Synthese (2023) 202:113
https://doi.org/10.1007/s11229-023-04316-x

ORIGINAL RESEARCH

The influence of philosophical training on the evaluation of philosophical cases: a controlled longitudinal study

Bartosz Maćkiewicz[1] · Katarzyna Kuś[1] · Witold M. Hensel[1]

Received: 2 June 2023 / Accepted: 21 August 2023 / Published online: 29 September 2023

Abstract
According to the expertise defense, practitioners of the method of cases need not worry about findings that ordinary people's philosophical intuitions depend on epistemically irrelevant factors. This is because, honed by years of training, the intuitions of professional philosophers likely surpass those of the folk. To investigate this, we conducted a controlled longitudinal study of a broad range of intuitions in undergraduate students of philosophy (n = 226), whose case judgments we sampled after each semester throughout their studies. Under the assumption, made by proponents of the expertise defense, that formal training in philosophy gives rise to the kind of expertise that accounts for changes in the students' responses to philosophically puzzling cases, our data suggest that the acquired cognitive skills only affect single case judgments at a time. There does not seem to exist either a general expertise that informs case judgments in all areas of philosophy, or an expertise specific to particular subfields. In fact, we argue that available evidence, including the results of cross-sectional research, is best explained in terms of differences in adopted beliefs about specific cases, rather than acquired cognitive skills. We also investigated whether individuals who choose to study philosophy have atypical intuitions compared to the general population and whether students whose intuitions are at odds with textbook consensus are more likely than others to drop out of the philosophy program.

Keywords Experimental philosophy · Method of cases · Expertise defense

1 Introduction

The expertise defense is a common reply to concerns about the legitimacy of appeals to intuition in philosophical arguments. It asserts that we have good reason to expect research conducted on ordinary people not to generalize to professional philosophers.

✉ Bartosz Maćkiewicz
b.mackiewicz@uw.edu.pl

[1] Faculty of Philosophy, University of Warsaw, Warsaw, Poland

The reason is that, given their training, professional philosophers tend to make more reliable judgments about philosophically puzzling cases than the folk. This is plausible as far as it goes. However, even if there are differences in intuitions between philosophers and laypeople, it is possible that they do not arise from differences in expertise. For one thing, people who enroll to study philosophy may have atypical intuitions to begin with. For another, philosophy students whose intuitions are at odds with the community-wide consensus may either try to conform, or else withdraw from the philosophy program altogether.

In this paper, we present the results of a controlled longitudinal study designed to shed light on how the putative cognitive skills invoked by the expertise defense develop over the course of a philosophical education. We begin by sketching the two standard challenges to the method of cases that the expertise defense is intended to address (Sect. 2). Both are based on experimental philosophy findings which indicate that ordinary people's judgments about philosophically puzzling cases depend to a significant extent on epistemically irrelevant factors. We then present the expertise defense, taking care to reconstruct its assumptions as to the nature of the cognitive skills it attributes to professional philosophers and the likely time at which they develop. We also characterize three hypothetical models of the scope of philosophical expertise that are amenable to empirical investigation. Next, in Sect. 3, we review existing research aimed at testing the expertise defense. All of this research follows a cross-sectional design, which has certain limitations. We argue that a longitudinal design is poised to address some of them. The main part of the article (Sect. 4) is devoted to presenting the results of our study and Sect. 5—to discussing their philosophical implications.

2 The expertise defense as a reply to the Diversity Challenge and the Questionable Evidence Challenge

Empirical findings have been interpreted as raising two kinds of concerns about using judgments of philosophically puzzling cases as premises in philosophical arguments. The Diversity Challenge (as it is called by Mortensen & Nagel, 2016) appeals to correlational studies indicating that such judgments depend on a variety of demographic variables (Machery, 2017). For example, intuitions about the reference of proper names and natural kind terms differ across cultures (Machery et al., 2004; Beebe & Undercoffer, 2015, Beebe & Undercoffer, 2016; Machery et al., 2010; Machery, 2017, see Dongen et al., 2021 for a meta-analysis), young people are more likely to ascribe knowledge in Fake Barns scenarios than older ones (Colaço et al., 2014), and judgments about free will and moral responsibility vary according to where the person is on the introversion–extraversion dimension (Feltz & Cokely, 2016, 2019).

Findings of this kind are disconcerting because not only are demographic variables epistemically irrelevant, but they are also unchangeable. It is impossible for an extravert to become an introvert, or for someone raised in India to be raised in the US. Nor should such accidental features matter to what kind of philosophical views one espouses. The upshot is that, when a philosophical intuition of one demographic group conflicts with that of another, it makes little sense to say that either one is mistaken.

This obviously complicates conceptual analysis, suggesting that different groups disagree on philosophically puzzling cases by virtue of having different concepts. It also frustrates attempts at exploiting philosophical intuitions to discover the nature of causation, knowledge, free will, etc., as these attempts assume that case judgments are objectively right or wrong (Machery, 2017; Stich & Tobia, 2016).

The Questionable Evidence Challenge, by contrast, appeals to research into effects of manipulable variables, such as the ordering and framing of the stories and questions used in the surveys (again, we borrow the name of the challenge from Mortensen & Nagel, 2016). Some experimental evidence indicates, for example, that people are more likely to ascribe knowledge in the Truetemp case when it is preceded by a clear lack-of-knowledge case than by a clear knowledge case (Swain et al., 2008; Wright, 2010, but see failed replication attempts in Ziółkowski, 2021 and Ziółkowski et al., 2023) and attribute responsibility, free will (Nichols & Knobe, 2007) and causal bases of behavior (Kim et al., 2017) depending on whether the question is asked in abstract or concrete terms. The concern here is that, being susceptible to such cognitive distortions, judgments about cases are not as trustworthy as many philosophers take them to be. Because it is possible to separate the influence of manipulable variables on case judgments from what may be regarded as unadulterated philosophical intuitions, evidence of this kind need not automatically threaten objectivist projects undermined by the Diversity Challenge, but it does suggest caution in pursuing them.

The strength of the Diversity Challenge is the subject of an ongoing debate. As Mortensen and Nagel (2016) point out, many of the demographic effects reported in the older literature have not been detected by subsequent studies, and Knobe (2019) even goes as far as to claim that the robustness of a wide range of philosophical intuitions, as indicated by 30 studies performed on a total of 12,696 subjects, suggests that the intuitions are innate. Yet Stich and Machery (2023) take a very different view of the literature, citing 100 studies, done on a total sample of over 40 million participants, that report differences in philosophical intuitions between various populations.

The expertise defense—a term coined by Weinberg et al. (2010)—sidesteps much of this debate because a vast majority of the findings under discussion concern the philosophical intuitions of the folk. This creates a problem for both challenges, for according to the proponents of the expertise defense, data about case judgments collected from the general population cannot undermine any aspect of philosophical research practices because philosophers have mastered a set of cognitive skills that enable them to outperform ordinary people on tasks involved in doing philosophy. Since construction and evaluation of thought experiments are arguably tasks involved in doing philosophy, it stands to reason that philosophical expertise, fostered by formal training and professional experience, encompasses some cognitive competences that improve performance on those tasks.

While it is not exactly clear what those competences are, they have been suggested to include sensitivity to the structure of philosophical concepts (Ludwig, 2007), and the abilities to closely analyze philosophical texts, construct and evaluate arguments, apply general concepts to specific situations with an eye to relevant detail (Williamson, 2011), and use the tools of formal logic (Weinberg et al., 2010, p. 335). Importantly, the cognitive skills posited by the expertise defense are by no means mysterious or occult. They are much like other familiar kinds of expertise, exhibited by, say, lawyers

(Williamson, 2007), physicists (Hales, 2006) and mathematicians (Ludwig, 2007), who all cope better than the untutored person with problems representative of their domains.

2.1 When philosophical expertise is formed

In psychological research, the term "expertise" denotes the ability to consistently deliver outstanding performance in a given domain. Shanteau (1992, pp. 255–256), a prominent expertise researcher, writes "a naïve decision maker has little or no skill in making decisions in a specific area. For example, graduate students generally are naïve about the kinds of decisions made by experts. Novices are intermediate in skill and knowledge; they frequently have studied for years and may even work at subexpert levels.... Typically, advanced (graduate students) are novices in making skilled decisions."

By contrast, proponents of the expertise defense tend to assume philosophical expertise to arise much earlier than it would if it were defined within psychology. Williamson (2007, p. 191) says "philosophy students have to learn how to apply general concepts to specific examples with careful attention to the relevant subtleties, just as law students have to learn how to analyze hypothetical cases. Levels of disagreement over thought experiments seem to be significantly lower among fully trained philosophers than among novices." And, in a later paper, asks rhetorically "But who ever claimed that the difference in skill at thought experimentation between a professional philosopher and an undergraduate is as dramatic as the difference in skill at chess between a grandmaster and a beginner?" (Williamson, 2011, p. 224).

The reason for this discrepancy is that the expertise defense imposes a special constraint on the cognitive skills involved in making case judgments: that they are sufficiently developed in most members of the philosophical community for appeals to intuition in philosophical arguments to be effective. To put this the other way around, if only a minority of philosophers had the cognitive skills necessary to make credible case judgements, philosophical expertise could not account for community-wide agreement about those judgments.

If, however, philosophical expertise were defined as the set of cognitive skills that enable outstanding performance on philosophical tasks, including tasks involved in thought experimentation, then expert philosophers, thus understood, would not be sufficiently numerous. People of outstanding ability are simply in the minority regardless of the domain, and philosophy is no exception.

