Synthese (2023) 201:217
https://doi.org/10.1007/s11229-023-04182-7

ORIGINAL RESEARCH

Arguing about thought experiments

Joachim Horvath[2] · Alex Wiegmann[1]

Received: 28 September 2022 / Accepted: 3 May 2023 / Published online: 12 June 2023

Abstract
We investigate the impact of informal arguments on judgments about thought experiment cases in light of Deutsch and Cappelen's mischaracterization view, which claims that philosophers' case judgments are primarily based on arguments and not intuitions. If arguments had no influence on case judgments, this would seriously challenge whether they are, or should be, based on arguments at all—and not on other cognitive sources instead, such as intuition. In Experiment 1, we replicated Wysocki's (Rev Philos Psychol 8(2):477–499, 2017) pioneering study on a Gettier-style case, and we confirmed that the informal arguments used by him had no significant effect. However, we also included an improved argument for ascribing knowledge, which did have a significant effect even in Wysocki's original design. We therefore followed up with Experiment 2 on three Gettier-style cases, where we used a more natural dialogical format for presenting both case descriptions and informal arguments. Overall, we found a clear impact of prima facie good pro and con arguments on case judgments. The issue of argument impact is thus no obstacle to arguing about thought experiments.

Keywords Thought experiments · Arguments · Mischaracterization view · Argument efficacy · Case judgments · Gettier-style cases · Experimental philosophy

1 Introduction

For a long time, it has been a metaphilosophical commonplace that judgments about thought experiments are intuition-based, or at least not based on explicit reflection and arguments (see, e.g., Booth & Rowbottom, 2014; Pust, 2017). This consensus has been shaken by the work of Max Deutsch (e.g., 2010, 2015) and Herman Cappelen (e.g., 2012, 2014a), who, in a number of case studies, bring out that, on the

✉ Joachim Horvath
joachim.horvath@rub.de

[1] Emmy Noether Group EXTRA, Institute for Philosophy II, Ruhr University Bochum, Bochum, Germany

[2] Emmy Noether Group EXTRA, Institute for Philosophy II, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

one hand, explicit appeals to intuition are rare in seminal presentations of philosophical thought experiments, but that, on the other hand, philosophers typically argue for their judgments about hypothetical cases. As a descriptive claim, this leads to the charge that the common intuition-based view about the method of (hypothetical) cases is largely a mischaracterization of actual philosophical practice. According to this *mischaracterization view*, metaphilosophers should therefore acknowledge that judgments about thought experiments are mostly just as argumentation-based as other non-trivial philosophical claims (see also Horvath, 2022).

One striking consequence of this view would be that the (in)famous experimental philosophy challenge to the method of cases (see, e.g., Alexander, 2012; Sytsma & Livengood, 2016) largely comes to nothing. For, the challenge crucially rests on experimental findings which suggest that intuitive judgments about thought experiments vary with factors that are irrelevant to their truth, such as *order of presentation* (see, e.g., Wiegmann et al., 2012, 2020) or well-known *cognitive biases* (see, e.g., Horvath & Wiegmann, 2022). From this, proponents of the program of *negative experimental philosophy* conclude that intuitive judgments about philosophical cases are unreliable (see, e.g., Machery, 2017) or at least not epistemically trustworthy (see, e.g., Alexander & Weinberg, 2014), and that philosophers should thus restrict or even abandon their practice of relying on intuitive case judgments as philosophical evidence. The mischaracterization view has an easy reply to this troublesome challenge: given that philosophers typically do not rely on intuitive case judgments as evidence, but instead argue for their case judgments, they can largely ignore the challenge and more or less continue with their long-standing practice of thought experimentation as it is (see, e.g., Cappelen, 2014a; Deutsch, 2015; Horvath, 2022).

In all fairness, it should be mentioned that the first wave of reactions to the mischaracterization view has been highly critical and unsympathetic (see, e.g., Chalmers, 2014; Devitt, 2015; Nado, 2016; Weinberg, 2014), although extant objections to Deutsch and Cappelen's case for the mischaracterization view turn out to be inconclusive at best (see, e.g., Cappelen, 2014a, 2014b; Deutsch, 2016, 2017; Horvath, 2022), and seem primarily aimed at defending the widely accepted "dogma" of a purely intuition-based method of cases "come what may" (but see Landes, 2023). However, even staunch critics of the mischaracterization view should appreciate that philosophers sometimes *do* argue for their case judgments, and that these arguments—insofar as they are good—surely have *some evidential weight* (even if intuitions may play an evidential role as well). Moreover, given that philosophers *can* argue for their judgments about hypothetical cases, this raises the natural question whether they *should* argue for them. Both the *descriptive fact* that philosophers do argue for their case judgments, and the *normative question* whether they should, give rise to a host of interesting metaphilosophical issues that are very much underexplored. For example, one might ask what *kinds* of arguments these arguments for case judgments really are, given that the relevant cases are purely hypothetical, and thus judgments about them cannot be true or false in any straightforward sense (see, e.g., Deutsch, 2016; Horvath, 2015, 2023a). One might further ask how *argumentation*—understood as a key philosophical method—plays out in the peculiar arena of arguing about thought experiments (cf. Horvath, 2022). Finally, one might also consider the *normative challenge* that philosophers—irrespective of what they have been doing so far—should primarily argue for their judgments about thought experiment cases because this approach would be methodologically more fruitful and transparent than merely appealing to intuition (cf. Horvath, 2023b).

In this paper, we want to focus on another important question about arguments for case judgments, namely, whether people in general and philosophers in particular are in fact *moved* by such arguments. For, if they are not, one might first raise the *descriptive objection* that the cognitive basis of judgments about hypothetical cases is likely something other than arguments, with intuition being the most plausible candidate (see also Nado, 2016). Second, one might raise the *normative objection* that if arguments for case judgments are psychologically inert, we should set them aside as, for example, mere post hoc rationalizations and instead base our case verdicts on something that has more *psychological efficacy*—with intuition being the most plausible candidate again.

The available evidence concerning whether people are actually moved by arguments for case judgments is still quite limited, however. The only study that directly addresses this issue against the background of the mischaracterization view is Wysocki (2017), who investigates whether explicitly given informal arguments for and against knowledge-ascriptions in a Gettier-style case of the stopped-clock type (cf. Russell, 1948)[1] have a significant influence on people's case judgments.[2] Wysocki ran his study in three different languages (English, Polish, Spanish) and four different countries (USA, India, Poland, Spain). The upshot is that neither an explicit informal argument for the canonical case judgment about Gettier-style cases—that the protagonist lacks knowledge—nor against that judgment had any significant effect on people's judgments about the case.

[1] Apart from Russell (1948), Gettier's (1963) famous presentation of such cases was also predated by similar cases in European medieval philosophy (see, e.g., Hilpinen, 2017) and classical Indian philosophy (see, e.g., Matilal, 1991). The reason why they are commonly referred to as 'Gettier-style cases' or simply (but somewhat misleadingly) as 'Gettier cases' is that Gettier's (1963) presentation had by far the largest impact on contemporary theory of knowledge, and might easily be the single most influential paper in epistemology.

[2] Focusing on the psychological efficacy of arguments about thought experiment cases is a relatively novel topic in experimental philosophy. Apart from Wysocki (2017), there are just a few studies from various disciplinary backgrounds that test the influence of arguments, reasons, or reflection on people's judgments about moral cases and other philosophical dilemmas, with quite mixed results (Ervin & Corral, 2022; Herec et al., 2022; Kneer et al., 2021; Paxton et al., 2012; Stanley et al., 2018). An interesting case is the computational NLP-based study by Na & DeDeo (2022), who investigate the efficacy of a number of broad argument-types in contexts of informal argumentation—from the classic deductive and inductive arguments to causal arguments and arguments from personal experience. What they found is that non-classic types, such as causal or example-based arguments, tend to be more effective than the classic deductive or inductive arguments. In contrast to this study, our own experiments focus on differences in argument efficacy between pro and con arguments of the same general type. Another noteworthy study is Hansen & Chemla (2015), who investigate the effect of argumentative glosses on classic philosophical thought experiments (with somewhat mixed results), which seems fairly closely related to the issue of argument impact—as is the more recent study "Socratic Questionnaires" by Hansen et al. (forthcoming). Other previous experimental philosophy studies on argumentation address a wide range of different issues, e.g., the pitfalls of metaphors in arguments (e.g., Ervas et al., 2018), the impact of linguistic bias on the method of cases (e.g., Fischer & Engelhardt, 2020), or issues about the content and rationality of reasoning with conditionals (e.g., Pfeifer & Tulkki, 2017).

