the Control, Incentive, or Reasons conditions. The vignette, instructions, and procedure were otherwise identical to those of Experiment 1.

5.2 Results and Discussion

5.2.1 Main Results

Participants in the three conditions answered differently ($\chi^2(2, 264)=8.2$, $p=.017$)[7], but the two manipulations did *not* lead participants to agree with philosophers about the clock case; rather, they led them to judge that the character in the clock case *knows* that it is 3:00pm. Figure 4 visualizes the results.

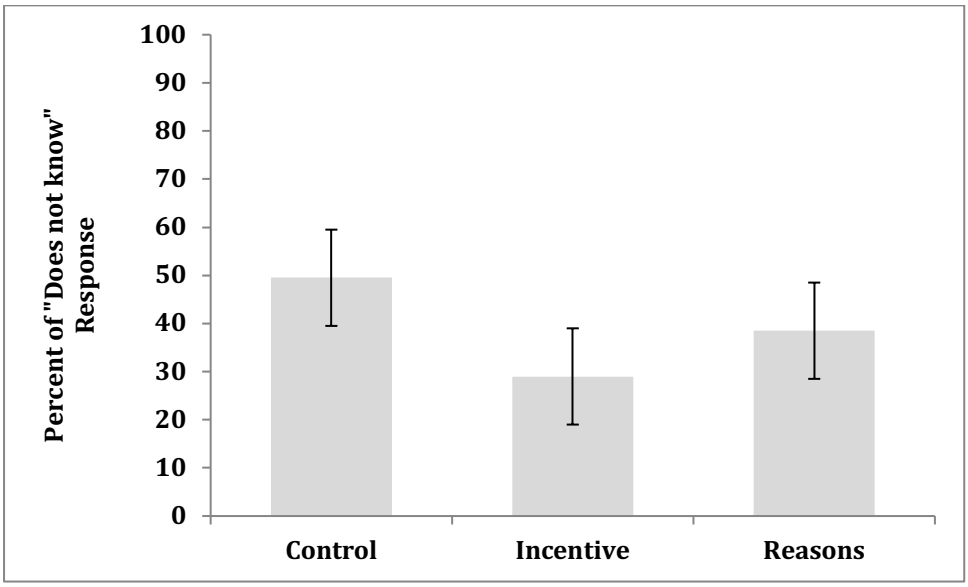

Fig. 4: Percentages of participants who judge that Wanda does not know the time in the 3 conditions of Experiment 2 (bars: 95% confidence intervals)

One may be surprised by the difference between the results in the Incentive and Reasons conditions in this study and in Study 1. We do not have a ready explanation, except for the fact it may be simply random sampling variation.

[7] Control vs. Incentive: $\chi^2(1, 186)=8.1$, $p=.004$; Control vs. Reasons: $\chi^2(1, 171)=2.1$, $p=.15$.

5.2.2 Response Time

We also examined whether people who answer more slowly answer differently. Averaging across the three conditions, we did not find any evidence that slower participants do so ($r(264)=.02$, $p=.77$; see Figure 2).[8] The results are similar when one looks at each condition separately (Control: $r(91)=.05$, $p=.64$; Incentive: $r(93)=0.0$, $p=.99$; Reasons: $r(78)=-.10$, $p=.38$). Thus, as was the case in Experiment 1, we failed to find any evidence that people who answer more slowly answer differently.

5.2.3 Analytic Thinking

In addition, we examined again whether people who report a preference for thinking answer differently. Averaging across the three conditions, we did not find any evidence that REI scores predict participants' response to the Clock case ($r(264)=-.09$, $p=.16$; see Figure 3). The results are similar in the Incentive condition ($r(93)=.06$, $p=.56$) and the Reasons condition ($r(78)=-.11$, $p=.32$). By contrast, participants in the control condition with higher REI scores (participants who report a taste for thinking) were more likely to *disagree* with philosophers and judge that the character in the Gettier case knows that it is 3:00pm ($r(93)=-.21$, $p=.05$). Again, there is little evidence that people who have a preference for thinking answer differently from people who don't have such preference; and when they do, the evidence suggests they tend to disagree *more* with philosophers. Differently put, the influence of increased reflection is limited, and where it does produce a difference, it *decreases* alignment with textbook epistemology.

[8] The results are similar if one excludes the reaction times two standard deviations below and above the mean RT ($r(258)=.01$, $p=.92$).

6. Experiment 3: Knowledge and Belief

So far, the results suggest that reflective judgments do not foster increased alignment with philosophical orthodoxy. In fact, reflection does not have much of an influence in the first place. But our results so far are limited to a single case drawn from one area of philosophy (the Clock case in epistemology). The following studies examine whether our findings generalize to other thought experiments and other areas of philosophy, starting with another case in epistemology. Experiment 3 focuses on the question of whether knowledge entails belief. The issue was first raised by Radford (1966), whose central thought experiment, *Queen Elizabeth* (described above), was previously tested by Myers-Schulz and Schwitzgebel (2013).

6.1 Participants and Materials

Participants were recruited on Amazon Mechanical Turk in exchange of a small compensation. Participants who failed the attention check or answered the comprehension question incorrectly were removed. Our final sample consisted of 385 respondents (male: 35.6%; mean age: 35.4; age SD: 16.2; age range: 18-83).

Our study had a 5x2 between-subjects design. Participants were randomly assigned to one of the ten conditions invoking five manipulations (Control, Delay, Incentive, Reasons, and Priming) and two epistemic states (knowledge and belief). Participants in the Knowledge conditions had to decide whether the character in the vignette knew that Queen Elizabeth died in 1603, participants in the Belief conditions whether she believed it. The instructions and procedures were identical to those of Experiment 1. The only difference consisted in the delay in the Delay condition. Participants had to wait 60 seconds before they could register their response, which we estimated was twice as long as it would take to read the case leisurely.

6.2 Results and Discussion

6.2.1 Main Results

A logistic regression was performed to ascertain the effect of our manipulations and of the Knowledge vs. Belief factor on the probability that participants judge that the character knows or believes that Queen Elizabeth died in 1603. The logistic regression model was statistically significant, $\chi^2(4) = 112.5$, $p<.001$. It explained 33.8% (Nagelkerke R^2) of the variance in participants' answers and correctly classified 74.5% of the data points. The Knowledge vs. Belief factor was statistically significant: Participants were significantly less likely to answer that the character believes that Queen Elizabeth died in 1603 than they were likely to answer that she knows that Queen Elizabeth died in 1603 (Wald=85.9, $p<.001$). By contrast, the manipulations were not statistically significant (Wald=1.1, $p=.86$). With standard assumptions of $\alpha=.05$ and a moderate effect size ($w=.3$), the power of our χ^2-test is very high (>.99); power remains high (>.7) for small to moderate effect sizes ($w\geq.16$), but is low for small effect sizes (Faul et al., 2007). Figure 5 presents the proportion of "knows" and "believes" answer for the five conditions.