A second problem is that, given psychological evidence, it is doubtful whether philosophical training gives rise to genuine expertise, as it is understood in psychology. Weinberg et al. (2010) point out that while training in some domains, including chess, mathematics, physics and meteorology, clearly gives rise to genuine expertise, there are also areas where no amount of experience seems to improve performance—for example, clinical psychology, psychiatry, polygraph testing, and stock brokerage (see, e.g., Shanteau, 1992, p. 258). The crucial difference is that domains where there is genuine expertise rely on well-developed training regimens characterized, among other things, by the availability of large amounts of clear and reliable feedback. Since training in thought experimentation does not rely on such a training regimen, it is unlikely to give rise to genuine expertise.

Given the constraint imposed on the notion of philosophical expertise, it is reasonable to resist the temptation to adapt the standard psychological notion of expertise to the skill of thought experimentation. Accordingly, it is reasonable to suppose that philosophical expertise, in the appropriate sense, develops already during philosophical studies, rather than, say, within the first 10 years after the person has obtained a PhD in philosophy. This, as we shall see, is the approach taken by existing experimental studies into the impact of philosophical expertise on case judgments.

2.2 The likely scope of philosophical expertise

A consequence of the decision to define philosophical expertise in terms of acceptable performance is that one has to be cautious about exploiting psychological expertise research when considering the question of the likely scope of the cognitive skills involved in making case judgments. This means that, at this stage, we know very little indeed about the likely mechanisms underlying case judgment evaluation as well as their domains of operation. The situation is not hopeless, though. It is reasonable to suppose that the putative cognitive abilities developed through philosophical training are more or less restricted to a domain. The only problem is that the models of philosophical expertise that we can propose are tentative and somewhat speculative.

We propose to distinguish three distinct possibilities. First, formal instruction in philosophy may enable students to master a set of skills whose exercise affects case judgments in all areas of philosophy. We may provisionally identify these putative skills with the method of philosophical thought experimentation and suppose that students of philosophy become increasingly adept at making relevant intuitive judgments because, in the course of their education, they encounter and engage with many thought experiments. On this *Method Model of Expertise*, we would expect all case judgments to vary together with the level of competence in appraising thought experiments.

Second, the relevant cognitive skills developed by virtue of studying philosophy may be specific to a subfield. This kind of competence may arise from learning to deploy appropriate theories or having developed domain-specific conceptual schemata. If this *Subfield Model* is accurate, then we would expect training in a given subfield of philosophy, such as epistemology or ethics, to affect all judgments about cases relevant to that subfield, without necessarily influencing judgments in other subfields.

Third, the cognitive skills making up philosophical expertise may be even more specific than that, perhaps being restricted to only one concept or even part of a concept. If this *Restricted Expertise Model* is accurate, then we would expect case judgments to change piecemeal as the person gradually acquires a rich mental representation of the structure of a particular concept.

3 Testing the expertise defense

One way to find out if philosophical training enhances the capacity to make judgments of philosophically puzzling cases is to compare the responses of philosophers and non-philosophers. The expertise defense would be blocked if the responses in both groups were the same.

According to available data, they are not. It has been reported that philosophers' intuitions about phenomenal consciousness do not coincide with those of ordinary people: while philosophers tend to treat diverse experiences such as feeling pain and seeing red as belonging to a single class of phenomenal mental states, the folk tend to distinguish between mental states that essentially have a valence (e.g. feeling pain), those that do not have a valence (e.g. being angry) and those that have both a valence and a perceptual component (e.g. smelling bananas) (Sytsma & Machery, 2010). Similarly, according to Machery (2012), although most people regardless of education are Kripkeans about the reference of proper names, the proportions of the causal–historical vs. descriptivist case judgments vary depending on background. Philosophers of language and semanticists have been found to have more Kripkean intuitions than comparably educated laypeople, whereas the judgments of linguists specializing in discourse analysis, historical linguistics, anthropological linguistics and sociolinguistics are more descriptivist. Differences associated with education have also been discovered in the area of knowledge attribution. According to Starmans and Friedman (2020), subjects holding a PhD in philosophy are less likely than non-philosophy academics or laypeople to attribute knowledge in some Gettier-style scenarios than in standard true justified belief situations. Knowledge attributions made by professional philosophers are also less sensitive to skeptical pressure than those made by either other academics or laypeople. Lastly, non-philosophy academics exhibit more skepticism about knowledge attributions than do philosophers and laypeople.

Although such findings serve to keep the expertise defense in the game, they are also consistent with the claim that philosophers' intuitions are in fact no better than those of the folk. Consequently, studies detecting a significant difference between philosophers and ordinary subjects are often followed up with an investigation into susceptibility to various forms of bias.

The only study focused on a demographic variable we know of speaks against the expertise defense: compatibilist intuitions of professional philosophers, like those of the folk (Feltz & Cokely, 2016), are positively correlated with extraversion (Schulz et al., 2011). As for research into the influence of manipulable variables on case judgments, a majority of studies so far have focused on ethics. Most of their findings also undermine the expertise defense. Professional philosophers engaged in moral reasoning have been found to be affected by persistent ordering effects (Schwitzgebel & Cushman, 2012, 2015), the cleanliness bias (Tobia et al., 2013a, 2013b), the "Asian disease" framing bias (Horvath & Wiegmann, 2022; Schwitzgebel & Cushman, 2015), and the actor–observer bias (Tobia et al., 2013a, 2013b), though it must be noted that this last effect did not replicate (Horvath & Wiegmann, 2022). Other empirical results that weaken the expertise defense indicate that philosophers are subject to the status quo bias in experience machine scenarios (Löhr, 2019) and are as susceptible to certain modal illusions as other academics except mathematicians (Kilov & Hendy, 2022). But there are also data that confirm a positive influence of philosophical training on intuition: philosophers gave more consistent responses to different versions of the experience machine than laypeople (Löhr, 2019), and ethicists, unlike the folk, were unaffected by question-focus bias (Horvath & Wiegmann, 2022). Furthermore, subjects with a PhD in philosophy outperformed laypeople at identifying information relevant to judgments elicited by thought experiments modeled on Gettier cases, the Chinese room, Mary, Fake Barns, and Twin Earth, though the effect was small (Schindler & Saint-Germier, 2022).

However, all the studies to date suffer from an important limitation of all cross-sectional research, in which subjects are compared at a single point in time. Cross-sectional studies can provide a snapshot of dependencies between selected variables in the sample, but they are silent on what kind of processes caused those dependencies to arise. In the case of comparisons made between samples drawn from different populations, there is an indefinite number of differences between the samples that may contribute to the study's outcome. The upshot is that, when subjects from a sample of philosophers exhibit a different pattern of responses from laypeople, this need not be due to a discrepancy in expertise. Because people do not choose their studies at random, philosophy students may have atypical intuitions to start with, or students whose intuitions conflict with those of their teachers may either drop out or strive to align their responses with what they perceive as the mainstream view.

In order to assess whether observed differences between philosophers and laypeople result from training or social selection, it is therefore necessary to conduct longitudinal studies, in which the intuitions of subjects are probed repeatedly over an extended period of time, providing a diachronic picture of variation in the responses. Although observational rather than experimental in character, longitudinal studies can provide invaluable information for causal inference. When employed with this aim in mind, they feature a control group selected in such a way as to resemble the experimental group as closely as possible but not be affected by factors hypothesized to influence the variables of interest.

In what follows, we report the results of a longitudinal study in which two cohorts of undergraduate students in philosophy (the experimental group) and in cognitive science (controls) were tested every semester for 3.5 years on their intuitions regarding ten widely discussed philosophical cases taken from a broad range of subfields. If the assumptions of the expertise defense are true, we should observe changes in intuitions resulting from training in the group of philosophy students. The predicted direction of these expected changes is, at least according to the proponents of philosophical expertise, pretty clear. The intuitions should become increasingly aligned with the consensus in a given area because only then the expertise defense could succeed. Because the courses are spread over time, we can also assess the generality of the putative competences developed by the training (see Sect. 3.2 for the discussion). By comparing the responses of philosophy students with those of the students of cognitive science, we can assess the extent to which observed patterns of changes in the experimental group could be explained in terms of philosophical training as opposed to being the result of a general academic education, age or some other factor present in both groups.

In the study, we investigated (1) whether formal training in philosophy affects case judgments and, if so, whether they are stable over time and whether they converge on textbook consensus, (2) whether the effects of formal training, if any, apply to all case judgments or only to some, and (3) whether the differences in case judgments, if any, between philosophers and laypeople can be explained by appeal to two types of social selection mechanisms: (a) people who enroll to study philosophy already have different intuitions from others, and (b) philosophy students whose intuitions do not conform to the community consensus tend to withdraw from the philosophy program.

4 Longitudinal study

4.1 Method

4.1.1 Materials

We selected from the philosophical literature ten classical cases to be tested. The choice was based on three criteria. First, the cases were selected from a wide range of philosophical subfields. Second, we included cases widely recognised in the philosophical literature. These are either thought experiments backed by a well-established philosophical theory that resolves what the judgment evoked by the case should be (e.g., Gettier cases undermining knowledge defined as a belief that is true and justified) or cases that are related to a certain well-known theory, albeit one that has its competitors in the philosophical market (e.g., Truetemp case as an argument against externalist conceptions of knowledge). Third, we were mainly interested in scenarios that had already been the subject of experimental research. Thus, we chose the Gettier case, Fake Barns, and Truetemp scenarios from epistemology, Putnam's Twin Earth and Kripke's Gödel/Schmidt cases from the philosophy of language, a Knobe-like harm scenario[1] from the philosophy of action, Nozick's Experience Machine, Thompson's Violinist and Frankfurt's (1969) case from ethics and moral philosophy, and Parfit's Teleportation case from metaphysics.