However, Wysocki's study has several methodological shortcomings that lower the evidential value of his results (see also Horvath, 2022, sec. 3.6). For example, he uses the problematic answer-format "really knows" versus "only believes" (cf. Cullen, 2010), his presented argument for ascribing knowledge is clearly not a good one (see below), and the case description itself is also not as precise as it could be (see below). Finally, drawing substantive conclusions from just a single study on a single thought experiment would be premature even if the study in question were absolutely flawless.

The main aim of this paper will therefore be to improve and extend the empirical basis for assessing the psychological impact of informal arguments on judgments about hypothetical cases. To this end, we will **first** report a *replication* of Wysocki's original study in English, because the robustness of findings of this kind cannot simply be taken for granted in light of the ongoing replication crisis in psychology and other disciplines (see, e.g., Open Science Collaboration, 2015). **Second**, we will report—on the basis of one additional condition—whether merely replacing Wysocki's bad argument for ascribing knowledge with a prima facie *good argument* already makes a significant difference for argument-impact. And **third**, we will present the results of a *statistically high-powered online experiment* that tests the impact of informal arguments on case judgments with three different types of Gettier-style cases: a version of Wysocki's stopped-clock case, a scenario of the fake-barn type, and a scenario that is structurally more similar to Gettier's (1963) own cases. In this experiment, we also implemented several improvements over Wysocki's original experimental design. For example, we replaced his problematic answer-format in terms of "really knows" vs. "only believes" with a simple choice between the relevant knowledge-ascription and its negation (in randomized order), and we included comprehension checks for each case. Most importantly, however, we also used a more natural presentation for both the case description and the relevant pro and con arguments in the form of a *short dialogue* between two friends.

This more natural presentation format also connects our experiment with a longstanding theme of *ordinary language philosophy*, namely, that the primary loci of linguistic understanding and interpretation are the ordinary, real-life contexts in which the linguistic items in question are used by competent speakers of the relevant language (see, e.g., Baz, 2012; Hansen, 2020; Travis, 1989). While our experiment might still be "unnatural" in other respects—maybe in part because of being a controlled experiment as such—it does take a significant step towards investigating argumentation in its most natural "linguistic habitat", the *dialogue*, and thereby helps to increase the "ecological validity" of its results.

2 Experiment 1: Replication and extension of Wysocki (2017)

Our first experiment serves two purposes. First, it is a high-powered replication of the English version of the experiment reported in Wysocki (2017), which tests whether knowledge attributions in Wysocki's *Tower Clock* case can be influenced by informal arguments for and against the claim that the agent in the case knows that it is 12 pm. In the original experiment by Wysocki, arguments did not significantly affect knowledge attributions (for neither of the samples recruited in four different countries). Second, we test the hypothesis that an arguably improved version of Wysocki's argument for attributing knowledge would increase the level of knowledge ascriptions in comparison to the two conditions in which Wysocki's argument against ascribing knowledge and no argument at all are presented (with results of the latter two conditions taken from our replication study). Study materials, data, and analysis code for the present and also the following experiment are publicly available at: https://osf.io/y8hck

2.1 Participants

The participants of this and all other experiments reported in this paper were recruited on *Prolific* (https://www.prolific.co/), completed an online survey in *Unipark* (https://www.unipark.com/en/), and were required to be native speakers of English. As preregistered (https://osf.io/mtgjp), the experiment was run until valid[3] responses of 740 participants (185 in each condition) were collected, which results in 90% power for detecting a 15% difference (65% vs. 50%; one-tailed z-proportion test at a 0.05 significance level) between two conditions. The average participant age was 39 years, and 47% reported to be male, 51% female, 1% non-binary/other, with 0% choosing the remaining option "prefer not to say".[4] Participants received a compensation of £0.25 for estimated 2 minutes of their time (£7.5/h), in accordance with the fair pay guidelines of Prolific.

2.2 Design, procedure, and materials

Participants were randomly assigned to one of four conditions (*Baseline*, *Pro*, *Con, Improved Pro*). In all four conditions, participants were presented with the *Tower Clock* case, worded exactly as in Wysocki (2017), which implements the basic pattern of Russell's (1948) famous stopped-clock case that contemporary epistemologists typically treat as a kind of Gettier-style case (see, e.g., Truncellito, 2007):

> John walks through the market square, and wonders what time it is. He looks at the clock on the town hall tower, and sees that the clock shows 12 pm. And indeed, it is 12 pm. However, John doesn't realize that the clock stopped exactly twelve hours ago, and this is why it shows the correct time.

In *Pro*, *Con*, and *Improved Pro* (see below), participants were in addition presented with an argument for or against the claim that John knows that it is 12 pm, which participants had to consider for at least 25 s before they were able to proceed to the test question (as in Wysocki's original experiment). The *Pro* and *Con* conditions were worded exactly as in the original experiment in order to replicate the English version of Wysocki's (2017) study, while the *Improved Pro* condition used an improved version of Wysocki's original pro condition, in which he presented a prima facie bad argument for attributing knowledge. The two main problems with the argument in *Pro* (see below)

[3] 980 participants started the experiment and 140 were excluded from the analysis for not completing the experiment, failing to answer the check-question correctly, or not being among the first 185 valid responses in their condition.

[4] Due to rounding, the percentages do not exactly add up to 100%.

are, first, that true belief is clearly not sufficient for knowledge (contrary to what is suggested by the first sentence of *Pro*), and, second, that the fact that it might be useful for John to believe that it is 12 pm is irrelevant to the question whether John indeed knows that it is 12 pm (contrary to what is suggested by the remainder of *Pro*). In contrast, the argument that is presented in favor of John's knowing that it is 12 pm in *Improved Pro* (see below) is at least prima facie convincing, even if it fails to be ultima facie compelling, due to Gettier-style counterexamples (this is not a problem, however, because we generally do not require that philosophical arguments must be ultima facie compelling in order to be presentable and worthy of discussion in the first place, which would be an excruciatingly high standard). More specifically, the improved pro argument suggests that true belief that is based on strong evidence is good enough for knowledge (in the first sentence of *Improved Pro*), and it cites two features of the case (in the remainder of *Improved Pro*) that plausibly render John's evidence strong, namely, the excellent visual conditions for reading the tower clock, and the fact that tower clocks are generally very reliable.

Pro:

> John knows that it's 12 pm, because he believes that it's 12 pm, and indeed it's 12 pm. Moreover, it might be useful for him that he thinks it's 12 pm. For example, imagine that John wants to catch a train at 12:10 pm. He looks at the clock, thinks that it's 12 pm, hurries to the railway station, and catches the train.

Con:

> John doesn't know that it's 12 pm, because it was an accident that he checked the time exactly when the clock was showing the correct time. For example, if John looked at the clock 5 minutes later, he would be mistaken about the time.

Improved Pro:

> John knows that it's 12 pm, because it is indeed 12 pm and his belief that it's 12 pm is based on strong evidence. For example, John sees the clock on the town hall tower in bright daylight, and town-hall clocks are usually very reliable.

The *test question* and *binary response options* were as follows:

Does John really know that it's 12 pm or does he only believe it?

- Really knows …
- Only believes …

To make sure that participants paid sufficient attention to their task and understood crucial features of the case, the following *two check questions* were included:

Did the clock show the correct time when John looked at it?

- yes
- no

Did the clock stop ten hours before John looked at it?