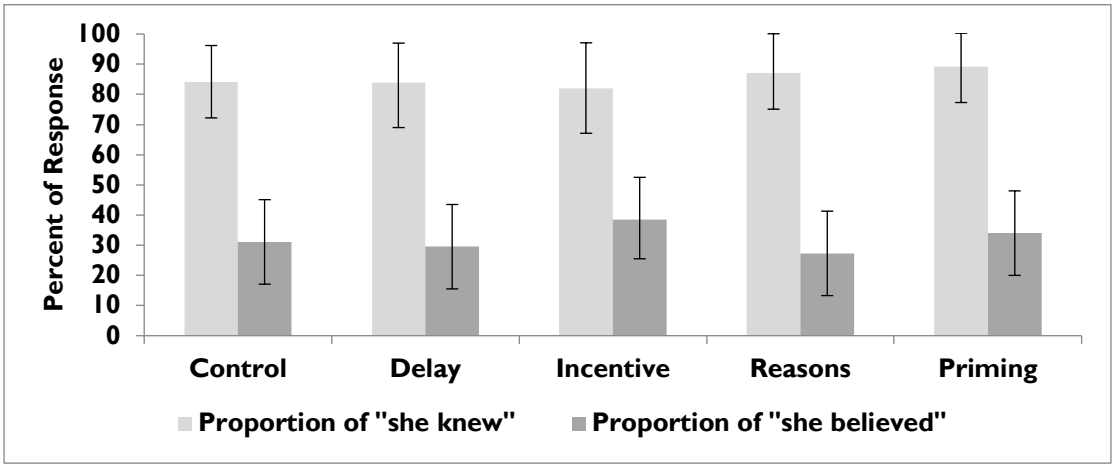

Fig. 5: Percentages of knowledge ascription and belief ascription in the 5 conditions of Experiment 3 (bars: 95% confidence intervals)

In our experiment, we replicated the results reported by Schwitzgebel and Myers-Schulz (2012), which cast doubt on the entailment thesis (but see Rose and Schaffer, 2013; Buckwalter, Rose, and Turri, 2015). We failed to find any evidence that compelling people to take their time in answering, telling them in advance that they will have to justify their answers, paying them to be accurate, *or* priming them to embrace an analytic cognitive style make any difference in their ascription of either knowledge or belief to the character in Schwitzgebel and Myers-Schulz's case.

6.2.2 Response Time

We also examined whether people who answer more slowly answer differently, excluding participants in the Delay condition. Averaging across the four other conditions, we did not find any evidence that in the Knowledge condition or in the Belief condition slower participants answer differently (respectively, $r(125)=.02$, $p=.81$ and $r(191)=.02$, $p=.75$; see Figure 2).[9] The results are largely similar when one looks at each condition separately (Belief condition: Control: $r(45)=.112$, $p=.44$; Delay: $r(44)=-.26$, $p=.09$; Incentive: $r(52)=-.14$, $p=.33$; Reasons: $r(44)=.17$, $p=.28$; Priming: $r(50)=.001$, $p=.99$; Knowledge condition: Control: $r(38)=-.18$, $p=.28$; Delay: $r(25)=.26$, $p=.21$; Incentive: $r(28)=.29$, $p=.14$; Reasons: $r(31)=-.06$, $p=.77$; Priming: $r(28)=-.16$ $p=.41$). Thus, we failed to find any evidence that people answer differently when, on their own, they take their time in considering Schwitzgebel and Myers-Schulz's case and in providing an answer.

[9] The results are similar if one excludes the reaction times two standard deviations below and above the mean RT ($r(145)=-.07$, $p=.40$ and $r(288)=-.09$, $p=.19$).

6.2.3 Analytic Thinking

Finally, we examined whether people who report a preference for thinking analytically answer differently. In the Knowledge condition, averaging across the five conditions, we did not find any evidence that REI scores predict participants' response to the target question ($r(150)=.07$, $p=.42$; see Figure 3). In the Belief condition, averaging across the five conditions, we did not find any evidence that REI scores predict participants' response ($r(235)=-.06$, $p=.37$; see Figure 3). For the individual conditions, none of the Bonferroni-corrected *p*-values attained significance (all $ps>.1$). The results are largely similar when one looks at uncorrected *p*-values. Belief condition: Control: $r(45)=.01$, $p=.93$; Delay: $r(44)=-.33$, $p=.03$; Incentive: $r(52)=-.02$, $p=.90$; Reasons: $r(44)=-.08$, $p=.62$; Priming: $r(50)=.01$, $p=.92$. Knowledge condition: Control: $r(38)=-.09$, $p=.61$; Delay: $r(25)=.20$, $p=.34$; Incentive: $r(28)=.37$, $p=.05$; Reasons: $r(31)=-.02$, $p=.91$; Priming: $r(28)=-.17$ $p=.40$.

In two sub-conditions – Incentive/Knowledge and Delay/Belief – the uncorrected *p*-values just about reach significance. Given that none of the overall p-values, or any of the individual Bonferroni-corrected p-values attain significance, this clearly does not constitute systematic evidence that people drawn to more analytic thinking answer differently from those who are not so disposed.

7. Experiment 4: Epistemic Side Effect Effect

Experiment 4 investigates further whether lay people's reflective judgments in reaction to epistemological cases vary from the judgments reported so far by experimental philosophers. In this case we focused on the asymmetric ascriptions of knowledge and belief regarding differently desirable side effects. The findings of previous studies by Beebe and Buckwalter (2010) and Beebe (2013) are just as much

at odds with standard epistemological doctrine as those reported by Myers-Schulz and Schwitzgebel (2013).

7.1 Participants and Materials

Participants were recruited on Amazon Mechanical Turk in exchange of a small compensation. Participants who failed the attention check or answered the comprehension question incorrectly were removed. Our final sample consisted of 701 respondents (male: 41.7%; mean age: 35.4; age SD: 11.6; age range: 18-73).

Our study used the *Movie Studio* scenario (described above, for details see the Appendix), and was a 5x2x2 between-subjects design. Each participant was assigned to one of the 20 conditions differing with respect to manipulation (Control, Delay, Incentive, Reasons, Priming), desirability of the side effect (better movies, worse movies), and epistemic state (knowledge, belief). Answers were collected on a 7-point Likert scale; participants reported to what extent they agreed or disagreed that the protagonist believed or knew that the newly adapted policy would make the movies better or worse from an artistic standpoint. The instructions and procedures were identical to those of Experiment 1. The only difference consisted in the delay in the Delay condition. Participants had to wait 40 seconds before they could register their response, which we estimated was twice as long as it would take to read the case leisurely.

7.2 Results and Discussion

7.2.1 Main Results

An ANOVA with the five manipulations, the Better vs. Worse factor, and the Knowledge vs. Belief factor was performed to ascertain the effect of our

manipulations on the probability that participants judge that the character knows or believes that the movies were made worse or better. The Better vs. Worse factor was significant ($F(1, 681)=262.76$, $p<.001$, $\eta^2=.28$) as was the Knowledge vs. Belief factor ($F(1, 681)=159.87$, $p<.001$, $\eta^2=.07$). Participants are more likely to ascribe knowledge and belief in the Worse condition than in the Better condition, and they are more likely to ascribe knowledge than belief. By contrast, our manipulations did not produce any significant effect ($F(1, 681)=.35$, $p=.85$). With standard assumptions of $\alpha=.05$ and a moderate effect size ($f=.25$), the power of an F-test is very high (>.99); assuming a small effect size ($f=.10$), power (.75) is still high (Faul et al., 2007). Figure 6 presents the means of the "knows" and "believes" answers for the five conditions for the worse and better conditions.