Because many of the philosophical thought experiments in their original form were unsuitable for a questionnaire study,[2] we adapted experimental materials from previous experimental philosophy studies whenever possible. An additional advantage of this solution is that it enables us to use existing data as a baseline for interpretation purposes because we can compare obtained results to the known estimates in the general population. Table 1 presents a list of the scenarios we used in the study together with their original sources. Appendix 1, Table 15 contains the full text of the scenarios in Polish (the language of the survey) and their English translations.

[1] We did not use the Polish translation of the original Knobe scenario, in which the protagonist was the chairman of a board of a company, because at the time of our study, a larger-scale research on the Knobe effect was being conducted at our university in which the translation of the original vignettes was being used.

[2] The majority of the thought experiments in their original form (as they are presented in the primary philosophical literature) are too long, embellished with unnecessary narrative details or too closely connected to philosophical arguments in which they are used.

Table 1 Scenarios used in the study together with their original sources of the thought experiments and empirical studies from which scenarios were adapted for the present study

Scenario	Source of the original scenario	Source of the experimental materials
Gettier case	Gettier (1963)	Weinberg et al. (2001)
Fake Barns	Goldman (1976)	Colaço et al. (2014) and Turri (2017)
Truetemp	Lehrer (1990)	Weinberg et al. (2001) and Swain et al. (2008)
Knobe harm case	Knobe (2003)	Bochyńska (2021)
Twin Earth	Putnam (1974, 1975)	Adapted from the original Putnam's case, because existing experimental materials (e.g., Genone & Lombrozo, 2012; Jylkkä et al., 2009; Nichols et al., 2016) contained too many changes to the original case
Gödel/Schmidt	Kripke (1980)	Machery et al. (2004)
Experience Machine	Nozick (1974)	De Brigard (2010)
Violinist	Thompson (1971)	Adapted from the original case
Frankfurt case	Frankfurt (1969)	Nahmias and Murray (2011) and Miller and Feltz (2011)
Teleportation	Parfit (1984)	Weaver and Turri (2018)

Each scenario was followed by three questions. The first concerned the philosophical intuition elicited by the scenario. It was presented in a forced-choice format (yes/no or choose one from several possibilities). For example, in Fake Barns case, we asked participants whether the protagonist knew that near the road there was a barn. The second question was concerned with subjective confidence in the answer ("What level of confidence would you ascribe to your answer?") and was answered on a pseudo-Likert 5-point scale ranging from "very low" to "very high". From the fourth semester onwards, we also asked participants a yes/no question about whether they had discussed this kind of thought experiment in class. The data is presented in Appendix 2, Table 16.

4.1.2 Translation procedure

The scenarios, originally in all but one case formulated in English, were translated into Polish by a person with a formal education in English–Polish translation. The translations were then reviewed by two members of our team who have a background in philosophy and experience in translating philosophical texts from English to Polish. In a few cases, the scenarios were slightly altered to address the exact problem we intended to study, or to ask the type of question we chose. When necessary, the scenarios were adapted to the Polish participants' knowledge (for example, in the Gettier case, information was added that the Buick is an American car). Suggested corrections were then consulted with the translator.

4.1.3 Procedure

The study consisted of seven measurement points: six at the beginning of each semester of the undergraduate program and the seventh at the beginning of the academic year following participants' completion of the program. At each measurement point, the same questionnaire was administered with minor changes aimed to shorten the duration of the study (several repeated demographic questions were dropped).

In order to increase the response rate, we employed a mixed-mode survey design that combined a traditional pen & paper questionnaire with an online survey. At the first and second measurement points, a paper questionnaire was administered during an obligatory class. Participation was voluntary. For those who were unable to participate in the pen & paper version of the study, an online survey was also available. During the rest of the study, an online survey was the dominant mode of participation. The second author stayed in touch with all participants and reminded them each semester via e-mail to complete the next part of the study.

The research received ethical approval from the appropriate University Research Ethics Committee and informed consent was obtained from all participants. The surveys were anonymized by assigning each participant a unique identifier that participants needed to sign each completed questionnaire.

4.1.4 Participants

The participants were undergraduate students of philosophy and undergraduate students of cognitive science at the University of Warsaw, Poland. The sample of cognitive science students was used as a matching control group in order to evaluate the confounding effects of age and education.

The philosophy program at the University of Warsaw is relatively fixed during the first four semesters. In the first year, students are required to participate in a two-semester epistemology course. In the second year, there are obligatory two-semester courses in ethics and ontology. In the third year, philosophy of language and philosophy of mind are discussed as part of a compulsory course in the history of analytic philosophy. Third-year students are also offered an elective course in philosophy of action, philosophy of language, and philosophy of mind.

Courses offered at the University of Warsaw typically consist of two parts: lectures and tutorials in small groups. In the philosophy program, many classical philosophical thought experiments are discussed in lectures and especially tutorials. Gettier cases and Fake Barns are discussed thoroughly in epistemology classes from the middle to the end of the first semester. Criteria of identity over time are discussed in detail during the ontology tutorial (fourth semester), where students are also introduced to various thought experiments designed to elicit intuitions about personal identity. During the third semester, also in ontology classes, when discussing the concept of particular objects and modal notions, students learn about Gödel/Schmidt and Twin Earth thought experiments, the latter of which is briefly described earlier in an epistemology lecture on externalism. The Violinist case is discussed during the second year in ethics classes, whereas the Frankfurt case receives a cursory mention in the ethics lecture at the end of the fourth semester. In the third year (fifth and sixth semesters), students

Table 2 The number of participants who took part in each stage of the study

Measurement point	Fall 2017 cohort							Fall 2018 cohort						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Control (Cog-Sci)	54	45	42	40	39	32	32	28	30	26	18	25	24	18
Experimental (Phil)	68	57	34	23	25	23	21	30	28	20	18	16	19	16

discuss in depth the Gödel/Schmidt and Twin Earth cases in the history of analytic philosophy. The Knobe experiment is mentioned (but not discussed in detail) in facultative philosophy of action classes (third year). As far as we know, the Truetemp and Experience Machine thought experiments are not covered in class at an undergraduate level.

Our control group was not perfect. Ideally, the students from the control group should not take any courses in philosophy. Unfortunately, this was not the case. The cognitive science program includes some elements of philosophy, which is common for all undergraduate programs at the University of Warsaw. The students are required to take a short introductory course in philosophy in the second semester, followed by obligatory courses in philosophy of language (where both Gödel/Schmidt and Twin Earth are discussed) and philosophy of mind in the third semester. They can also choose advanced courses in philosophy of mind and philosophy of language during the fourth semester. In addition, both cognitive science and philosophy students take a compulsory 120-h logic course during the first year of their studies. Hence, if learning logic has a significant influence on the formation of philosophical expertise (cf. Weinberg et al., 2010, p. 335), the influence should be the same in both groups.

Two cohorts participated in the study: students who started their program in Fall 2017 and those who started in Fall 2018. Table 2 presents sample sizes for each of the measurement points. In total, 226 students took part in the study [112 men, 107 women, 1 other gender and 6 participants who refused to answer the question; mean age at the first measurement point: 20.0 years old (SD = 1.48)]. 180 Subjects participated in the study from the beginning, 33 students started at the second measurement point, 4 students at the third, 3 at the fourth, 4 at the fifth, and 1 student at the sixth.[3]

4.2 Results

For all scenarios, we computed a combined score (e.g., Turri, 2016a) in the following manner: if the answer was "yes," we multiplied the confidence rating by 1, and if the answer was "no," we multiplied the rating by − 1. For answers that were neither "yes" or "no," we multiplied the confidence rating by one just in case the answer was

[3] All participants declared their readiness to take part in the study in October of their first semester of study, but not all students took part in all stages of the study. In particular, despite their declaration, some students started their participation at a later stage.

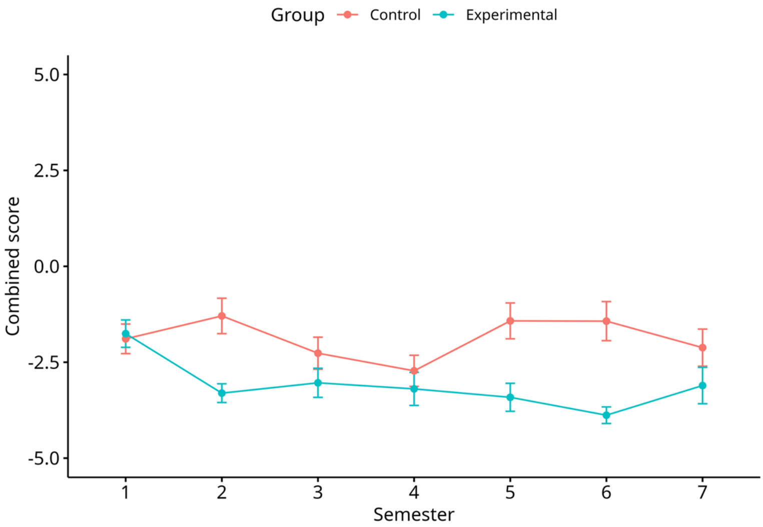

Fig. 1 Changes in intuitions about the Gettier case over time. The combined score of + 5 represents attribution of knowledge with maximum confidence and − 5 represents a denial of knowledge with maximum confidence. Error bars correspond to the standard error of the mean

consistent with philosophical consensus. We thereby obtained the main numerical dependent variable that ranged from − 5 to + 5. For each such measurement, we fitted a linear mixed model using *R lme4* package (Bates et al., 2015) with a group (cog-sci students vs. philosophy students) and time of the measurement (1–7) as predictors. For the time, we have successive differences contrast coding (cf. Venables & Ripley, 2002, pp. 147–149). The motivation behind this choice was that we were mainly interested in changes from semester to semester and not general linear trends. In cases where closer inspection of the results was called for, we fitted two additional models to the data. The first was a linear mixed model but, instead of using successive differences contrast coding, we coded the time of the measurement as a numeric variable. This allowed us to investigate simple linear trends that would be not visible in a semester-by-semester analysis. The second additional model was the same as the main one but fitted only to the data from philosophy students. It was used to investigate data in cases when the main analysis yielded interaction effects that were difficult to interpret. We also conducted an analysis of bare categorical responses for which we fitted a generalized linear mixed model with logit as a link function. The results of these analyses, together with additional data, are reported in Appendix 3, Figs. 12, 13, 14, 15, 16, 17, 18, 19, 20, and 21 and Tables 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26.[4]

4.2.1 Gettier case

The results are presented in Fig. 1. The participants in both groups began with similar confidence scores and their responses were consistent with a large body of available

[4] All data, experimental matirials and code are available at: https://osf.io/gu6x5/.