- yes
- no

On the final page, we asked *standard demographic questions* and provided an *optional text field* for participants in case they encountered any problems with the study (e.g., technical issues, typos, unclear questions/formulations, etc.).

2.3 Results

We only included the data of those participants in the analysis who answered both check questions correctly. The results are summarized in Fig. 1. Knowledge ascriptions were lowest in *Con*, 11.35% [Ci 95%; 7.17%; 16.83%], followed by *Baseline*, 16.76% [11.68; 22.93], *Pro*, 18.38% [23.08; 24.72] and *Improved Pro*, 25.41% [19.30; 32.31].

Let us now consider our preregistered tests. As in Wysocki (2017), and as preregistered by us, we found no significant effect between *Con*, *Baseline*, and *Pro*, $\chi^2_{2, N=555} = 3.825$, $p = 0.148$. Furthermore, we predicted that knowledge ascriptions in the new *Improved Pro* condition would be significantly higher than in both *Con* and *Baseline*. This prediction was confirmed. Significantly more participants ascribed knowledge in *Improved Pro* than in *Con*, $\chi^2_{1, N=370} = 12.18$, $p < 0.001$, Cohen's $h = 0.37$, as well as in *Baseline*, $\chi^2_{1, N=370} = 4.159$, $p = 0.021$, $h = 0.21$.

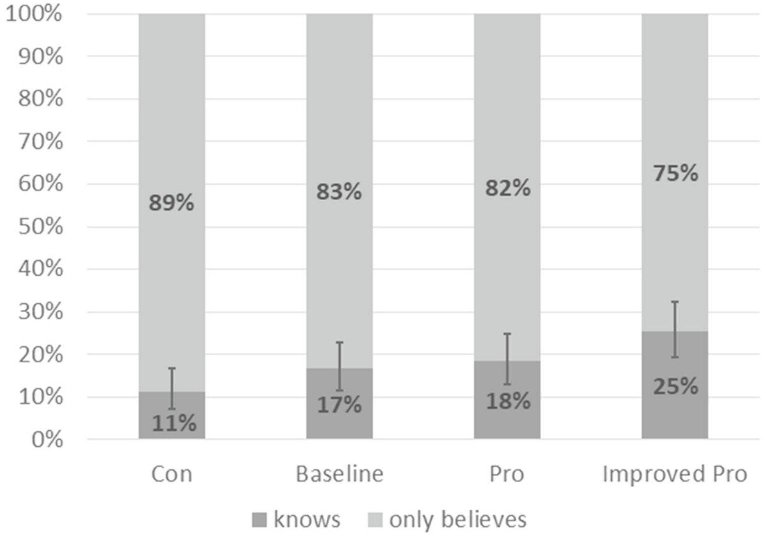

Fig. 1 Percentages of participants ascribing either knowledge or mere belief to the agent in Wysocki's *Tower Clock* case. Error bars represent 95% confidence intervals

2.4 Discussion

We replicated Wysocki's original finding that knowledge ascriptions in the *Tower Clock* case were not affected by the presented informal arguments for and against attributing knowledge. However, using a new and improved argument in favor of ascribing knowledge in *Improved Pro* did significantly increase knowledge attributions as predicted. This initial finding brings us to our second experiment, in which we test—for a broader range of Gettier-style scenarios—whether knowledge ascriptions in thought experiment cases are influenced by prima facie good informal arguments.

3 Experiment 2: Three Gettier-style cases in a dialogical design

In this experiment, we tested for three scenarios (*Dollar*, *Clock*, *Hospital*) whether prima facie good informal arguments for or against the claim that an agent knows a certain proposition can influence participants' knowledge ascriptions. One of the scenarios that we used, *Clock*, is a slightly improved and more accurate version of Wysocki's *Tower Clock* case. The other two scenarios are adapted from other Gettier-style cases that were previously tested in experimental philosophy.

The *Dollar* case, which is a variation on the fake-barn theme (cf. Goldman, 1976), was already used in a study by Horvath and Wiegmann (2016). Interestingly, and contrary to the "textbook consensus" in epistemology, what they found is that laypeople and even the majority of expert epistemologists tend to ascribe knowledge in such cases. Our experiment confirms this surprising finding even more strongly than in the original study (see below), which may in part be due to our binary response format versus the originally used Likert item. This issue—philosophically interesting as it may be—is largely orthogonal to our main question in the present study, however, which is to investigate the influence of arguments on case judgments, irrespective of where the majority in the relevant cases lies.

The third case that we used, *Hospital*, is more similar to Gettier's (1963) own cases insofar as it involves a justified false assumption on behalf of the protagonist of the case. In previous studies (e.g., Machery et al., 2017), this case turned out to robustly replicate Gettier's canonical judgment that the epistemic subject fails to know the relevant proposition in such cases (even across several different cultures). In our adaptation of the hospital case, however, we strengthened the justification for one of the (false) premise beliefs of the protagonist of the case (see footnote 7). For, first, it is an important aspect of Gettier-style cases that the justification-condition is clearly satisfied—given their philosophical purpose of refuting the *justified*-true-belief account of knowledge. Second, if the justification of the protagonist is relatively weak, this might bias the case against ascribing knowledge to the protagonist, and thus favor the canonical judgment about Gettier-style cases. Third, and most importantly, such a biased presentation would make it less plausible that the case can be reasonably judged as both a case of knowledge and a case of non-knowledge (by different people or at different times)—which is crucial for our investigation of the effect of pro versus con arguments for attributing knowledge.

Therefore, one reason why we selected the two additional cases *Dollar* and *Hospital* is that they contain features that prima facie speak both in favor of and against attributing knowledge to the protagonist. In the *Pro* conditions, the features that favor ascribing knowledge are argumentatively exploited by emphasizing the *strong evidence* that the protagonist has for his target true belief. In the *Con* conditions, on the other hand, the argumentative focus is on the protagonist's *luck* in ending up with a true belief. Such an appeal to epistemic luck is still the most commonly presented reason against ascribing knowledge in Gettier-style cases even in the contemporary epistemological literature (see, e.g., Engel Jr., 2015; Pritchard, 2005). In this way, the tested cases allow for at least a prima facie reasonable disagreement—and thus for related argumentation—about whether the protagonist knows the proposition in question or not. For, the natural expectation that argumentation can make a reason- and argument-based difference to people's case judgments is only warranted if people's case judgments can in fact be reasonably pulled in different directions (see also Machery, 2017). Given that, as (meta)philosophers, we are not interested in arguments as mere rhetorical devices for manipulating people's judgments, this is in fact one of the key features of the tested cases.

Finally, we also changed and improved how the informal arguments in the *Pro* and *Con* conditions are presented to the participants of the experiment. Wysocki (2017) simply added the arguments immediately after the case description in the "voice of the experimenter", so to speak. This mode of argument-presentation has two potential disadvantages. First, given that an experimental questionnaire is already *not a very natural linguistic context* (see, e.g., Hansen, 2020), presenting an argument in the anonymous and often hard to interpret experimenter's voice even increases this tendency; second, and relatedly, this might induce *demand characteristics* in the participants (see, e.g., Orne, 1962), such that they would think about the test question more in terms of what the experimenter expects to hear from them than in terms of what they themselves think about the case. In order to avoid, or at least ameliorate, these potential problems of unnaturalness and demand characteristics, we presented both the case descriptions and the informal arguments in our *Pro* and *Con* conditions in a *short dialogue* between two friends or good colleagues (i.e., two people who are on friendly terms with each other), who we uniformly named 'Tom' and 'Jill' across all conditions. In this way, participants' attention is directed away from the experimenter and the artificiality of the experimental context, and towards the more natural situation of a casual exchange between two friends or colleagues. Moreover, this mode of argument-presentation is also nicely consonant with one of the major approaches in *argumentation theory*, which holds that various types of *dialogues* are the proper linguistic format for studying informal argumentation (see, e.g., Walton, 2008).