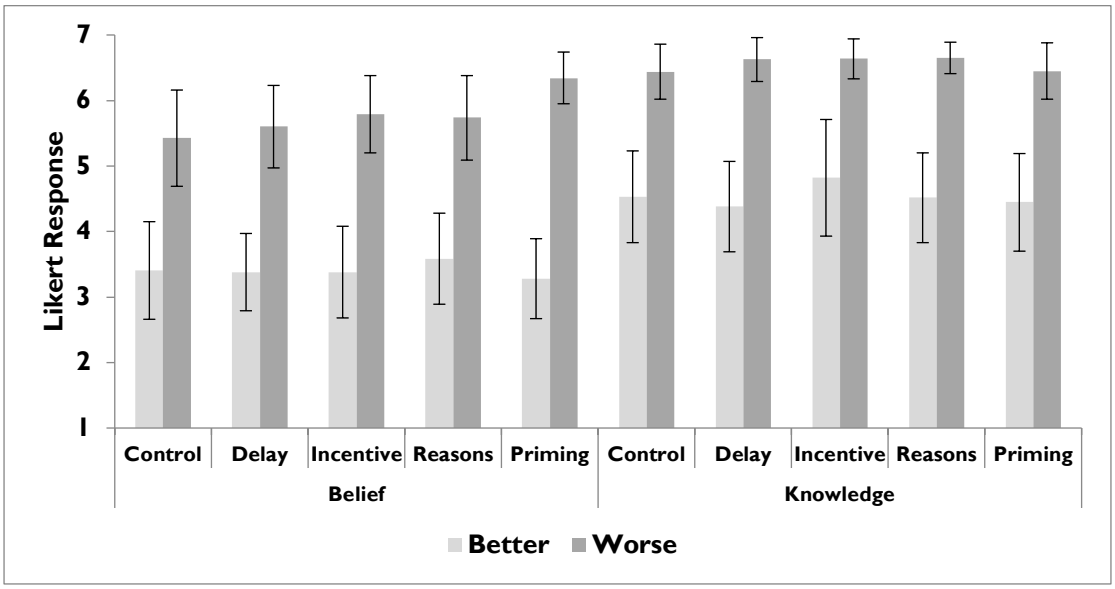

Fig. 6: Mean agreement with the claim that the director knew or believed the movies would become better or worse from an artistic standpoint in the 20 conditions of Experiment 4 (bars: 95% confidence intervals)

Thus, we replicated Beebe and Buckwalter's (2010) and Beebe's (2013) findings: The desirability of an action influences the ascription of knowledge and belief. In addition, we failed to find any evidence that compelling people to take their

time in answering, telling them in advance that they will have to justify their answers, paying them to be accurate, or priming them to embrace a reflective cognitive style make any difference in their answers, and it is likely that if there were small or moderate effects to be found, we would have found them.

7.2.2 Response Time

We also examined whether people who answer more slowly answer differently, excluding participants in the Delay condition. Averaging across the four other conditions, we did not find any evidence that they do (Harm and Knowledge conditions: $r(142)=.06$, $p=.51$; Help and Knowledge conditions: $r(136)=-.02$, $p=.80$; Harm and Belief conditions: $r(145)=-.05$, $p=.56$; Help and Belief conditions: $r(140)=-.07$, $p=.43$).[10] Figure 7 reports the scatterplot for these four conditions.

[10] The results are similar if one excludes the reaction times two standard deviations below and above the mean RT (Help and Knowledge conditions: $r(134)=-.07$, $p=.42$; Harm and Belief conditions: $r(140)=-.07$, $p=.39$; Help and Belief conditions: $r(139)=-.07$, $p=.39$), except for the Harm and Knowledge conditions: $r(139)=.20$, $p=.02$.

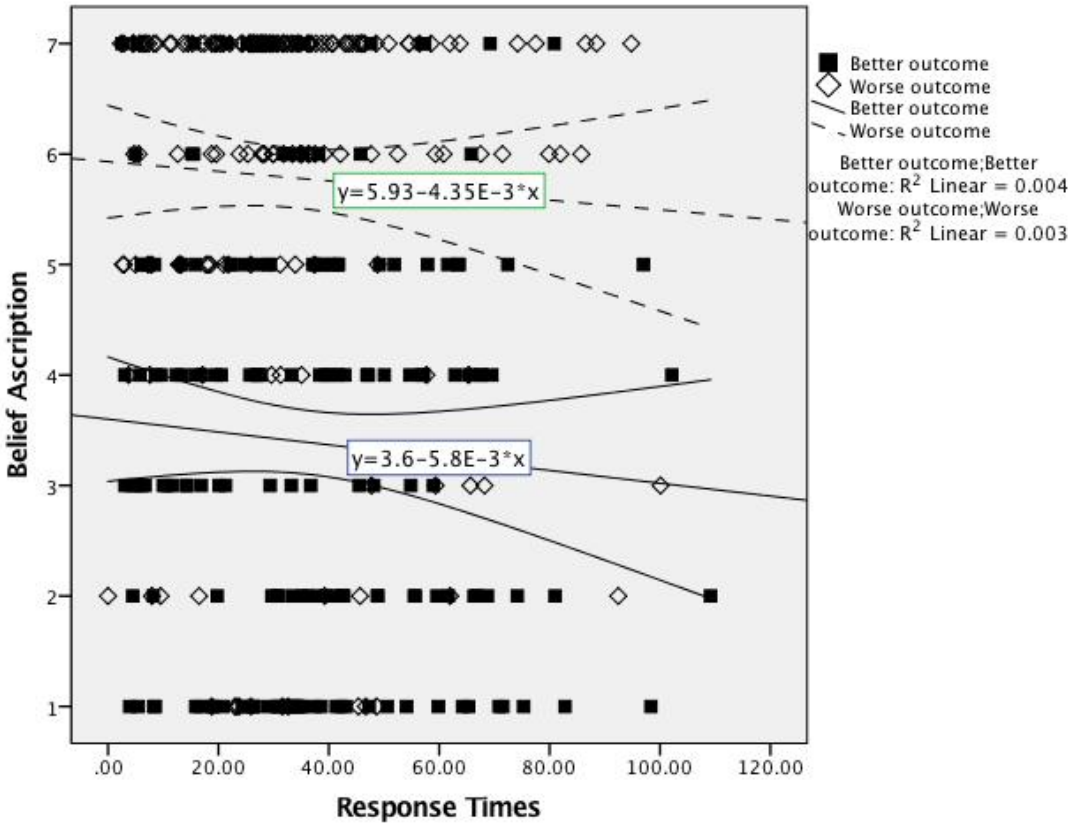

Figure 7a: Scatterplot and regression lines for the ascription of belief in experiment 4 as a function of participants' response time (after exclusion of data points larger than 2 SD above the mean)