Table 3 Linear mixed-effects model for the combined scores in the Gettier case

Predictors	Combined score (Gettier case)		
	Estimates	CI	p
(Intercept)	− 1.78	− 2.24 to − 1.31	**< 0.001**
Semester 2-1	0.60	− 0.24 to 1.43	0.159
Semester 3-2	− 0.96	− 1.83 to − 0.09	**0.030**
Semester 4-3	− 0.07	− 0.99 to 0.85	0.877
Semester 5-4	0.94	0.00 to 1.87	**0.050**
Semester 6-5	− 0.19	− 1.14 to 0.76	0.692
Semester 7-6	− 0.28	− 1.29 to 0.73	0.586
Group [Experimental]	− 1.39	− 2.06 to − 0.72	**< 0.001**
Semester 2-1: GroupExperimental	− 2.26	− 3.41 to − 1.12	**< 0.001**
Semester 3-2: GroupExperimental	1.16	− 0.11 to 2.42	0.073
Semester 4-3: GroupExperimental	− 0.17	− 1.59 to 1.25	0.816
Semester 5-4: GroupExperimental	− 0.98	− 2.47 to 0.51	0.196
Semester 6-5: GroupExperimental	− 0.25	− 1.74 to 1.24	0.741
Semester 7-6: GroupExperimental	0.92	− 0.62 to 2.46	0.241
Random Effects			
σ^2	6.66		
τ_{00} ident	3.86		
ICC	0.37		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.063/0.407		

Bold is used to indicate statistically significant predictors

research on the epistemic intuitions of subjects from the general population (e.g., Machery et al., 2017, 2018; Nagel et al., 2013; Turri, 2013). The majority shared Gettier intuitions, refusing to ascribe knowledge to the protagonist of the story. The situation started to differ from the second measurement point onwards. We observed a large change in scores between the first and second semesters in the experimental group (philosophy students), whereas in the controls (cognitive science students) the ratings remained stable.[5] The difference in the experimental group between the first and second measurements was statistically significant (interaction between Group and Semester 2-1: $b = -2.26$, $p < 0.001$; see Table 3). Overall scores in the experimental group were significantly lower than those in the control group ($b = -1.39$, $p < 0.001$). A closer examination of the data based on a model with semester coded as a numeric variable revealed a statistically significant interaction between Semester and Group ($b = -0.27$, $p = 0.005$). Together with no statistically significant first-order effects

[5] A model fitted only to data from the group of philosophy students confirms this observation (Semester 2-1: $b = -1.63$, $p < 0.001$).

in this model, this should be interpreted as evidence for a negative linear trend in the experimental group that is not present in the controls.

4.2.2 Fake Barns

Although there are doubts as to the replicability of these results, Fake-barn cases have been found to be sensitive to some demographic variables, such as age (Colaço et al., 2014) and gender (Bergenholtz et al., 2023), and presentation order (Wright, 2010). Nonetheless, the results so far have regularly shown that lay people tend to attribute knowledge to the protagonist of the Fake-barn scenario (Colaço et al., 2014; Turri et al., 2015; Turri, 2016b, 2017). Our results are presented in Fig. 2. At the first measurement point, participants in both groups tended to attribute knowledge to the protagonist although the mean combined score was slightly lower in the experimental group. The tendency to attribute knowledge reversed for the experimental group from second semester onward ($b = -1.86$, $p = 0.005$, see Table 4). For the rest of the study, philosophy students tended to refrain from attributing knowledge to the protagonist (mean score below the midpoint) whereas scores for the cognitive students remained stable (and positive). The overall difference between the groups was statistically significant ($b = -2.42$, $p < 0.001$). No general linear trend was found ($p > 0.05$), which suggests that the observed change in intuitions occurred mainly between the first two measurement points.

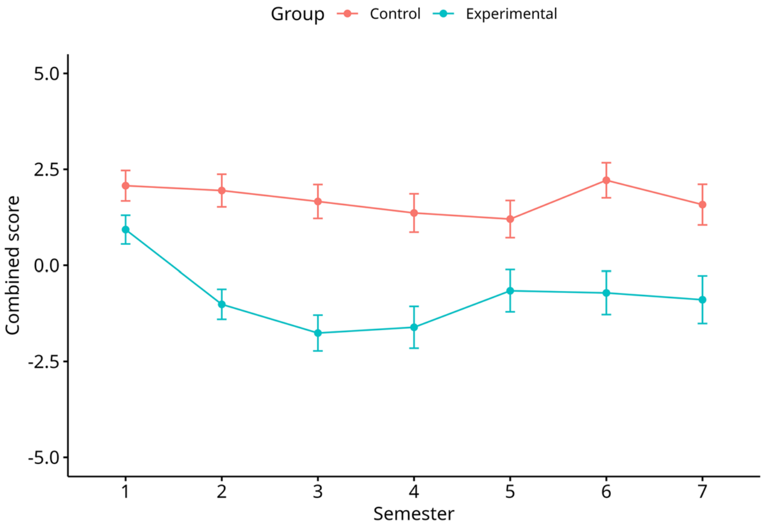

Fig. 2 Changes in intuitions about the Fake Barns case over time. The combined score of + 5 represents attribution of knowledge with maximal confidence and − 5 represents a denial of knowledge with maximal confidence. Error bars correspond to the standard error of the mean

Table 4 Linear mixed-effects model for the combined scores in the Fake Barns case

Predictors	Combined score (Fake Barns)		
	Estimates	CI	p
(Intercept)	1.76	1.24 to 2.27	**< 0.001**
Semester 2-1	− 0.14	− 1.09 to 0.80	0.764
Semester 3-2	− 0.07	− 1.06 to 0.91	0.882
Semester 4-3	− 0.37	− 1.41 to 0.68	0.489
Semester 5-4	− 0.17	− 1.23 to 0.89	0.753
Semester 6-5	0.73	− 0.35 to 1.80	0.186
Semester 7-6	− 0.34	− 1.48 to 0.81	0.563
Group [Experimental]	− 2.42	− 3.16 to − 1.67	**< 0.001**
Semester 2-1: GroupExperimental	− 1.86	− 3.15 to − 0.56	**0.005**
Semester 3-2: GroupExperimental	− 0.59	− 2.02 to 0.84	0.420
Semester 4-3: GroupExperimental	0.50	− 1.10 to 2.11	0.539
Semester 5-4: GroupExperimental	1.00	− 0.68 to 2.69	0.242
Semester 6-5: GroupExperimental	− 0.62	− 2.30 to 1.07	0.474
Semester 7-6: GroupExperimental	0.23	− 1.52 to 1.97	0.796
Random Effects			
σ^2	8.54		
τ_{00} ident	4.66		
ICC	0.35		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.115/0.427		

Bold is used to indicate statistically significant predictors

4.2.3 Truetemp

The results of previous studies show that lay people tend to disagree with the knowledge attribution in Truetemp cases or the average knowledge ratings are close to the midpoint of the scale (Swain et al., 2008; Wright, 2010; Ziółkowski, 2021). In our study, we observed an overall difference between groups ($b = 0.91, p = 0.24$, see Table 5). The mean rating in both groups was close to the midpoint (see Fig. 3), but the ratings were generally lower in the experimental group. Interestingly, this was the other way around at the first measurement point. The change from the first to the second semester ($b = 1.67$, $p < 0.001$) and an interaction between the semester factor and group ($b = -2.25, p = 0.002$) were statistically significant. After the first two periods, the intuitions remained more or less stable. An analysis with measurement points coded as numbers revealed a small but statistically significant positive effect of the semester variable ($b = 0.23$, $p = 0.003$) and an interaction between semester and experimental group but with the opposite sign (b = − 0.30, p = 0.009), which means that the linear trend was present only in the control group.