3.1 Participants

As preregistered (https://osf.io/8jfzn), the experiment was run until valid[5] responses of enough participants were collected (185 in each condition, 1665 overall), resulting in 90% power for detecting a 15%-difference (65% vs. 50%; one-tailed z-proportion test at a 0.05 significance level) between two conditions. The average age of participants was 39 years, and 45% reported to be male, 54% female, 1% non-binary/other, with 1 participant choosing the remaining option "prefer not to say". Participants received £0.25 for estimated 2 minutes of their time (£7.5/h).

3.2 Design, procedure, and materials

Participants were randomly assigned to one of nine conditions in a 3 (argument: *Baseline*, *Pro*, and *Con*) * 3 (scenario: *Dollar*, *Clock*, and *Hospital*) between-subjects design.

The scenarios were worded as follows. *Dollar*[6]:

> Tom: "Hi Jill, what do you think about the following case: 'A waiter was recently hired by a restaurant in a remote part of the country where, without anyone's knowledge, most of the circulating dollar notes are fake. The restaurant manager

[5] 2994 participants started the experiment, of which 1329 were excluded. 662 of these exclusions were due to a change in the *Dollar* vignette that we made in reaction to an open feedback comment from a participant (we excluded everyone who saw the initial wording and run all three *Dollar* conditions again with the changed wording; see footnote 5 for details about the wording). The data of the other 667 participants were excluded from the analysis because they did not complete the experiment, failed to answer the check-questions correctly, or were not part of the first 185 valid responses in their respective condition.

[6] This is the final and revised version of the *Dollar* vignette. The initial wording of the *Dollar* scenario in the *Baseline* condition was as follows:

> *Tom: "Hi Jill, what do you think about the following case: 'A waiter was recently hired by a restaurant in a remote part of the country where, without anyone's knowledge, most of the circulating dollar notes are fake. The restaurant manager still owes the waiter ten dollars for an extra hour that he worked. She takes a ten-dollar note from the cash register and hands it over to the waiter, who thanks her and puts the ten dollars in his wallet. In fact, this particular ten-dollar note happens to be one of the few genuine dollar notes in the area.' So, Jill, do you think that the waiter in this case knows that the restaurant manager has given him a ten-dollar note?".*
> *Jill: "Well, Tom, I'm in a bit of a hurry at the moment – but I will tell you over lunch what I think about the case.".*

We are now interested in your own opinion about the case. Please choose the option that better describes Tom's case:

- *The waiter knows that the restaurant manager has given him a ten-dollar note.*
- *The waiter does not know that the restaurant manager has given him a ten-dollar note.*

We then slightly modified the wording in all *Dollar* conditions because one participant noted that they chose the first option because it seemed to them that a fake ten-dollar note is still a ten-dollar note. While this understanding is surely debatable, we nevertheless decided to make the case less ambiguous in this respect and changed the potentially misleading occurrences of "ten-dollar-note" to "ten dollars". As it happened, the change had virtually no effect on knowledge ascriptions in the *Dollar* conditions (see analyses file at https://osf.io/y8hck). Moreover, even the unmodified *Dollar* case produced very similar results as other cases of the fake-barn-type in previous research (cf. Horvath & Wiegmann, 2016). Therefore, whatever the precise experimental impact of this issue may be, it is not likely to be a crucial factor. Thanks also to Michelle Liu for discussion.

> still owes the waiter ten dollars for an extra hour that he worked. She takes a ten-dollar note from the cash register and hands it over to the waiter, who thanks her and puts the ten dollars in his wallet. In fact, this particular ten-dollar note happens to be one of the few genuine dollar notes in the area.' So, Jill, do you think that the waiter in this case knows that the restaurant manager has given him ten dollars?"

In the *Baseline* version of *Dollar* (and in all other *Baseline* conditions as well), Tom gets the following neutral response from Jill:

Jill: "Well, Tom, I'm in a bit of a hurry at the moment – but I will tell you over lunch what I think about the case."

The argument for attributing knowledge in the *Pro* condition was worded as follows:

Jill: "Well, Tom, here's what I think about the case. The waiter's belief that the restaurant manager has given him ten dollars is in fact true and based on strong evidence. After all, the waiter clearly sees the ten-dollar note when the restaurant manager hands it over to him, and a ten-dollar note is pretty easy to recognize. Therefore, the waiter knows that the restaurant manager has given him ten dollars, because if you believe something that is true on the basis of strong evidence, then you also know it."[7]

Finally, the argument in the *Con* condition of the *Dollar* case was the following:

Jill: "Well, Tom, here's what I think about the case. Although the waiter's belief that the restaurant manager has given him ten dollars is in fact true, the waiter only hit on the truth by luck. For, the restaurant manager could have easily given him one of the many fake ten-dollar notes that circulate in the area, and the waiter wouldn't have noticed the difference. Therefore, the waiter does not know that the restaurant manager has given him ten dollars, because one cannot know something if one only gets it right by luck."

For each scenario, we included two check questions to ensure data quality.[8] In the *Dollar* case, the two questions were:

[7] In this and the following argument-conditions, one might worry that the evaluative language that is used in describing the arguments (e.g., "strong evidence", "easy to recognize", "only by luck") makes them seem artificially strong and hard to resist. But first, the very point of our study is to investigate whether arguments have an impact on case judgments *at all*, and so the most natural starting point is to investigate both pro and con arguments *in their strongest form*. For, if people are not even moved by the strongest versions of these arguments, then they are probably not moved by weaker versions either. Moreover, giving reasons and arguments for a judgment is a *constitutively normative practice*, and so trying to keep evaluative language out of presenting one's reasons or evidence for a certain judgment is both artificial and ill-motivated. It would only be a problem if the evaluations in question were not adequate to the situation described, which might weaken the normative force of the argumentation in question. However, this is not the case in our argument-conditions. For example, one plausibly has "strong evidence" (by ordinary standards) if one has seen something that is easy to recognize from a close distance, such as a ten-dollar note. Similar points apply to other evaluative features of our vignettes (thanks to Eugen Fischer and Kevin Reuter for discussion on this point).

[8] See the online materials at https://osf.io/y8hck for the check questions that we used in the *Clock* and *Hospital* case.

Was the dollar note that the waiter received from the restaurant manager a genuine note?

- yes
- no

Did the waiter receive a twenty-dollar note from the restaurant manager?

- yes
- no

The *target question* in all *Dollar* conditions was the following:

We are now interested in your own opinion about the case. Please choose the option that better describes Tom's case:

- The waiter knows that the restaurant manager has given him ten dollars.
- The waiter does not know that the restaurant manager has given him ten dollars.

Clock scenario:

Tom: "Hi Jill, what do you think about the following case: 'John walks through the market square, and wonders what time it is. He looks at the clock on the town hall tower, sees that the clock shows 11, and so believes that it's 11 am, since he looks at the clock in bright daylight. And indeed, it is 11 am. However, John doesn't realize that the clock stopped exactly twenty-four hours ago, and this is why it shows the correct time.' So, Jill, do you think that John in this case knows that it's 11 am?"[9]

Argument in the *Pro* condition of *Clock*:

Jill: "Well, Tom, here's what I think about the case. John's belief that it's 11 am is in fact true and based on strong evidence. After all, John sees the clock on the town hall tower in bright daylight, and town-hall clocks are usually very reliable. Therefore, John knows that that it's 11 am, because if you believe something that is true on the basis of strong evidence, then you also know it."[10]

[9] *Clock* is a slightly improved version of Wysocki's *Tower Clock*, which was prompted by a participants' comment in the open feedback text field at the end of the questionnaire (see above). The worry is that a tower clock that stopped only 12 hours ago, as in *Tower Clock*, would actually not show the correct time, 12 pm, but rather the incorrect time, 12 am. In addition, we changed the hours from 12 to 11, because the pm/am-convention is somewhat confusing in the case of 12 pm/am.