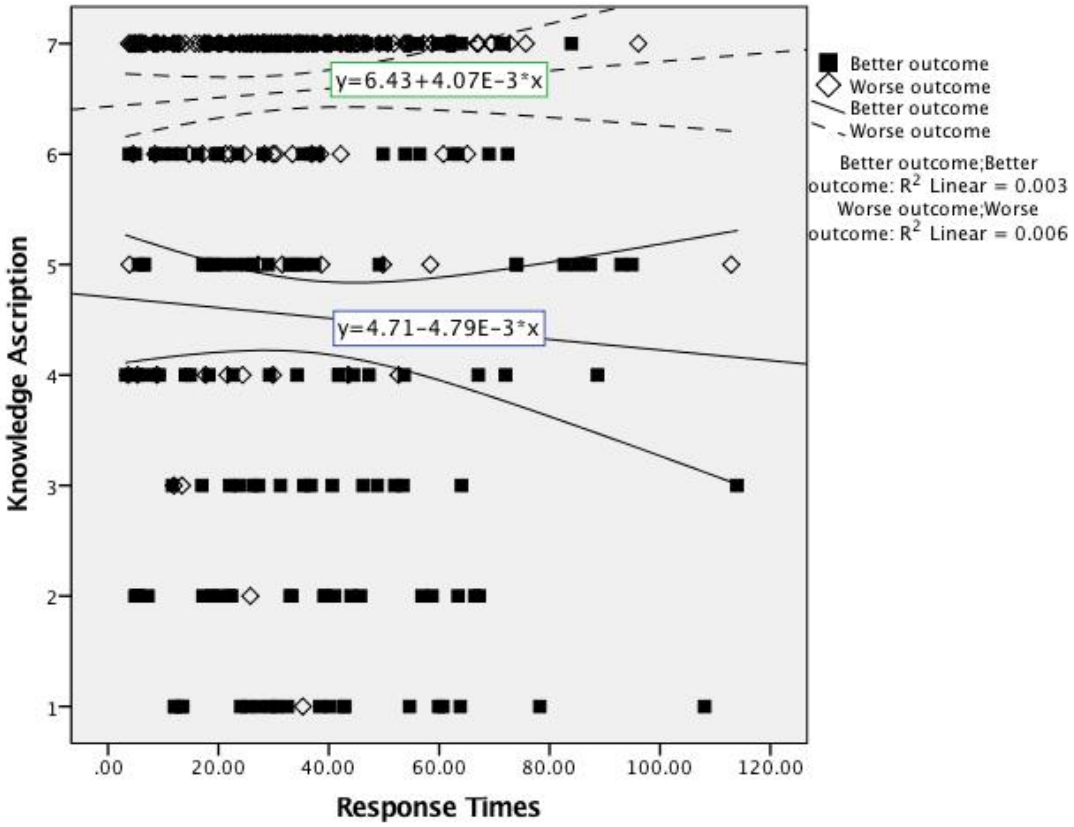

Figure 7b: Scatterplot and regression lines for the ascription of knowledge in experiment 4 as a function of participants' response time (after exclusion of data points larger than 2 SD above the mean)

In sum, we failed to find any evidence that people who, on their own, take their time in considering the relevant case and in answering judge differently from people who don't.

7.2.3 Analytic Thinking

Finally, we examined whether people who report a preference for thinking answer differently. Averaging across the four other conditions, we did not find any evidence that participants with higher REI scores answer differently (Harm and Knowledge conditions: $r(231)=-.09$, $p=.23$; Help and Knowledge conditions: $r(119)=-.02$, $p=.77$;

Harm and Belief conditions: $r(180)=.02$, $p=.77$; Help and Belief conditions: $r(177)=-.05$, $p=.53$). Figure 8 reports the scatterplot for these four conditions.

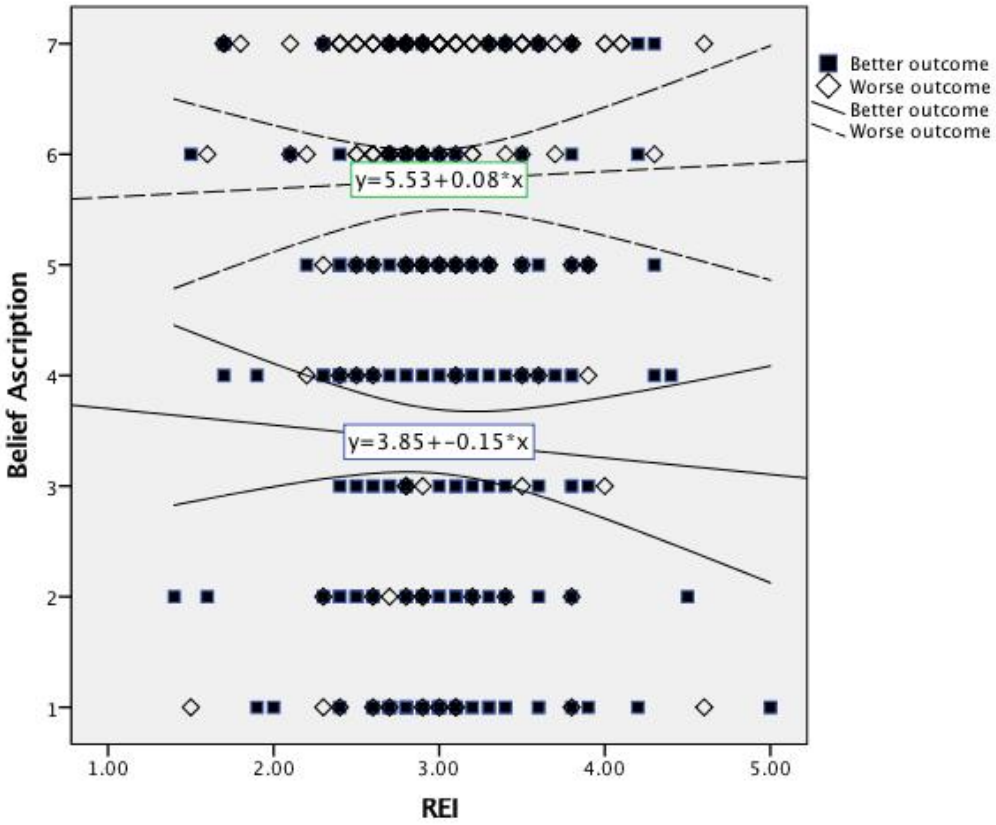

Figure 8a: Scatterplot and regression lines for the ascription of belief in experiment 4 as a function of participants' REI scores