Table 5 Linear mixed-effects model for the combined scores in the Truetemp case

Predictors	Combined score (Truetemp)		
	Estimates	CI	p
(Intercept)	0.36	− 0.18 to 0.91	0.190
Semester 2-1	1.67	0.65 to 2.68	**0.001**
Semester 3-2	− 0.20	− 1.25 to 0.86	0.715
Semester 4-3	0.40	− 0.72 to 1.51	0.488
Semester 5-4	− 0.24	− 1.38 to 0.89	0.677
Semester 6-5	0.42	− 0.73 to 1.57	0.476
Semester 7-6	− 0.38	− 1.61 to 0.84	0.540
Group [Experimental]	− 0.91	− 1.69 to − 0.12	**0.024**
Semester 2-1: GroupExperimental	− 2.25	− 3.65 to − 0.86	**0.002**
Semester 3-2: GroupExperimental	− 0.41	− 1.94 to 1.13	0.602
Semester 4-3: GroupExperimental	− 0.07	− 1.79 to 1.65	0.939
Semester 5-4: GroupExperimental	1.39	− 0.41 to 3.19	0.131
Semester 6-5: GroupExperimental	− 1.65	− 3.46 to 0.16	0.073
Semester 7-6: GroupExperimental	0.62	− 1.25 to 2.49	0.518
Random Effects			
σ^2	9.81		
τ_{00} ident	5.07		
ICC	0.34		
N_{ident}	226		
Observations	847		
Marginal R^2/Conditional R^2	0.033/0.362		

Bold is used to indicate statistically significant predictors

4.2.4 Knobe harm case

Given the robustness of the Knobe effect, we expected a positive answer—that is, an attribution of intentionality to the action's side-effect.[6] Surprisingly, neither philosophy nor cognitive science students tended to ascribe intentionality in the Knobe-like scenario (see Fig. 4). The intuitions were stable over time, but the means were much closer to the midpoint in the experimental group than in the control group, where we observed a moderately strong negative response ($b = 1.55$, $p < 0.001$, see Table 6). Analyzing only the group of philosophy students, we found a statistically significant difference between the fourth and the fifth semesters in the opposite direction of orthodox theory of intentional action ($b = 1.34$, $p = 0.035$). No simple linear trends were observed.

[6] The vignette we used, with a similarly worded question, was previously tested on the Polish population. In the harm scenario, a large majority of subjects were willing to attribute intentionality to the protagonist's action (cf. Bochyńska, 2021).

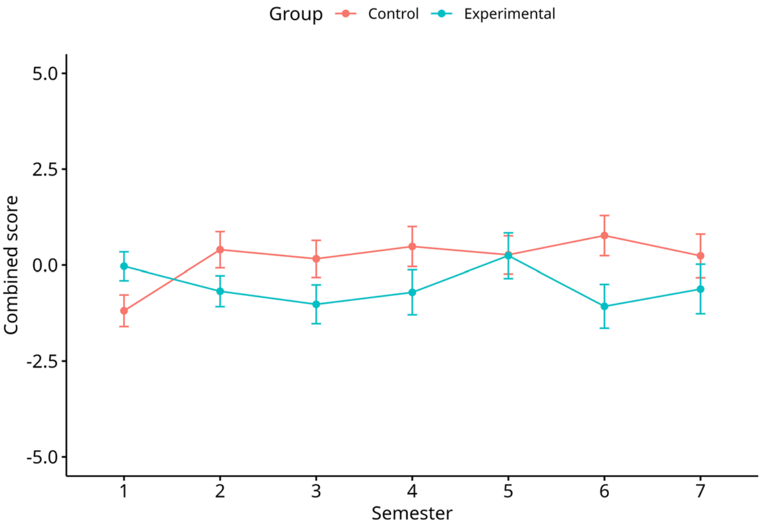

Fig. 3 Changes in intuitions about the Truetemp case over time. The combined score of + 5 represents attribution of knowledge with maximum confidence and − 5 represents a denial of knowledge with maximum confidence. Error bars correspond to the standard error of the mean

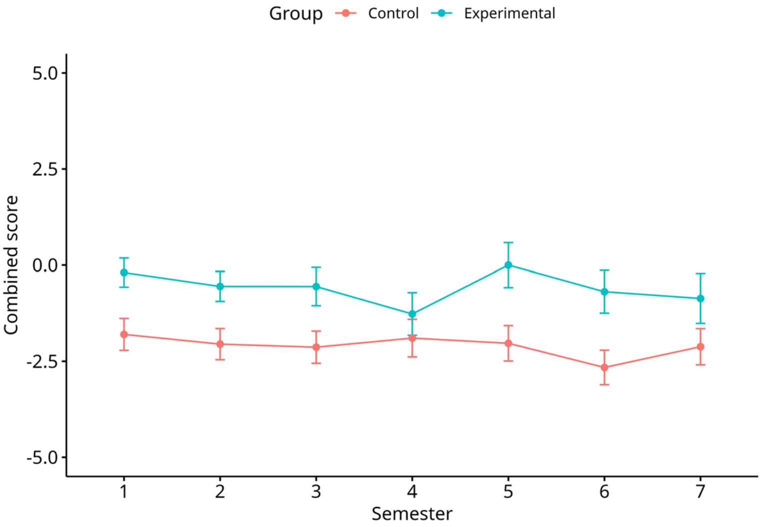

Fig. 4 Changes in intuitions about the Knobe harm case over time. The combined score of + 5 represents attribution of intentionality with maximum confidence and − 5 represents a denial of intentionality with maximum confidence. Error bars correspond to the standard error of the mean

Table 6 Linear mixed-effects model for the combined scores in the Knobe harm case

Predictors	Combined score (Knobe)		
	Estimates	CI	p
(Intercept)	− 2.04	− 2.59 to − 1.49	**< 0.001**
Semester 2-1	− 0.24	− 1.12 to 0.64	0.588
Semester 3-2	0.00	− 0.90 to 0.91	0.995
Semester 4-3	0.35	− 0.61 to 1.31	0.474
Semester 5-4	− 0.15	− 1.13 to 0.82	0.755
Semester 6-5	− 0.40	− 1.39 to 0.59	0.424
Semester 7-6	0.14	− 0.91 to 1.19	0.793
Group [Experimental]	1.55	0.76 to 2.34	**< 0.001**
Semester 2-1: GroupExperimental	− 0.09	− 1.30 to 1.11	0.877
Semester 3-2: GroupExperimental	0.07	− 1.25 to 1.38	0.921
Semester 4-3: GroupExperimental	− 0.99	− 2.47 to 0.48	0.186
Semester 5-4: GroupExperimental	1.49	− 0.05 to 3.04	0.058
Semester 6-5: GroupExperimental	− 0.49	− 2.04 to 1.06	0.531
Semester 7-6: GroupExperimental	− 0.07	− 1.67 to 1.53	0.930
Random Effects			
σ^2	7.15		
τ_{00} ident	5.92		
ICC	0.45		
N_{ident}	226		
Observations	849		
Marginal R^2/Conditional R^2	0.050/0.481		

Bold is used to indicate statistically significant predictors

4.2.5 Twin Earth

Most previous studies have used more complex scenarios than ours. Their results undermine both pure internalism and pure externalism as two opposing folk theories of natural kind terms regardless of the type of natural kinds used in the scenario, and suggest the need for some hybrid account, according to which natural kind terms are ambiguous or polysemous (Jylkkä et al., 2008; Genone & Lombrozo, 2012; Nichols et al., 2016; Tobia et al., 2020). The participants of our study were asked whether XYZ is water. In both groups, the dominant answer was "no" and the intuitions about the case were relatively strong (see Fig. 5). The only statistically significant effect was a positive change at the second measurement point ($b = 0.98$, $p = 0.007$, see Table 7 for detailed results), which is also statistically significant for philosophy students analyzed separately ($b = 1.01$, $p = 0.004$). An analysis with semester coded as a numeric variable revealed a weak although statistically significant overall linear trend in the direction of agreeing with the claim that XYZ is water ($b = 0.20$, $p < 0.001$).

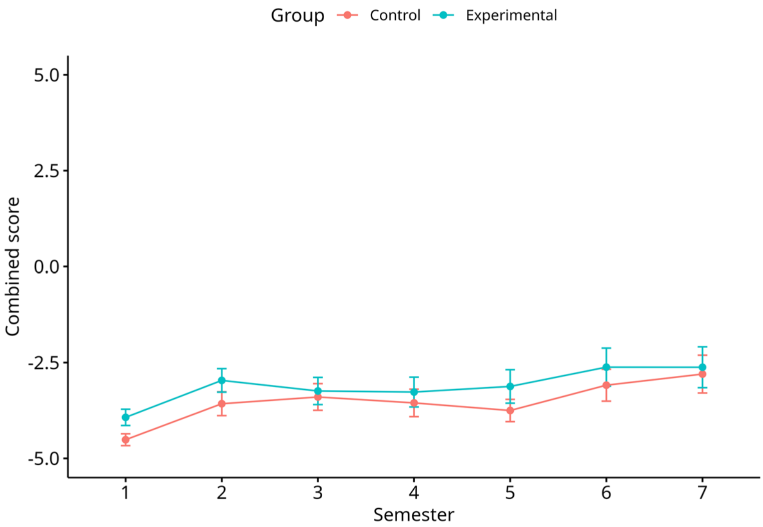

Fig. 5 Changes in intuitions about the Twin Earth case over time. The combined score of + 5 represents a belief in the statement that XYZ is water with maximum confidence and − 5 represents a denial of that statement with maximum confidence. Error bars correspond to the standard error of the mean

All in all, the participants were less likely over time to agree with the answer implied by Putnam's account of natural kinds.