[10] One might worry here that Jill's argumentation adds new details that change crucial epistemic features of the case described by Tom, and that Jill therefore does not really argue about the case as stated by Tom, but rather about a somewhat different case of her own making (thanks to an anonymous reviewer for raising this issue). For example, in the *Pro* condition of *Clock*, Jill backs up her claim that John's belief is "based on strong evidence" by pointing out that "John sees the clock on the town hall tower in bright daylight, and town-hall clocks are usually very reliable". However, Tom's case description neither includes that John looks at the clock "in bright daylight", nor that town-hall clocks are "usually very reliable". So, what is going on here? Nothing extraordinary, we believe. For, Jill arguably applies common pragmatic mechanisms in interpreting Tom's case description, such as broadly Gricean principles and reliance on the common ground of the conversation (likewise in the *Pro* and *Con* conditions of *Dollar* and *Hospital*). When she assumes, e.g., that John looks at the clock "in bright daylight", she merely relies on the common knowledge that there is usually bright daylight at 11 am, which can be taken to be part of the common ground of any ordinary

Argument in the *Con* condition of *Clock*:

Jill: “Well, Tom, here’s what I think about the case. Although John’s belief that it’s 11 am is in fact true, he only hit on the truth by luck. For, John could have easily looked at the clock at some other moment when it didn’t show the correct time, and he wouldn’t have noticed the difference. Therefore, John does not know that it’s 11 am, because one cannot know something if one only gets it right by luck.”

Target question of *Clock*:

We are now interested in your own opinion about the case. Please choose the option that better describes Tom’s case:

- John knows that it is 11 am.
- John does not know that it is 11 am.

Hospital scenario:

Tom: “Hi Jill, what do you think about the following case: ‘Paul Bigelmair[11] is worried because it is 10 pm and his wife Mary is not home from work yet. Usually she is home by 6 pm. He tries her cell phone but just keeps getting her voicemail. Starting to worry that something might have happened to her, he decides to call some local hospitals to ask whether any patient by the name of “Mary Bigelmair” was admitted that evening. At the University Hospital, the person who answers his call confirms that someone by that name was admitted with major but not life-threatening injuries following a car crash. Paul grabs his coat and rushes out to drive to University Hospital. As it turns out, the patient at University Hospital is not Paul’s wife, but another woman with the same name. In fact, Paul’s wife had a heart attack as she was leaving work, and is actually receiving treatment in Metropolitan Hospital, a few miles away.’ So, Jill, do you think that when Paul rushes out to drive to University Hospital, he knows that his wife was hospitalized?”

Argument in *Pro* condition of *Hospital*:

Footnote 10 continued
conversation, and she reasonably expects that Tom, for Gricean reasons, would have explicitly said so if he had intended the case to be understood differently in this respect. Likewise, Jill can also regard it as being in the common ground of her conversation with Tom that town-hall clocks are usually very reliable, and that Tom would have said so if he had wanted her to understand the case differently. So, Jill plausibly does not add any new or unintended details to Tom’s description of the *Clock* case, but simply relies on features of the case that are pragmatically implicated, based on the common ground of the conversation (for a pragmatic account of thought experiments along those lines, see Saint-Germier, 2021). Of course, we—as experimenters—also need to rely on our participants’ shared understanding of the dialogue between Tom and Jill. This, however, is a common feature of most experimental studies that work with verbal prompts. And while things surely can go wrong in the pragmatics of such experiments (see, e.g., Wiegmann, 2022), the assumption that something will go wrong, without offering any specific evidence or hypothesis to this effect, would amount to an unduly skeptical attitude towards linguistic communication.

[11] In comparison to the original study by Machery et al. (2017), we changed the last name of Paul and his wife Mary from the ubiquitous English last name “Smith” to the more unusual “Bigelmair” (a name of German-Swabian descent). For, in combination with the fact that “Mary” is one of the most common first names of English, the fact that University Hospital confirms to Paul that someone with the name “Mary Smith” was admitted that evening is not clearly strong evidence in favor of Paul’s belief that his wife was hospitalized—given the countless other Mary Smiths out there, one of which might also have been admitted to the presumably fairly large University Hospital on the same evening.

Jill: "Well, Tom, here's what I think about the case. Paul's belief that his wife was hospitalized is in fact true and based on strong evidence. After all, hospitals are highly trustworthy institutions, and so University Hospital's explicit confirmation that a Mary Bigelmair was admitted that evening is a clear reason for Paul to believe that his wife was hospitalized. Therefore, Paul knows that his wife was hospitalized when he rushes out, because if you believe something that is true on the basis of strong evidence, then you also know it."

Argument in *Con* condition of *Hospital*:

Jill: "Well, Tom, here's what I think about the case. Although Paul's belief that his wife was hospitalized is in fact true, he only hit on the truth by luck. For, given that the Mary Bigelmair at University Hospital is not Paul's wife, he would have still believed that his wife was hospitalized even if she hadn't been hospitalized at all, for example, because she died from her heart attack at work. Therefore, Paul does not know that his wife was hospitalized when he rushes out to drive to University Hospital, because one cannot know something if one only gets it right by luck."

Target question of *Hospital*:

> We are now interested in your own opinion about the case. Please choose the option that better describes Tom's case:
>
> - When Paul Bigelmair rushes out to drive to University Hospital, he knows that his wife was hospitalized.
> - When Paul Bigelmair rushes out to drive to University Hospital, he does not know that his wife was hospitalized.

In all nine conditions, participants could only submit their responses after a forced waiting time of 25 s, as in *Experiment 1* and Wysocki's (2017) original study.

3.3 Results

The results are summarized in Figs. 2 and 3. On average and across scenarios, knowledge ascriptions were—as predicted—lowest in *Con*, 43.96% [Ci 95%; 39.79%; 48.21%], followed by Baseline, 53.69% [49.44; 57.90], and *Pro*, 65.77% [61.65; 69.71]. All three preregistered z-proportion tests were significant, and effect sizes ranged from small (*Pro* vs. *Baseline*, and *Baseline* vs. *Con*) to medium (*Pro* vs. *Con*). Detailed statistics can be found in Table 1.

On the level of individual scenarios, seven out of the nine z-proportion tests were significant,[12] with the exception of the comparison of *Baseline* vs. *Con* in the *Clock* and *Hospital* case. Effect sizes ranged from small to large, with most effect sizes being small or medium (for detailed statistics, see again Table 1).

[12] Or five out of nine z-proportion tests if the alpha level is conservatively adjusted for multiple comparisons (0.05 divided by number of tests).

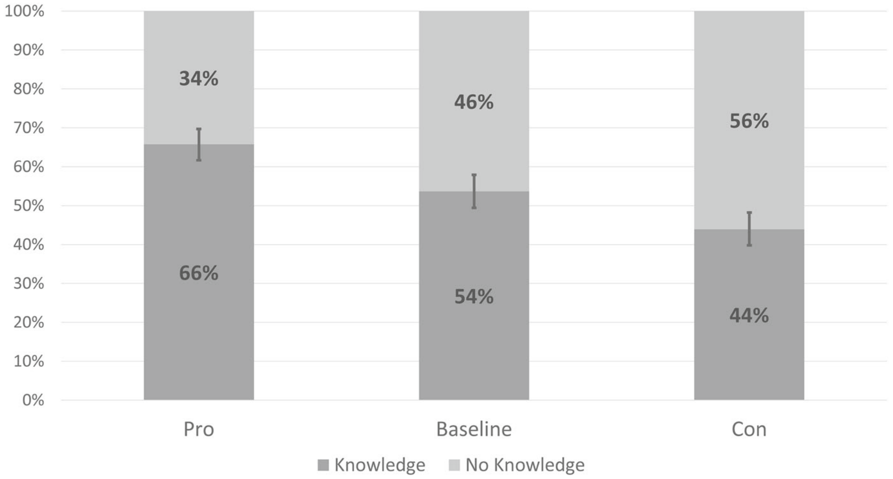

Fig. 2 Percentages of participants ascribing either knowledge or non-knowledge to the agent across all scenarios. Error bars represent 95% confidence intervals