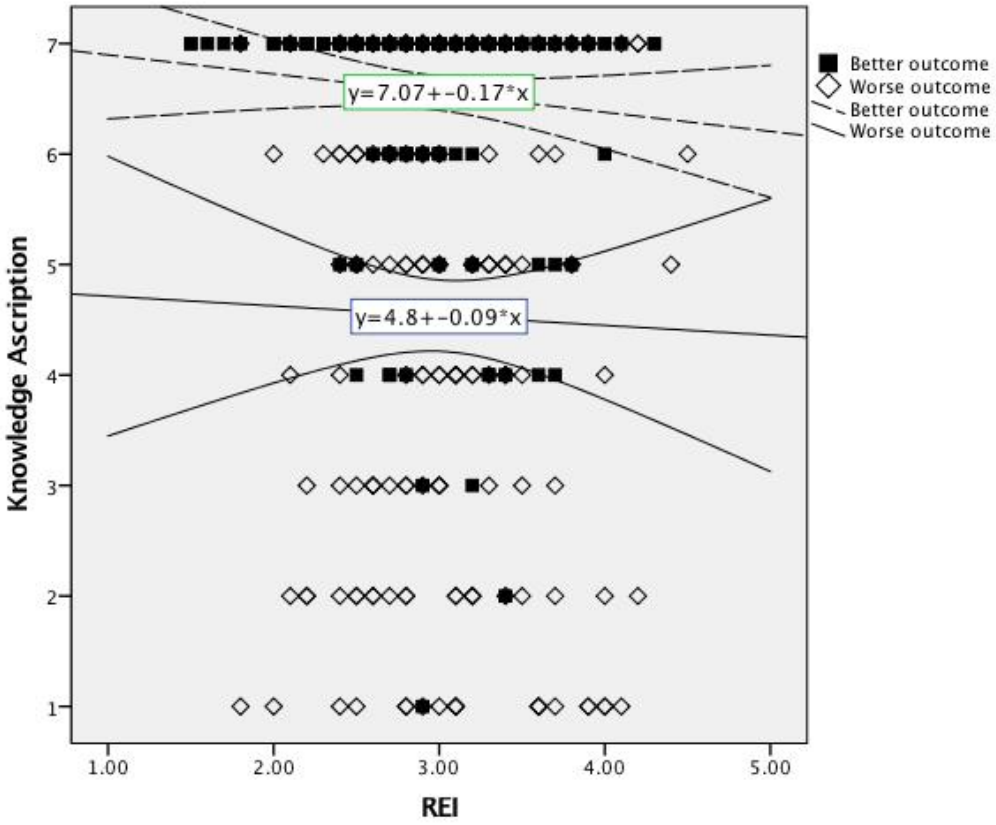

Figure 8b: Scatterplot and regression lines for the ascription of knowledge in experiment 4 as a function of participants' REI scores

So, there is no evidence that people who have a preference for analytic thinking answer differently from people who don't have such preference.

8. Experiment 5: The Gödel Case

Our final experiment examined whether the findings reported so far generalize to another area of philosophy: the philosophy of language. Following Kripke (1972), most philosophers assume that in the Gödel case the proper name "Gödel" refers to the man who stole the theorem. Previous work suggests however that for a substantial proportion of Americans (between 25% and 40%), the Gödel case elicits judgments more in line with the descriptivist theory of reference (Machery et al., 2004, 2010,

2015). Experiment 5 examined whether lay people agree more with philosophers when they report their reflective judgment.

8.1 Participants and Materials

Participants were recruited on Amazon Mechanical Turk in exchange of a small compensation. Participants who failed the attention check, answered the comprehension question incorrectly, or attempted to complete the survey multiple times (as evidenced by their IP address) were removed. Our final sample consisted of 274 respondents (male: 54.4%; mean age: 43.4; age SD: 13.5; age range: 21-79).

All participants were randomly assigned to one of five conditions: Control, Delay, Incentive, Reasons, Priming. The instructions and procedures were identical to those of Experiment 1. Participants had to decide whether the protagonist in the case is talking about the man who stole the theorem (Kripkean answer) or about the man who discovered the theorem (descriptivist answer). The only difference consisted in the delay in the Delay condition. Participants had to wait 60 seconds, which we estimated was twice as long as it would take to read the Gödel case leisurely.

8.2 Results and Discussion

8.2.1 Main Results

A logistic regression was performed to ascertain the effect of our manipulations on the probability that participants judge that the character is talking about the character originally called "Gödel" when he uses the proper name "Gödel." The logistic regression model was not statistically significant, $\chi^2(4) = 5.10$, $p = .28$. The model explained 2.5% (Nagelkerke R^2) of the variance in participants' answers and correctly classified 60.9% of the data points. The power of the χ^2 test, assuming a moderate

effect size (w=.3) was very high (.99); power remains high (>.7) for small to moderate effect sizes (w≥.19), but is low for small effect sizes (Faul et al., 2007). We also note that all the manipulations *decreased* the proportion of Kripkean responses, although not significantly so. Figure 9 presents the proportion of the "Gödel" answer for the five conditions.

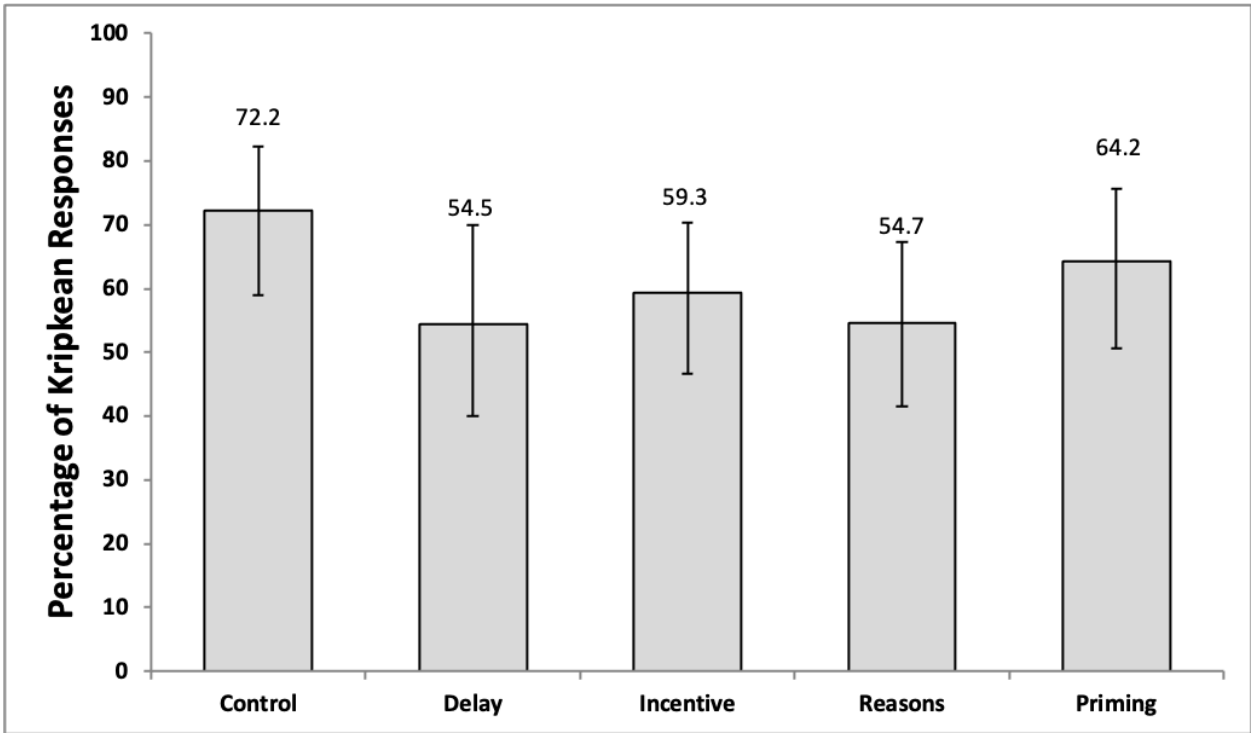

Fig.9: Percentages of Kripkean response in the 5 conditions of Experiment 5 (bars: 95% confidence intervals)