4.2.6 Gödel/Schmidt

Previous studies (e.g., Beebe & Undercoffer, 2015, 2016; Machery et al., 2004, 2009, 2015; Sytsma et al., 2015) have shown that, in the western cultures, people's judgments about the reference of proper names in Gödel/Schmidt cases are consistent with Kriple's causal-history theory. In our study, this was the first case where the question was not in the "yes/no" format. Instead, we asked the participants who the protagonist of the story was talking about when he used the name "Gödel". Following the standard description of this case, there were two possible answers: the author of the theorem or the fraud. For analysis purposes, we decided to code the answer consistent with the causal theory of reference (the fraud) as 1 and the descriptivist answer (the author) as − 1. Thus, the answers combined with confidence ratings form a variable ranging from − 5 (strong confidence and descriptivist intuitions) to + 5 (strong confidence and causal–historical intuitions). Overall, the intuitions of philosophy students were more in line with the Kripkean theory of reference ($b = 0.92$, $p = 0.022$, see Table 8) than the intuitions of the controls, but the means in both groups were very similar at the first and sixth measurement points (see Fig. 6). Statistically significant change towards the negative answer was observed between the sixth and the seventh semesters ($b = -1.97$, $p = 0.001$). However, a marginally significant interaction with the opposite sign ($b = 1.83$, $p = 0.051$) indicates that the drop in ratings occurred predominantly in the control group. A separate analysis of the data from the experimental group revealed a statistically significant change in the direction of philosophical orthodoxy between

Table 7 Linear mixed-effects model for the combined scores in the Twin Earth case

Predictors	Combined score (Twin Earth)		
	Estimates	CI	p
(Intercept)	− 3.51	− 3.88 to − 3.15	**< 0.001**
Semester 2-1	0.98	0.27 to 1.69	**0.007**
Semester 3-2	0.14	− 0.60 to 0.88	0.705
Semester 4-3	− 0.11	− 0.90 to 0.67	0.778
Semester 5-4	− 0.23	− 1.03 to 0.57	0.573
Semester 6-5	0.63	− 0.18 to 1.44	0.125
Semester 7-6	0.26	− 0.60 to 1.13	0.548
Group [Experimental]	0.36	− 0.17 to 0.89	0.183
Semester 2-1: GroupExperimental	0.04	− 0.94 to 1.01	0.942
Semester 3-2: GroupExperimental	− 0.48	− 1.55 to 0.60	0.382
Semester 4-3: GroupExperimental	0.03	− 1.18 to 1.24	0.964
Semester 5-4: GroupExperimental	0.44	− 0.83 to 1.71	0.495
Semester 6-5: GroupExperimental	− 0.25	− 1.52 to 1.02	0.703
Semester 7-6: GroupExperimental	− 0.23	− 1.54 to 1.09	0.735
Random Effects			
σ^2	4.85		
τ_{00} ident	2.20		
ICC	0.31		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.035/0.336		

Bold is used to indicate statistically significant predictors

the second and the third measurement points ($b = 1.23$, $p = 0.013$). No statistically significant simple linear trends were found.

4.2.7 Experience Machine

According to available research, a large majority of ordinary people who respond to vignettes modeled closely on Nozick's original scenario refuse to be connected to the experience machine (de Brigard, 2010; Hindriks & Douven, 2018; Weijers, 2014). In our study, participants were presented with two choices: remain in the real world or plug into the experience machine. Following Nozick's original analysis, we coded the former option as 1 and the latter as − 1. Combined with the confidence ratings, the dependent variable ranged from − 5 (a strong preference for connecting to the machine) to + 5 (a strong preference for staying in the real world). We did not find any statistically significant effects (see Table 9) and, as Fig. 7 shows, the intuitions remained rather stable across all measurement points. The participants tended to agree

Table 8 Linear mixed-effects model for the combined scores in the Gödel/Schmidt case

Predictors	Combined score (Gödel/Schmidt)		
	Estimates	CI	p
(Intercept)	– 0.76	– 1.30 to – 0.22	**0.006**
Semester 2-1	– 0.66	– 1.65 to 0.34	0.194
Semester 3-2	0.82	– 0.21 to 1.86	0.120
Semester 4-3	– 0.19	– 1.29 to 0.91	0.731
Semester 5-4	0.66	– 0.45 to 1.78	0.244
Semester 6-5	0.84	– 0.30 to 1.97	0.147
Semester 7-6	– 1.97	– 3.18 to – 0.77	**0.001**
Group [Experimental]	0.92	0.13 to 1.70	**0.022**
Semester 2-1: GroupExperimental	1.35	– 0.02 to 2.71	0.053
Semester 3-2: GroupExperimental	0.47	– 1.04 to 1.98	0.539
Semester 4-3: GroupExperimental	0.35	– 1.34 to 2.04	0.684
Semester 5-4: GroupExperimental	– 1.01	– 2.78 to 0.76	0.262
Semester 6-5: GroupExperimental	– 0.71	– 2.49 to 1.06	0.432
Semester 7-6: GroupExperimental	1.83	– 0.00 to 3.67	0.051
Random Effects			
σ^2	9.47		
τ_{00} ident	5.14		
ICC	0.35		
N_{ident}	225		
Observations	850		
Marginal R^2/Conditional R^2	0.043/0.380		

Bold is used to indicate statistically significant predictors

with Nozick's interpretation of this case and rather confidently said that they would remain in the real world. Again, no overall linear trends were found.

4.2.8 Violinist

Subjects were asked whether they had a moral duty to stay connected to the violinist. Mean scores below midpoint indicate that they disagree with the claim, but as can be seen in Fig. 8, they were relatively close to the midpoint. The scores did not differ significantly between the two groups and no general linear trends were observed. In the experimental group, we observed an increase in scores between the third and fourth semesters (interaction: $b = 1.74$, $p = 0.013$, separate model: $b = 1.05$, $p = 0.045$) and a decrease between the fourth and fifth semesters (interaction: $b = -1.53$, $p = 0.038$, separate model: $b = -1.13$, $p = 0.043$, see Table 10 for detailed results).

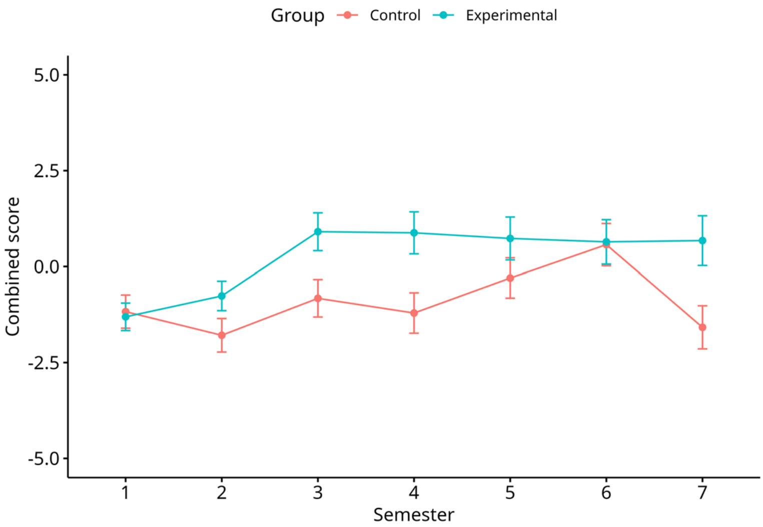

Fig. 6 Changes in intuitions about the Gödel/Schmidt case over time. The combined score of + 5 represents a belief in the statement that the name "Gödel" refers to the fraud with maximum confidence and − 5 represents a belief in the statement that it refers to the author of the proof with maximum confidence. Error bars correspond to the standard error of the mean

4.2.9 Frankfurt case

A study by Miller and Feltz (2011) indicates that people are willing to ascribe moral responsibility and blameworthiness in cases where there were no alternative possibilities available to an agent. In our study, subjects were asked three questions: whether it was possible for Frank not to kill Furt (*Possible not to kill?*), whether he was responsible for Furt's death (*Responsible?)* and whether he was blameworthy for killing Furt (*Blameworthy?*). The only statistically significant effect found in the responses to the first question was a difference between sixth and seventh measurement points in the model fitted only to responses by philosophy students ($b = -1.50, p = 0.032$). Participants in both groups tended to believe that it was not possible for Frank not to kill Furt. They were also highly confident that Frank was both responsible and blameworthy for the killing. Interestingly, we found a statistically significant decrease in scores between the first and second semesters (*Responsible?*: $b = -0.75, p = 0.044$; *Blameworthy?*: $b = -0.89$, $p = 0.013$). A closer examination of this first-order effect suggests that the control group was responsible for it, which is reflected in a statistically significant interaction in the *Blameworthy?* question ($b = 1.31$, $p = 0.008$, see Fig. 9 for the visual comparison and Table 11 for detailed results). Again, analyzing only the group of philosophy students we found a statistically significant positive difference between the sixth and seventh semesters with regard to the *Responsible?* question ($b = 1.03$, $p = 0.040$), which is consistent with the previously noted negative difference with regard to the *Possible not to kill?* question.