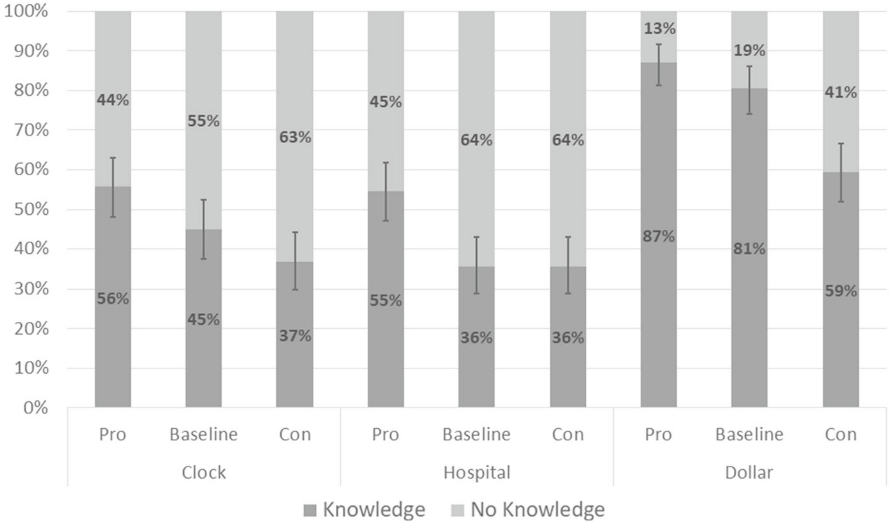

Fig. 3 Percentages of participants ascribing either knowledge or non-knowledge to the agent for all cases and conditions. Error bars represent 95% confidence intervals

3.4 Discussion

Although the effects of the presented informal arguments were mostly not very large and, in a few cases, even non-significant, the overall pattern of results strongly indicates that arguments can have a considerable influence on knowledge ascriptions in hypothetical cases.

Table 1 Results for one-tailed z-proportion tests

Scenario	Comparison	*N*	$X^2{}_1$	P value	Effect size
Dollar	Pro vs Baseline	370	2.865	0.045	0.18
Dollar	Baseline vs Con	370	19.575	< 0.001	0.47
Dollar	Pro vs Con	370	35.871	< 0.001	0.64
Clock	Pro vs Baseline	370	4.324	0.019	0.22
Clock	Baseline vs Con	370	2.517	0.056	0.17
Clock	Pro vs Con	370	13.320	< 0.001	0.38
Hospital	Pro vs Baseline	370	4.324	< 0.001	0.38
Hospital	Baseline vs Con	370	2.517	0.500	0.00
Hospital	Pro vs Con	370	13.370	< 0.001	0.38
Across all	Pro vs Baseline	1110	16.813	< 0.001	0.25
Across all	Baseline vs Con	1110	10.514	< 0.001	0.19
Across all	Pro vs Con	1110	53.265	< 0.001	0.44

P-values are unadjusted, effect size represents Cohen's h. Effect size conventions: 0.2 (small), 0.5 (medium), 0.8 (large)

One might wonder, however, whether these effects are really driven by the presented arguments *as such*, or rather by other features of our presentation apart from their argumentative content, such as making a certain response option pragmatically more salient, or facilitating reading comprehension of a philosophically relevant case feature.[13]

In order to address this worry, let us first clarify that, in the context of this study, we understand an *argument* in the fairly broad sense of a kind of conversational move in an *informal argumentation*, that is, as any (complex) speech act that provides some reason in favor of the relevant target claim or judgment. This seems to be the dominant understanding of arguments in the interdisciplinary field of *argumentation theory* (see, e.g., Eemeren et al., 2004; Walton, 2008), which, in its key respects, is also shared by proponents of the mischaracterization view (see, e.g., Cappelen, 2012; Deutsch, 2015; Horvath, 2022). Important corollaries of this informal understanding of argumentation are, first, that the arguments in question can (and often will) be *non-deductive*, and second, that they can (and frequently will) be *incomplete or enthymematic*, and thus need not resemble anything like the formal or semi-formal arguments in logic textbooks or philosophical papers.

For illustration, consider the argument in the *Con* condition of our improved *Clock* case above. Here, Jill informally appeals to a feature of the case described by Tom—that John merely got it right be luck that it is 11 am—as a reason for judging that John does not know that it is 11 am. At least prima facie, this is also—normatively speaking—a *good reason* for making that judgment (see above), and so Jill thereby presents a *prima facie good informal argument* for this judgment (even if it may be incomplete or enthymematic). And since we found a highly significant difference between this

[13] Thanks to an anonymous reviewer for articulating this potential worry.

Con condition and the opposing *Pro* condition of *Clock*, in which Jill offers a prima facie good informal argument for the judgment that John knows that it is 11 am, our experiment arguably indicates a significant influence of prima facie good informal arguments on participants' knowledge ascriptions.

Now, it could be that, for example, Jill's informal argument in the *Con* condition *also* makes John's lack of knowledge pragmatically salient to participants of the experiment, or that it facilitates reading comprehension of the crucial case feature of accidentally true belief. If so, however, Jill would do these further things *in virtue of* presenting a prima facie good informal argument against ascribing knowledge to John, and so, in a sense, presenting a good informal argument cannot be separated from these potential additional effects, since they are all different aspects of one and the same argumentative speech act.

That being said, our *Experiment 1* does provide some initial evidence that it is really the "rational properties" of the informal arguments presented, i.e., the normative quality of the reasons given by them, that is the driving force behind our results. For, what we found in the replication part of *Experiment 1* is that, although the contrast between the *Pro* and *Con* condition was non-significant, just as in Wysocki's (2017) original experiment, the contrast between our *Improved Pro* condition and the *Con* condition was indeed significant. By design, however, the only difference between Wysocki's *Pro* condition and our *Improved Pro* condition is that the latter presents a *prima facie good informal argument* for the claim that John knows that it is 11 am, while the former presents an informal argument that is even *prima facie bad* (because it merely appeals to the practical usefulness of John's true belief, which is prima facie irrelevant or insufficient for ascribing knowledge). Thus, in this case, it seems very plausible that the observed effect crucially depends on the prima facie quality of the presented informal arguments, given that there is no obvious difference in pragmatic salience-raising or the facilitation of reading comprehension.

Relatedly, we also do not think that our findings are grist to the mill of the experimental philosophy challenge to the method of cases, because this challenge crucially rests on findings that suggest a variation of case judgments with *philosophically irrelevant factors* (see above)—but the variation of case judgments with prima facie good arguments is clearly not irrelevant in this sense. It is true that some prima facie good arguments (i.e., those that are not ultima facie good) may lure us away from the truth about the issue at hand. But this is simply a particular instance of the more general fact that fallible or non-conclusive reasons can sometimes lure us away from the truth even if we respond to them in a perfectly rational way. And just as most epistemologists do not return to an infallibilist, Cartesian conception of reasons because of that fact, it would likewise not be a compelling reason to regard sensitivity to "merely" prima facie good arguments as a philosophically irrelevant factor. So, while we often may not know whether a given prima facie good philosophical argument is also *ultima facie* good or compelling, it would be wrong to conclude that only ultima facie good arguments are philosophically relevant. In fact, this would imply a rather sweeping indictment of our actual practice of philosophical argumentation, because all we can confidently say about most actual instances of philosophical argumentation is whether they prima facie bear on the philosophical claim in question, and are in this sense at least prima facie good arguments—but not whether they are also ultima facie good arguments, which we may only be able to tell at the end of a long and intricate philosophical debate. For this reason, to argue that philosophical judgments should not be influenced by "merely" prima facie good arguments might even be plainly self-refuting.[14]

Finally, the suggestive evidence from *Experiment 1* notwithstanding, we think that it is still an important desideratum for future research to investigate contrasts between informal arguments of different prima facie quality in a more systematic fashion.