Thus, we failed to find any evidence that compelling people to take their time in answering, telling people in advance that they will have to justify their answers, paying them to be accurate, or priming them to embrace a reflective analytic style improved people's responses to the Gödel case, and it is likely that if there were a moderate or even a small to moderate effect to be found, we would have found it.

8.2.2 Response Time

We also examined whether people answer differently when they answer more slowly, excluding participants in the Delay condition. Averaging across the four other conditions, we did not find any evidence that they do ($r(219)=-.007$, $p=.92$; see Figure 2).[11] None of the uncorrected (and a fortiori Bonferroni-corrected) p-values for the individual conditions attained significance: Control: $r(54)=-.10$, $p=.48$; Delay: $r(55)=-.12$, $p=.39$; Incentive: $r(59)=-.07$, $p=.62$; Reasons: $r(53)=.045$ $p=.76$; Priming: $r(53)=.08$, $p=.57$. Thus, we failed to find any evidence that people who, on their own, take their time in considering the Gödel case and in answering answer differently from people who don't.

8.2.3 Analytic Thinking

In addition, we examined whether people who report a preference for thinking are more likely to agree with philosophers. Averaging across the five conditions, we did not find any evidence that REI scores predict participants' response to the Gödel case ($r(274)=-.04$, $p=.512$; see Figure 3). None of the Bonferroni-corrected p-values for the individual conditions attained significance (all ps>.05). Except for the priming condition, the same held for uncorrected p-values: Control: $r(54)=-.14$, $p=.32$; Delay: $r(55)=-.23$, $p=.10$; Incentive: $r(59)-.04$, $p=.75$; Reasons: $r(53)=-.11$ $p=.45$; Priming: $r(53)=.33$, $p=.01$. So, there is no systematic evidence that people who have a preference for analytic thinking agree more with philosophers about the Gödel case than people who don't have such a preference. While suggestive, the correlation in the

[11] The results are similar if one excludes the reaction times two standard deviations below and above the mean RT ($r(213)=-.003$, $p=.97$).

Priming condition should not give solace to philosophers since, if it isn't a mere accident, it goes in the wrong direction: People who are less reflective are more likely to give the response in line with philosophers' judgments.

9. Discussion

9.1 Meta-Philosophical Implications of the Experimental Studies

The reflection defense assumes that increased reflection influences people's responses to philosophical cases and improves them, in the sense of bringing them more into alignment with philosophical orthodoxy. Since experimental philosophers do not ordinarily encourage extensive reflection characteristic of the philosophical method, the data thus collected—or so the argument goes—is of no use. We put the empirical adequacy of the reflection defense to the test with respect to four well-known thought experiments.[12] For each, philosophers agree as to what constitutes the correct response.

Focusing on a thin conception of reflection and reflective judgment, we have examined two types of factors that might be conducive to reflective deliberation: the *circumstances* under which the deliberative process takes place, and *individual dispositions* to engage in careful reflection. As regards the former, we adapted a variety of standard manipulations from social psychology and experimental economics to encourage diligent reflection: time delay, financial incentives, reason specification, and analytic priming. Out of the 18 conditions with manipulations contrasted with the respective control conditions across five experiments, we could

[12] We also note, though do not elaborate on this point, that we replicated all the original experimental-philosophy studies, in line with Cova et al. (2018).

only detect a significant difference—or some sign of "influence"—in two comparisons: In Experiment 2, financial incentives and reason specification somewhat changed responses *vis-à-vis* the control condition, yet in both cases the increased reflection produced results that were *less* in alignment with philosophical theory. As regards circumstances then, the Influence and Alignment assumption has proven a nonstarter: In nearly all conditions we failed to detect an influence of reflection in the first place, and in the few cases where an influence was detected, it decreased alignment.

Concerning individual dispositions to engage in reflection, we contrasted the responses of participants who, out of their own free will, spent more time with the task with those who responded quickly on the one hand. We couldn't detect a significant difference in slow vs. fast responses for a single condition of any of the five experiments. We also explored whether people with a penchant for more analytic thinking (i.e. subjects on the "rational" end of the Rational-Experiential Inventory) respond differently from those who tend towards a more intuitive thinking style (those on the "experiential" end of the REI spectrum). Averaging across conditions, we did not find a significant difference in any of the five experiments. The Bonferroni corrected p-values for each of the 23 conditions were also nonsignificant. In short, participants who have a natural disposition to engage in more analytic thinking responded the same as those who do not. These results are consistent with the findings reported in previous studies, which attempted to measure the disposition to engage in reflection by means of the Need for Cognition inventory or the Cognitive Reflection Task.

Taking stock: In a series of five experimental studies with a total of over 1800 individual subjects, we found that neither a disposition to engage in reflection nor

circumstantial factors conducive to reflective judgment bring folk judgments into alignment with philosophical orthodoxy. In nearly all cases tested, they fail to have in impact entirely. Our studies had sufficient power to detect medium-sized effects with a very high probability, and small to medium effects with high probability. We cannot exclude the possibility of small effects induced by reflection. Note, however, that even if those were to be found, it's far from clear that this would make the reflection defense any more convincing. Take, for instance, Radford's unconfident examinee case, where in the control condition we found more than 80% of the participants to ascribe knowledge, while a mere 30% ascribed belief. The difference, defying orthodox epistemology, constitutes a large effect (h=1.15). Or consider the well documented epistemic side-effect effect: In the control conditions, the effect size of the divergence between positively and negatively valenced outcomes was very large for both belief (d=1.01) and knowledge (d=1.02). Now assume it could be shown that extensive reflection produces small effects in line with philosophical orthodoxy, e.g. increasing epistemic state ascriptions somewhat in the positively valenced Knobe-type cases. The epistemic side-effect effect will still be of at least moderate size, and it is more likely that they will remain large. The overall conclusion – that folk judgments frequently differs strongly from philosophical consensus and that extensive reflection does not bring the two into alignment – remains the same.