Table 9 Linear mixed-effects model for the combined scores in the Experience Machine case

Predictors	Combined score (Experience Machine)		
	Estimates	CI	p
(Intercept)	3.14	2.69 to 3.60	**< 0.001**
Semester 2-1	0.27	− 0.33 to 0.88	0.372
Semester 3-2	0.27	− 0.35 to 0.89	0.397
Semester 4-3	− 0.01	− 0.67 to 0.65	0.981
Semester 5-4	− 0.28	− 0.95 to 0.39	0.409
Semester 6-5	− 0.55	− 1.23 to 0.13	0.113
Semester 7-6	− 0.10	− 0.83 to 0.62	0.784
Group [Experimental]	− 0.20	− 0.84 to 0.45	0.546
Semester 2-1: GroupExperimental	− 0.82	− 1.66 to 0.01	0.053
Semester 3-2: GroupExperimental	− 0.09	− 1.01 to 0.82	0.844
Semester 4-3: GroupExperimental	0.11	− 0.91 to 1.12	0.835
Semester 5-4: GroupExperimental	0.13	− 0.94 to 1.19	0.816
Semester 6-5: GroupExperimental	0.63	− 0.44 to 1.70	0.249
Semester 7-6: GroupExperimental	0.27	− 0.84 to 1.37	0.635
Random Effects			
σ^2	3.37		
τ_{00} ident	4.50		
ICC	0.57		
N_{ident}	225		
Observations	847		
Marginal R^2/Conditional R^2	0.011/0.576		

Bold is used to indicate statistically significant predictors

4.2.10 Teleportation

Results of six studies by Weaver and Turri (2018) suggest that people allow for the possibility that one and the same individual can be in two different places at the same time. Our Teleportation case, which was closely modeled on Parfit’s thought experiment, involved split teleportation, where a malfunctioning teleporter reconstructed the teleported person in two copies. Participants were able to select one from four possible answers. Two answers implied that one of the two copies was identical to the original person, the third implied that both copies were identical, and the fourth—that neither copy was the original person. We coded the last option as + 1 and the rest of the possible answers as − 1. We did not observe any statistically significant effects, regardless of the model used to analyze the data. Participants did not exhibit a strong preference for any of the answers (see Fig. 10). The results are presented in Table 12. Appendix 4, Fig. 22 and Table 27 contain a detailed breakdown of the answers. The participants were fairly equally split between “neither” and “both” answers, which is consistent with previous research.

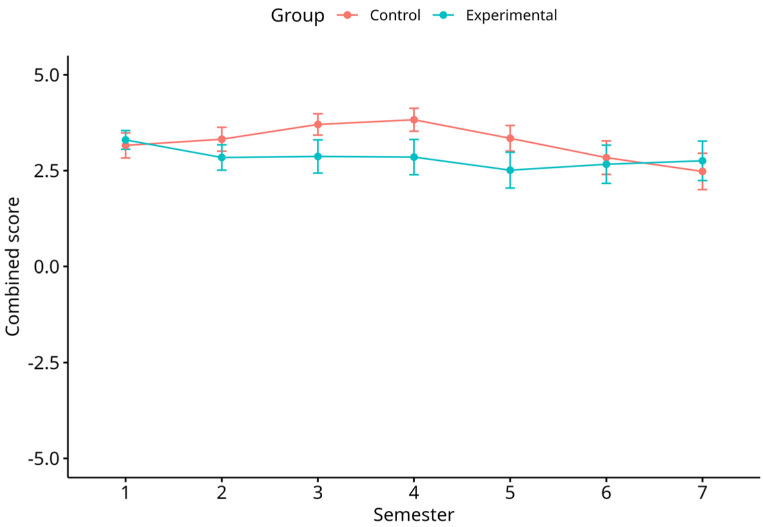

Fig. 7 Changes in intuitions about the Experience Machine case over time. The combined score of + 5 represents a preference for staging in the real world with maximum confidence and − 5 represents a preference for connecting to the machine with maximum confidence. Error bars correspond to the standard error of the mean

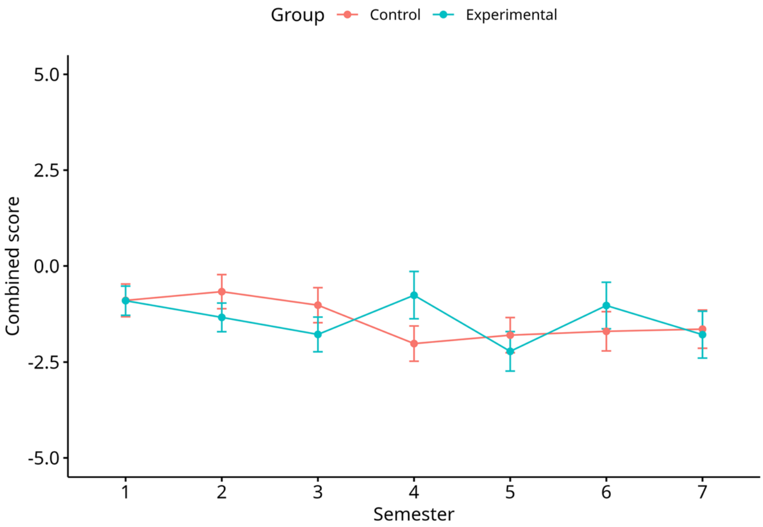

Fig. 8 Changes in intuitions about the Violinist case over time. The combined score of + 5 represents a belief that one has a moral duty to stay connected to the violinist with maximum confidence and − 5 represents a denial of the existence of this duty with maximum confidence. Error bars correspond to the standard error of the mean

Table 10 Linear mixed-effects model for the combined scores in the Violinist case

Predictors	Combined score (Violinist)		
	Estimates	CI	p
(Intercept)	− 1.51	− 2.10 to − 0.92	**< 0.001**
Semester 2-1	0.17	− 0.66 to 0.99	0.691
Semester 3-2	− 0.49	− 1.34 to 0.35	0.253
Semester 4-3	− 0.70	− 1.59 to 0.20	0.125
Semester 5-4	0.40	− 0.51 to 1.31	0.384
Semester 6-5	− 0.47	− 1.39 to 0.46	0.323
Semester 7-6	− 0.00	− 0.99 to 0.98	0.996
Group [Experimental]	0.25	− 0.59 to 1.09	0.564
Semester 2-1: GroupExperimental	− 0.69	− 1.82 to 0.44	0.231
Semester 3-2: GroupExperimental	0.13	− 1.11 to 1.37	0.837
Semester 4-3: GroupExperimental	1.74	0.36 to 3.12	**0.013**
Semester 5-4: GroupExperimental	− 1.53	− 2.98 to − 0.09	**0.038**
Semester 6-5: GroupExperimental	1.38	− 0.07 to 2.84	0.062
Semester 7-6: GroupExperimental	− 1.03	− 2.53 to 0.47	0.178
Random Effects			
σ^2	6.23		
τ_{00} ident	7.48		
ICC	0.55		
N_{ident}	226		
Observations	849		
Marginal R^2/Conditional R^2	0.017/0.554		

Bold is used to indicate statistically significant predictors

4.2.11 Confidence ratings

We also analyzed confidence ratings separately. To that end, we fitted a linear mixed-effects model for each scenario, in a way analogous to the previous analyses. Instead of using the combined score, we entered raw confidence ratings as a dependent variable. Table 13 presents the results for all scenarios. The overall pattern that can be seen in Fig. 11 is that philosophy students generally had slightly less confidence in their answers. In 9 out of the 12 analyzed questions, the effect of group (philosophy vs. cognitive science students) was statistically significant.

4.2.12 Attrition

We wanted to see whether students whose intuitions did not conform to the textbook consensus were more likely than their colleagues to withdraw from the philosophy program and, thus, to drop out of our study. To this end, we took all the responses

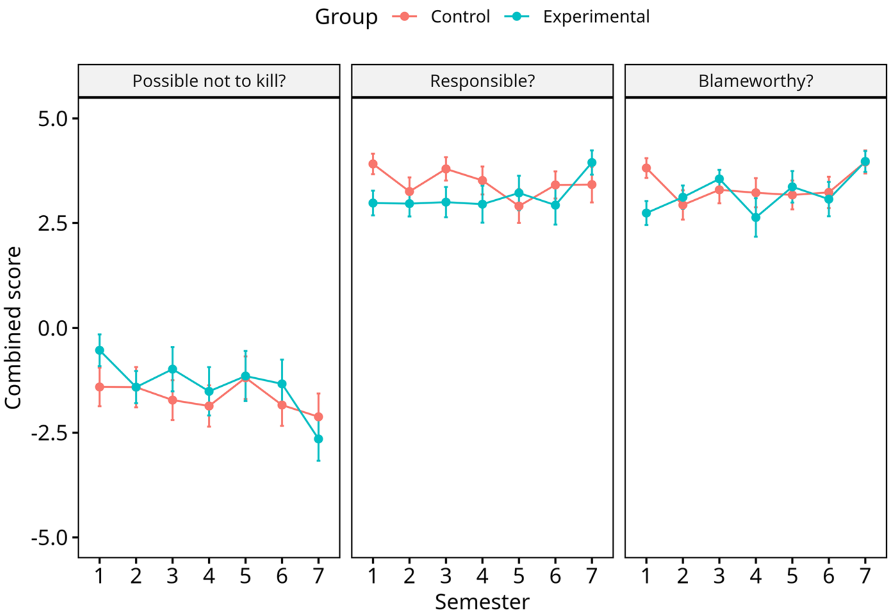

Fig. 9 Changes in intuitions about the Frankfurt case over time. Three panes of the plot correspond to the three questions that were asked. The combined score of + 5 represents a positive answer to a given question with maximum confidence and − 5 represents a negative answer with maximum confidence. Error bars correspond to the standard error of the mean

at the first measurement point and compared the answers given by participants who later completed the questionnaire at the third measurement point with those given by participants who did not. The third measurement point was chosen because it nicely splits the sample into two groups of comparable size. If students whose intuitions matched the literature consensus are more likely to become academically trained philosophers, we would expect differences at this stage. The results are presented in Table 14. The overall pattern of the responses is that there are no statistically significant differences between those two groups.

4.2.13 Analyses on a reduced dataset

To check the robustness of our findings, we decided to re-run the main part of the analysis (linear mixed-effects model with combined scored as a DV) on a reduced dataset. This dataset contains observations only for those participants who successfully completed the questionnaire at least six out of the seven times, and if one of the measurement points was missing, it came from either the sixth or seventh semester. The idea behind this analysis is that it enables us to completely eliminate the effect of selection bias, although at the cost of lower sample size and reduced statistical power.[7] The full models and plots can be found in Appendix 5. Here, we only summarize the main findings.