4 General discussion

The two main goals of *Experiment 1* were to *replicate* the English version of Wysocki's (2017) study of the impact of informal arguments on thought experiment judgments, and to test an *improvement* over Wysocki's original pro condition. Our high-powered replication was successful insofar as it confirmed Wysocki's finding that his argument conditions had no significant impact on participants' case judgments. However, the fact that our improved pro condition, *Improved Pro*, gave rise to a medium-sized significant effect in comparison to Wysocki's *Con* condition, and also to a small-sized significant effect in comparison to Wysocki's *Baseline*, already indicates that his original materials were probably not ideal for investigating the impact of informal arguments on case judgments. For, prior to testing this impact experimentally, the natural expectation would be that prima facie good informal arguments should at least have *some* measurable impact on people's case judgments, simply because such arguments can generally be expected to have some influence on people's judgments—unless one is an utter cynic about the impact of rational argumentation.

Therefore, our *Experiment 2* aims to improve on Wysocki's (2017) study, first, by only using *prima facie good arguments* in the *Pro* and *Con* conditions, and second, by presenting both arguments and case descriptions in the *more natural setting of a short dialogue* between two friends, and third, by adding *two further types of Gettier-style cases* to broaden the scope of the investigation. Gettier-style cases in particular are well-suited for studying the impact of arguments for and against ascribing knowledge because they combine features that rationally pull in different directions (see also Machery, 2017). Otherwise, it would be mysterious how the traditional account of knowledge in terms of justified true belief could have seemed so plausible to many philosophers in the first place, despite the fact that it is arguably refuted by the specific combination of justified true belief with epistemic luck in Gettier-style cases, which is absent in paradigm cases of justified true belief. So, it is the peculiar combination of justified true belief with accidentally true belief in Gettier-style cases that gives rise to their rational pull in different directions. One might therefore expect that unambiguous cases of both justified true belief and accidentally true belief will not be amenable to the influence of pro and con arguments in the same way as Gettier-style cases, which would be an interesting hypothesis for future research.

The results of *Experiment 2* suggest—pace Wysocki (2017)—that both pro and con arguments for ascribing knowledge do have a considerable effect on participants' case judgments. Averaged across all scenarios, knowledge ascriptions were 22 percentage

[14] Thanks to an anonymous reviewer for prompting these further clarifications.

points higher in *Pro* than in *Con* (Cohen's h = 0.44). The *Baseline* condition lies between these two conditions, with knowledge ascriptions being 12 percentage points higher than in *Con* (h = 0.25) and 10 percentage points lower than in *Pro* (h = 0.19). While there is certainly room for interpretation as to whether the presented informal arguments had a markedly strong impact in our study, their effect is clearly anything but negligible. For example, a recent meta-analysis in social psychology (Lovakov & Agadullina, 2021) found an average effect size of d = 0.36, with the 30th percentile being at d = 0.18. Accordingly, the overall effect we found for *Pro* versus *Con* comparisons was (descriptively) larger than the average effect size in social psychology, and even the relatively smaller effects for comparisons of *Pro* and *Con* with the *Baseline* condition were still larger than 30% of the effects published in the social psychology literature.[15]

Interestingly, the only comparisons at the level of individual scenarios that came out non-significant were between the *Con* and *Baseline* conditions of *Clock* and *Hospital*. One might speculate that the *Baseline* level for ascribing knowledge to the protagonist is already so low in these two types of Gettier-style cases that there is not much to be gained by presenting further con arguments. However, it should be noted that the level of knowledge ascriptions in the *Baseline* condition was surprisingly high for all three cases. With 45%, for example, it comes close to the 50% chance-level in *Clock*, and it still lies at 36% in *Hospital*, which is the scenario that most closely resembles Gettier's own cases. This stands in striking contrast to our replication of Wysocki's (2017) *Tower Clock* case in *Experiment 1*, in which knowledge was ascribed at the much lower level of 17% (and thus almost exactly at the 18% rate in Wysocki's U.S. sample of participants). The most likely reason for this large difference is Wysocki's different answering format in terms of "really knows" and "only believes", which tends to have a polarizing effect on participants' judgments (cf. Cullen, 2010). And while cases of the fake-barn type were already judged as cases of knowledge by the majority of lay and expert participants in previous research (cf. Colaço et al., 2014; Horvath & Wiegmann, 2016; Turri, 2017), the *Baseline* level of 81% for *Dollar* is still staggering, and may in part be due to our binary response options, which force participants to take sides even when they are less than fully confident—unlike the previously used Likert items.

It is important to note, however, that the issue of argument impact is independent of where exactly the baseline lies, as long as there is still room for an effect in both directions, which was clearly the case for all three tested cases in *Experiment 2*. Therefore, the overall result pattern of *Experiment 2*—both across cases and at the level of individual scenarios—strongly suggests that there is indeed a considerable impact of prima facie good informal pro and con arguments on knowledge ascriptions in Gettier-style cases. This also disconfirms Wysocki's (2017) earlier conclusion that arguments have no significant impact on judgments about such cases. Moreover, our findings are in line with the plausible prior expectation that prima facie good informal arguments should at least have *some* measurable impact on people's judgments about hypothetical cases.

[15] Cohen's d (as reported in the meta-analysis by Lovakov & Agadullina, 2021) and Cohen's h (from our own experiment) are directly comparable, since the conventions for a small (0.2), medium (0.5), and large effect (0.8) are exactly the same (see Cohen, 1992, for details).

5 Conclusion

In this paper, we investigated the issue of argument impact on thought experiment judgments against the backdrop of Deutsch and Cappelen's mischaracterization view about the method of cases. This view takes philosophers' judgments about hypothetical cases to be primarily based on arguments, and not intuitions. However, if arguments had no influence on judgments about hypothetical cases, then this would call into question whether these judgments are—or at least should be—based on arguments at all, and not on other available epistemic sources instead, with intuition being the most plausible candidate. In *Experiment 1*, we therefore replicated Wysocki's pioneering study on argument impact for knowledge-ascriptions in Gettier-style cases, and we confirmed his initial result that the informal arguments used by him had no significant effect. However, Wysocki's pro-argument condition was not ideal, and so we also included an improved pro condition with a prima facie good argument for ascribing knowledge, which did have a significant effect on knowledge ascriptions. In *Experiment 2*, we thus followed up with an improved and broadened study on the impact of prima facie good informal arguments about Gettier-style cases, and we found a significant effect of arguments across all cases and conditions and also at the level of individual scenarios. Overall, there was a clear pattern of considerable influence of prima facie good informal arguments on judgments about hypothetical cases—as one would have naturally expected prior to experimental testing. Therefore, the issue of argument impact is no obstacle to the mischaracterization view about the method of cases, except in the unlikely event that philosophers turn out to be much less sensitive to prima facie good informal arguments than laypeople. This last remark requires an important caveat, however: what we have primarily investigated is sensitivity to arguments about hypothetical cases that our lay participants had probably never thought about before. Thus, our lay participants were very likely influenced by these arguments without having any (settled) prior view of the matter. This condition will rarely be satisfied by professional philosophers when they consider well-known thought experiment cases from their own field. So, for studying the impact of informal arguments under conditions of (possibly quite strong) prior views on the cases in question, a different kind of experimental design would be needed. We note this, and also direct experimental research with philosophers, as further desiderata for future research.[16]

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by Deutsche Forschungsgemeinschaft (Grant No. 391304769).

[16] For very helpful comments on previous versions of the paper, we would like to thank two anonymous reviewers of *Synthese* as well as the participants of the workshop *Philosophy's Experimental Turn and the Challenge from Ordinary Language*, Free University Berlin, June 2022, and the EXTRA Workshop.4 *Thought Experiments and Arguments*, Ruhr University Bochum, October 2022. Joachim Horvath and Alex Wiegmann's work on this paper was generously funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project Number 391304769.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

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

Alexander, J., & Weinberg, J. M. (2014). The "unreliability" of epistemic intuitions. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 128–145). Routledge.

Baz, A. (2012). *When words are called for: A defense of ordinary language philosophy*. Harvard University Press.

Booth, A. R., & Rowbottom, D. P. (Eds.). (2014). *Intuitions*. Oxford University Press.