At this point, there are two responses available to proponents of the reflection defense: First, they could argue that the reason we did not find any effect is that our manipulations are poor means of leading people to engage in sufficiently reflective deliberation about philosophical cases even when reflection is thinly construed. Second, they could argue that the thin conception of reflective judgment that we have been working with here is not what they had in mind. We do not find either response

compelling. Let's start with the first kind of response. Combined with earlier studies, we now have eight different ways of inducing reflection thinly construed, and none of them seem to support the central presuppositions of the reflection defense, which are either that reflection sufficiently immunizes judgments about philosophical cases from sensitivity to allegedly irrelevant factors, or at least that it changes our judgments about those cases.

The second kind of response is no more compelling than the first. Of course, it is perfectly possible that reflection more thickly construed might lead people to change their judgments about philosophical cases, and we happily admit that we have done nothing to address this possibility. Having said that, it should be obviously unacceptable to attempt to rebut an empirical challenge to the way philosophers standardly use the method of cases by appealing to some unspecified account of reflection. A convincing response to the experimental challenge must explain not only *what* properties the judgments studied by experimental philosophers apparently lack, but also *why* these properties are important to the way philosophers standardly employ the method of cases. And here it is important that proponents of the reflection defense do not rely on a bait-and-switch strategy. If the reflection defense is deemed plausible and appealing at all, it is largely, we submit, because the notion of reflection is characterized thinly: Philosophical arguments, we agree, do not appeal to "gut reactions" to cases or to "shots from the hip" in response to these, but rather to careful and reflective judgments. However, the intuitive plausibility and appeal of the reflection defense thinly understood do not transfer to versions of the argument that appeal to thicker characterizations of reflection. If it's plausible that only careful, slow, reflective judgments about cases are philosophically relevant, is it equally plausible that only epistemically analytic judgments about cases are philosophically

relevant? Surely not. For one, many deny that any judgment is epistemically analytic (Williamson, 2007; Machery, 2017). And even if some are, it is not evident at all that the judgments made in response to cases are epistemically analytic. What's the upshot? Proponents of the reflection defense who appeal to thicker characterizations of reflection can't simply trade upon the initial plausibility and appeal of the reflection defense when reflection is characterized thinly on pain of engaging in bait-and-switch. What is required is a detailed characterization of reflection understood in some more substantial way, and clear arguments to the effect that the method of cases demands this type of thick reflection.

Until now, proponents of the reflection defense have not provided compelling arguments for thicker characterizations of the notion of reflection. Furthermore, we submit, compelling arguments will be hard to find. As noted in Section 2, an adequate characterization of reflection must be consistent with the way philosophers use thought experiments (the descriptive-inadequacy problem), but the thicker the characterization, the more likely it is that it will fall prey to the descriptive-inadequacy problem (Cappelen, 2012; Machery, 2017, chapter 1). For instance, Kauppinen's dialogical conception of reflection fails to capture how philosophers usually judge in response to cases. There is no doubt that, as Kauppinen insists, philosophers compare cases and try to identify ways in which particular cases are alike or differ, but they typically do not do this in the process of making a judgment about these cases. Rather, having made judgment about several cases, they try to identify potential reasons that explain their pattern of judgments. When Gettier (1963) proposes his ten-coin case, he does not compare it to other cases to conclude that the agent does not know that he has ten coins in his pocket. Nor do his readers. Rather, once we have judged in response to several cases, we compare those judgments to

identify the reasons that explain the relevant pattern of judgments. It is thus implausible that only judgments understood along Kauppinen's dialogical conception of reflection are philosophically relevant.

Furthermore, a thick conception of reflection must not imply by stipulation that the research done by experimental philosophers is not relevant for philosophical methodology (the stipulation problem). Differently put, simply stipulating a concept of reflection according to which the kinds of biases identified by experimental philosophers *cannot* arise is unhelpful. Instead, it must be shown by means of arguments or empirical evidence that reflection, understood in some thicker manner, results in judgments that do not fall prey to such biases, or at least do so at a much lower rate.

9.2 Why Doesn't Reflection Influence Judgment about Cases?

It is surprising that people who are disposed to engage in careful analytic thinking and people who are primed to engage in reflective deliberation do not judge differently in response to cases than people who read cases under conditions that are standardly used by experimental philosophers. Why is that? There are at least two answers to this question.

First, it could be that, contrary to what proponents of the reflection defense assume (premise 2 of the argument sketched in Section 2), participants in experimental-philosophy studies are already engaged, by themselves, in reflective deliberation when they respond to cases under conditions standardly used by experimental philosophers. If this were the case, then it would be unsurprising that priming people to be reflective or looking at people disposed to careful thinking would not make any difference, as we found.

While some subjects may, by themselves, engage in reflective deliberation on their own under conditions standardly used by experimental philosophers, we doubt that this is the case for all subjects, and while this explanation may be partly correct, it is incomplete. The reason is that many participants respond rather quickly to cases, too quickly for them to have had the time to engage in careful, reflective deliberation.

The second explanation of our surprising result is more radical: Typically, reflection does not change the judgments made in response to cases (for a similar view about moral judgments, see Haidt, 2001, and for a discussion of the limits of reflection, see Kornblith 2010). Rather, it merely leads people to find reasons for the judgments they made unreflectively in response to these cases. People who are inclined to engage in analytic thinking are merely better at finding arguments for the judgments they make about cases; people who are primed to engage in reflection are primed to find reasons for their judgments about cases. If finding arguments or justification is the product of reflection, then it is unsurprising that priming people to be reflective or looking at people disposed to careful thinking would not make any difference, as we found.

One may wonder why reflection does not change judgments about philosophical cases much, while it appears to influence judgment in the social-psychological and behavioral-economical literature, allowing people to overcome their spontaneous answers. We propose to explain the difference between our findings and the past research on reflection as follows. The judgments people make in the types of situations examined by social psychologists and behavioral economists are frequently mistaken by people's own lights; when this happens, they tend to change their responses when given an opportunity to reflect. By contrast, when judgments are made with confidence and are not erroneous by participants' own lights, reflection

does not influence judgment much; rather, it leads people to think of arguments for the judgments independently made.

The explanation proposed in this section, we submit, is the deepest reason why the reflection defense fails. It misunderstands the role of reflection. It assumes that reflection would lead us to judge differently in response to cases instead of, as our results suggest, prompting us to explore reasons, arguments, and justifications for them, while leaving the judgments themselves unchanged.

Conclusion

In this article, we have addressed the reflection defense put forward in response to the challenge against the use of cases inspired by experimental philosophy. We have shown experimentally that there is no systematic evidence which suggests that reflection, thinly understood, leads people to respond differently to cases than under standard experimental-philosophy conditions. This finding undermines the view according to which reflective and unreflective judgments about cases differ. Instead, we would like to suggest, people's response to thought experiments typically expresses deep-seated judgments, and reflection merely bolsters this judgment by pushing people to explore potential reasons for their judgment. While it is possible that the reflection defense might appeal to some thicker reflection, we see little reason for optimism. Given that both the expertise defense and the reflection defense have so far proven inadequate, we conclude that philosophers should take the experimentalist's challenge against the use of cases seriously.

Bibliography

Alexander, J. (2012). *Experimental philosophy: An introduction.* Cambridge: Polity.