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

Cappelen, H. (2014a). X-Phi without intuitions? In A. R. Booth & D. P. Rowbottom (Eds.), *Intuitions* (pp. 269–286). Oxford University Press.

Cappelen, H. (2014b). Replies to Weatherson, Chalmers, Weinberg, and Bengson. *Philosophical Studies, 171*(3), 577–600. https://doi.org/10.1007/s11098-014-0285-0

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

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

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology, 1*(2), 275–296. https://doi.org/10.1007/s13164-009-0016-1

Deutsch, M. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology, 1*(3), 447–460. https://doi.org/10.1007/s13164-010-0033-0

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

Deutsch, M. (2016). Gettier's method. In J. Nado (Ed.), *Advances in experimental philosophy and philosophical methodology* (pp. 69–97). Bloomsbury.

Deutsch, M. (2017). Replies to commentators. *Inquiry, 60*(4), 420–442. https://doi.org/10.1080/0020174X.2016.1220637

Devitt, M. (2015). Relying on intuitions: Where Cappelen and Deutsch go wrong. *Inquiry, 58*(7–8), 669–699. https://doi.org/10.1080/0020174X.2015.1084824

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Engel Jr., M. (2015). Epistemic luck. In J. Fieser, & B. Dowden (Eds.), *Internet Encyclopedia of philosophy*. http://www.iep.utm.edu/epi-luck/

Ervas, F., Ledda, A., Ojha, A., Pierro, G. A., & Indurkhya, B. (2018). Creative argumentation: When and why people commit the metaphoric fallacy. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.01815

Ervin, W. T., & Corral, D. (2022). The effects of reflective reasoning on philosophical dilemmas. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 7. https://escholarship.org/uc/item/41n0t4wg

Fischer, E., & Engelhardt, P. E. (2020). Lingering stereotypes: Salience bias in philosophical argument. *Mind & Language, 35*(4), 415–439. https://doi.org/10.1111/mila.12249

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121–123. https://doi.org/10.2307/3326922

Goldman, A. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy, 73*(20), 771–791. https://doi.org/10.2307/2025679

Hansen, N. (2020). "Nobody would really talk that way!": The critical project in contemporary ordinary language philosophy. *Synthese, 197*(6), 2433–2464. https://doi.org/10.1007/s11229-018-1812-x

Hansen, N., & Chemla, E. (2015). Linguistic experiments and ordinary language philosophy. *Ratio, 28*(4), 422–445. https://doi.org/10.1111/rati.12112

Hansen, N., Francis, K. B., & Greening, H. (forthcoming). Socratic questionnaires. *Oxford Studies in Experimental Philosophy*.

Herec, J., Sykora, J., Brahmi, K., Vondracek, D., Dobesova, O., Smelik, M., Vaculik, M., & Prochazka, J. (2022). Reflection and reasoning in moral judgment: Two preregistered replications of Paxton, Ungar, and Greene (2012). *Cognitive Science*. https://doi.org/10.1111/cogs.13168

Hilpinen, R. (2017). Sed ubi socrates currit? On the Gettier problem before gettier. In R. Borges, C. de Almeida, & P. D. Klein (Eds.), *Explaining knowledge: New essays on the Gettier problem* (pp. 135–151). Oxford University Press.

Horvath, J. (2015). Thought experiments and experimental philosophy. In C. Daly (Ed.), *The Palgrave handbook of philosophical methods* (pp. 386–418). Palgrave Macmillan. https://doi.org/10.1007/978-1-137-34455-7_16

Horvath, J. (2022). Mischaracterization reconsidered. *Inquiry*. https://doi.org/10.1080/0020174X.2021.2019894

Horvath, J. (2023a). Gettier's thought experiments. In A. Vaidya & D. Prelević (Eds.), *Epistemology of modality and philosophical methodology* (pp. 302–326). Routledge. https://doi.org/10.4324/9781003002192-17/gettier-thought-experiments-joachim-horvath

Horvath, J. (2023b). On the role of intuitions in experimental philosophy. In A. M. Bauer & S. Kornmesser (Eds.), *The compact compendium of experimental philosophy*. Walter de Gruyter.

Horvath, J., & Wiegmann, A. (2016). Intuitive expertise and intuitions about knowledge. *Philosophical Studies, 173*(10), 2701–2726. https://doi.org/10.1007/s11098-016-0627-1

Horvath, J., & Wiegmann, A. (2022). Intuitive expertise in moral judgments. *Australasian Journal of Philosophy, 100*(2), 342–359. https://doi.org/10.1080/00048402.2021.1890162

Kneer, M., Colaço, D., Alexander, J., & Machery, E. (2021). On second thought: Reflections on the reflection defense. *Oxford Studies in Experimental Philosophy, 4*, 257–296. https://doi.org/10.1093/oso/9780192856890.003.0010

Landes, E. (2023). Philosophical producers, philosophical consumers, and the metaphilosophical value of original texts. *Philosophical Studies, 180*(1), 207–225. https://doi.org/10.1007/s11098-022-01900-8

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology, 51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

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

Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Cova, F., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., et al. (2017). The Gettier Intuition from South America to Asia. *Journal of Indian Council of Philosophical Research, 34*(3), 517–541. https://doi.org/10.1007/s40961-017-0113-y

Matilal, B. K. (1991). Perception: An essay on classical Indian theories of knowledge. In *Perception*. Oxford University Press.

Na, R. W., & DeDeo, S. (2022). The diversity of argument-making in the wild: From assumptions and definitions to causation and anecdote in Reddit's "Change My View". *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/20n240qq

Nado, J. (2016). The intuition deniers. *Philosophical Studies, 173*(3), 781–800. https://doi.org/10.1007/s11098-015-0519-9

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*. https://doi.org/10.1126/science.aac4716

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776–783. https://doi.org/10.1037/h0043424

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Pfeifer, N., & Tulkki, L. (2017). Conditionals, counterfactuals, and rational reasoning: An experimental study on basic principles. *Minds and Machines, 27*(1), 119–165. https://doi.org/10.1007/s11023-017-9425-6

Pritchard, D. (2005). *Epistemic luck*. Oxford University Press.

Pust, J. (2017). Intuition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, (Summer 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/intuition/

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

Saint-Germier, P. (2021). Getting Gettier straight: Thought experiments, deviant realizations and default interpretations. *Synthese, 198*(2), 1783–1806. https://doi.org/10.1007/s11229-019-02166-0

Stanley, M. L., Dougherty, A. M., Yang, B. W., Henne, P., & De Brigard, F. (2018). Reasons probably won't change your mind: The role of reasons in revising moral decisions. *Journal of Experimental Psychology. General, 147*(7), 962–987. https://doi.org/10.1037/xge0000368

Sytsma, J., & Livengood, J. (2016). *The theory and practice of experimental philosophy*. Broadview Press.

Travis, C. (1989). *The uses of sense: Wittgenstein's philosophy of language*. Oxford University Press.

Truncellito, D. A. (2007). Epistemology. In J. Fieser, & B. Dowden (Eds.), *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/epistemo/

Turri, J. (2017). Knowledge attributions in iterated fake barn cases. *Analysis, 77*(1), 104–115. https://doi.org/10.1093/analys/anx036

Walton, D. (2008). *Informal logic: A pragmatic approach* (2nd ed.). Cambridge University Press.

Weinberg, J. M. (2014). Cappelen between rock and a hard place. *Philosophical Studies, 171*(3), 545–553. https://doi.org/10.1007/s11098-014-0286-z

Wiegmann, A. (2022). Lying with deceptive implicatures? Solving a puzzle about conflicting results. *Analysis*. https://doi.org/10.1093/analys/anac037

Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive Expertise and Irrelevant Options. *Oxford Studies in Experimental Philosophy, 3*, 275–310. https://doi.org/10.1093/oso/9780198852407.003.0012

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology, 25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995

Wysocki, T. (2017). Arguments over Intuitions? *Review of Philosophy and Psychology, 8*(2), 477–499. https://doi.org/10.1007/s13164-016-0301-8