Alexander, J. (2016). Philosophical expertise. In J. Sytsma and W. Buckwalter (eds.), *A companion to experimental philosophy* (pp. 555-567). Malden, MA: Wiley-Blackwell.

Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, *2*(1), 56-80.

Beebe, J. R. (2013). A Knobe effect for belief ascriptions. *Review of Philosophy and Psychology, 4*(2), 235-258.

Beebe, J. R. (2015). Do bad people know more? Interactions between attributions of knowledge and blame. *Synthese*, 1-25.

Beebe, J. R. (2016). Evaluative effects on knowledge attributions. In J. Sytsma and W. Buckwalter (eds.), *A companion to experimental philosophy* (pp. 359-367). Malden, MA: Wiley-Blackwell.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side‐effect effect. *Mind & Language, 25*(4), 474-498.

Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology, 25*(5), 689-715.

Beebe, J. R., & Shea, J. (2013). Gettierized Knobe effects. *Episteme, 10*(3), 219.

Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research, 86*(3), 495-532.

Buckwalter, W. (2014). The mystery of stakes and error in ascriber intuitions. *Advances in experimental epistemology*, 145-174.

Buckwalter, W., Rose, D., & Turri, J. (2015). Belief through thick and thin. *Noûs*, *49*(4), 748-775.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology, 42*(1), 116.

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of personality and social psychology, 51*(5), 1032.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, *19*(1-3), 7-42.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.

Colaço, D., & Machery, E. (2017). The intuitive is a red herring. *Inquiry*, *60*(4), 403-419.

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme*, *11*(02), 199-212.

Cova, F., et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*,

Dalbauer, N., & Hergovich, A. (2013). Is what is worse more likely?—the probabilistic explanation of the epistemic side-effect effect. *Review of Philosophy and Psychology, 4*(4), 639-657.

Deutsch, M. E. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method*. Cambridge, MA: MIT Press.

Dewey, J. (1910). *How we think*. Boston: D.C. Heath and Company.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of personality and social psychology, 71*(2), 390.

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.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences, 105*(45), 17538-17542.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25-42.

Garrett, H. E. (1922). *A study of the relation of accuracy to speed* (Vol. 8): Columbia university.

Gerken, M., and Beebe, J. R. (2014). Knowledge in and out of Contrast. *Nous*.

Gettier, E. L. (1963). Is justified true belief knowledge?. *analysis*, *23*(6), 121-123.

Gonnerman, C., Reuter, S., & Weinberg, J. M. (2011). *More oversensitive intuitions: Print fonts and could choose otherwise.* Paper presented at the One Hundred Eighth Annual Meeting of the American Philosophical Association, Central Division, Minneapolis, MN.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Hannon, M. (2018). Intuitions, reflective judgments, and experimental philosophy. *Synthese*, *195*(9), 4147-4168.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(03), 383-403.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*(1), 11-26.

Horvath, J. (2010). How (not) to react to experimental philosophy. *Philosophical Psychology*, *23*(4), 447-480.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical explorations, 10*(2), 95-118.

Kneer, M. (2018). Perspective and epistemic state ascriptions. *Review of Philosophy and Psychology*, *9*(2), 313-341.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190-194.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, *6*(2), 107.

Kornblith, H. (2010). What reflective endorsement cannot do. *Philosophy and Phenomenological Research*, *80*(1), 1-19.

Kripke, S. A. (1972). Naming and necessity. In *Semantics of natural language* (pp. 253-355). Springer Netherlands.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological bulletin, 125*(2), 255.

Liao, S. M. (2008). A defense of intuitions. *Philosophical Studies, 140*(2), 247-262.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*(1), 128-159.

Machery, E. (2011). Thought experiments and philosophical knowledge. *Metaphilosophy*, *42*(3), 191-214.

Machery, E. (2012). Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *27*(1), 37-54.

Machery, E. (2017). *Philosophy within its proper bounds.* London: Oxford University Press.

Machery, E., Deutsch, M., Sytsma, J., Mallon, R., Nichols, S., & Stich, S. P. 2010. Semantic intuitions: Reply to Lam. *Cognition*, 117, 361-366.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition, 92*(3), B1-B12.

Machery, E, Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2018). Gettier Was Framed! In McCready, M Mizumoto, J. Stanley, and S. Stich (eds.), *Epistemology for the rest of the world: Linguistic and cultural diversity and epistemology*. Oxford: Oxford University Press.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Usui, N., & Hashimoto, T. (2015). Gettier across cultures. *Noûs*.

Malmgren, A. S. (2011). Rationalism and the content of intuitive judgements. *Mind*, *120*(478), 263-327.

Mizrahi, M. (2015). Three arguments against the expertise defense. *Metaphilosophy*, *46*(1), 52-64.

Murray, D., Sytsma, J., & Livengood, J. (2013). God knows (but does God believe?). *Philosophical Studies, 166*(1), 83-107.

Myers‐Schulz, B., & Schwitzgebel, E. (2013). Knowing that P without believing that P. *Nous, 47*(2), 371-384.

Nado, J. (2015). Intuition, philosophical theorizing, and the threat of skepticism. *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*, 204.

Nado, J. (2016). The intuition deniers. *Philosophical Studies*, *173*(3), 781-800.

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science, 6*(4), 155-156.

Pachella, R. G. (1973). *The interpretation of reaction time in information processing research*. Michigan University Ann Arbor Human Performance Center. (No. TR-45)

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.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163-177.

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.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653-660.

Radford, C. (1966). Knowledge: By Examples. *Analysis, 27*(1), 1-11.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-430.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347-356.

Rose, D., & Schaffer, J. (2013). Knowledge entails dispositional belief. *Philosophical Studies*, *166*(1), 19-50.

Russell, B. (1948). *Human knowledge: Its scope and its limits*. London: George Allen & Unwin.

Sartwell, C. (1992). Why knowledge is merely true belief. *The Journal of Philosophy*, *89*(4), 167-180.

Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta psychologica, 27*, 143-153.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non‐philosophers. *Mind & Language*, *27*(2), 135-153.

Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, *51*(3), 416-446.

Stich, S. P., and Machery, E. (Forthcoming). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology*.

Strevens, M. (2019). *Thinking off your feet: How empirical psychology vindicates armchair philosophy*. Camrbidge, MA: Harvard University Press.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research, 76*(1), 138-155.

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), 1275-1289.

Turri, J. (2014). Knowledge and suberogatory assertion. *Philosophical Studies*, *167*(3), 557-567.

Weinberg, J. M., & Alexander, J. (2014). Intuitions through Thick and Thin. *Intuitions*, 187.

Weinberg, J. M., Alexander, J., Gonnerman, C., & Reuter, S. (2012). Restrictionism and reflection: Challenge deflected, or simply redirected? *The Monist, 95*(2), 200-222.

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters?. *Philosophical Psychology*, *23*(3), 331-355.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica, 41*(1), 67-85.

Williams, B. (1973). Deciding to believe. In Bernard Williams (ed.), *Problems of the Self*. Cambridge University Press 136-51.

Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.