[7] This analysis was suggested by the anonymous reviewer. We are grateful for this idea as we think that it improve the overall soundness of the analytical approach.

Table 11 Linear mixed-effects models for the combined scores in three questions about the Frankfurt case

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	− 1.67	− 2.22 to − 1.11	**< 0.001**	3.49	3.12 to 3.86	**< 0.001**	3.38	3.05 to 3.72	**< 0.001**
Semester 2-1	0.16	− 0.82 to 1.14	0.752	− 0.75	− 1.48 to − 0.02	**0.044**	− 0.89	− 1.60 to − 0.19	**0.013**
Semester 3-2	− 0.41	− 1.43 to 0.60	0.425	0.61	− 0.15 to 1.36	0.115	0.37	− 0.36 to 1.11	0.318
Semester 4-3	− 0.01	− 1.09 to 1.06	0.984	− 0.36	− 1.16 to 0.44	0.382	− 0.15	− 0.94 to 0.63	0.704
Semester 5-4	0.53	− 0.57 to 1.62	0.345	− 0.71	− 1.52 to 0.11	0.089	− 0.12	− 0.92 to 0.67	0.760
Semester 6-5	− 0.76	− 1.86 to 0.35	0.181	0.54	− 0.28 to 1.37	0.196	0.07	− 0.74 to 0.87	0.869
Semester 7-6	− 0.28	− 1.46 to 0.90	0.642	0.00	− 0.88 to 0.88	0.995	0.74	− 0.12 to 1.60	0.092
Group [Experimental]	0.48	− 0.32 to 1.28	0.238	− 0.41	− 0.95 to 0.12	0.129	− 0.29	− 0.79 to 0.20	0.243
Semester 2-1: Group-Experimental	− 0.88	− 2.22 to 0.45	0.195	0.78	− 0.22 to 1.78	0.125	1.31	0.34 to 2.28	**0.008**
Semester 3-2: Group-Experimental	0.93	− 0.54 to 2.41	0.215	− 0.59	− 1.68 to 0.51	0.293	0.01	− 1.06 to 1.08	0.986
Semester 4-3: Group-Experimental	− 0.30	− 1.95 to 1.36	0.726	0.05	− 1.18 to 1.28	0.934	− 0.81	− 2.01 to 0.39	0.186
Semester 5-4: Group-Experimental	− 0.33	− 2.06 to 1.40	0.709	1.12	− 0.17 to 2.41	0.089	0.81	− 0.45 to 2.07	0.208
Semester 6-5: Group-Experimental	0.62	− 1.12 to 2.36	0.485	− 0.83	− 2.12 to 0.47	0.211	− 0.38	− 1.65 to 0.88	0.551
Semester 7-6: Group-Experimental	− 1.24	− 3.04 to 0.56	0.176	1.03	− 0.31 to 2.36	0.133	0.18	− 1.13 to 1.49	0.788
Random Effects									

Table 11 (continued)

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
σ^2	9.05			5.04			4.83		
$\tau_{00\ \mathrm{ident}}$	5.60			2.20			1.71		
ICC	0.38			0.30			0.26		
N_{ident}	226			226			225		
Observations	850			846			848		
Marginal R^2/Conditional R^2	0.018/0.393			0.023/0.320			0.024/0.279		

Bold is used to indicate statistically significant predictors

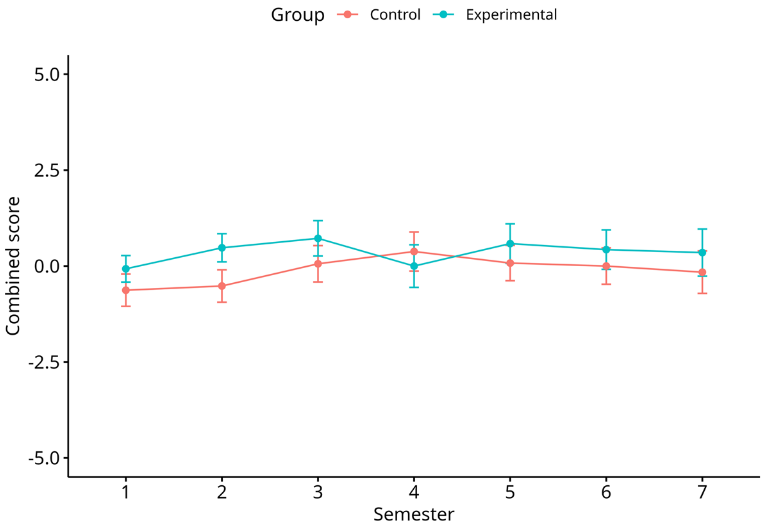

Fig. 10 Changes in intuitions about the Teleportation case over time. The combined score of + 5 represents an answer that neither copy was the original with maximal confidence and − 5 represents the opposite view (other answers) with maximal confidence. Error bars correspond to the standard error of the mean

For the Gettier case, in the philosophy group we found a clear drop in combined scores between the first and the second semester (see Fig. 23), which is indicated by a statistically significant interaction ($b = -2.28$, $p = 0.010$; see Table 28). This is congruent with the results of the original analysis. In the Fake Barns case, we observed an almost identical pattern of responses to the original analysis (see Fig. 24), with a very large drop in scores between the first and second semesters in the philosophy group ($b = -2.36, p = 0.013$). The overall difference between groups remained significant ($b = -2.79$, $p < 0.001$; see Table 29). For the Truetemp case, all the effects that were found in the main analysis remained significant, except for the overall difference between the two groups (see Fig. 25; Table 30).

For the Knobe harm case, all the effects that reached statistical significance remained such when the analysis was re-run on the reduced dataset (see Table 31). The overall pattern of the responses also did not change (see Fig. 26). In the Twin Earth case, the only effect that was statistically significant in the original analysis was the change between the first and second semesters in the direction opposite to philosophical orthodoxy. In the analysis on the reduced dataset, this effect ceased to be statistically significant ($b = 0.90$, $p = 0.090$, see Table 32). It must be noted, however, that the regression coefficient is virtually the same ($b = 0.98$ vs. 0.90, see Fig. 27) and we think that the fact that it did not reach the level of statistical significance is the consequence of reducing statistical power by limiting the number of observations. For the Gödel/Schmidt case, the overall shape of the results stayed the same (see Fig. 28), but the significance of the individual predictors changed a little bit (see Table 33). The large change in the intuitions of the philosophy students between the second and third semesters was not significant in the original analysis ($b = 0.47$, $p = 0.539$), probably due to a similar trend in the sample of cognitive science students. In the reduced dataset,

Table 12 Linear mixed-effects model for the combined scores in the Teleportation case

Predictors	Combined score (Teleportation)		
	Estimates	CI	p
(Intercept)	0.04	– 0.51 to 0.60	0.878
Semester 2-1	0.05	– 0.80 to 0.90	0.907
Semester 3-2	0.49	– 0.39 to 1.37	0.276
Semester 4-3	0.31	– 0.62 to 1.24	0.518
Semester 5-4	– 0.34	– 1.28 to 0.61	0.487
Semester 6-5	0.21	– 0.75 to 1.17	0.664
Semester 7-6	– 0.06	– 1.08 to 0.97	0.912
Group [Experimental]	0.25	– 0.55 to 1.04	0.542
Semester 2-1: GroupExperimental	0.62	– 0.55 to 1.79	0.298
Semester 3-2: GroupExperimental	– 0.29	– 1.57 to 1.00	0.661
Semester 4-3: GroupExperimental	– 0.88	– 2.32 to 0.55	0.226
Semester 5-4: GroupExperimental	0.60	– 0.91 to 2.10	0.437
Semester 6-5: GroupExperimental	– 0.21	– 1.72 to 1.30	0.786
Semester 7-6: GroupExperimental	0.04	– 1.52 to 1.60	0.961
Random Effects			
σ^2	6.76		
τ_{00} ident	6.17		
ICC	0.48		
N_{ident}	225		
Observations	849		
Marginal R^2/Conditional R^2	0.008/0.482		

Bold is used to indicate statistically significant predictors

this tendency is not present, and a large shift towards orthodoxy in the sample of philosophy students is now statistically significant ($b = 2.29$, $p = 0.030$). Interestingly, whereas in the original analysis no differences between groups were observed for the Experience Machine case, in the reduced dataset we can observe a clear and consistent difference (see Fig. 29)—philosophy students are much less likely to wholeheartedly decide to remain in the real world ($b = -1.35$, $p = 0.033$; see Table 34). In the Violinist case, the main effect that we found in the original analysis was a statistically significant interaction indicating change in intuitions between the third and fourth semesters in the philosophy group. This finding was replicated in the analysis on a reduced dataset ($b = 2.48$, $p = 0.004$, see Fig. 30; Table 35).

In the original analysis, we did not find any statistically significant predictor for the *Possible not to kill?* question to the Frankfurt case. In the reduced dataset, we observed a change between the first and second semesters towards the positive response ($b = 1.55$, $p = 0.015$; see Table 36), which is mainly driven by the control group. For the *Responsible?* question in the original analysis we observed a trend towards negative

Table 13 Linear mixed-effects model for the confidence scores in all tested cases

Predictors	Gettier (confidence)			Fake Barns (confidence)			Truetemp (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	3.96	3.83 to 4.09	**< 0.001**	3.98	3.84 to 4.11	**< 0.001**	3.86	3.72 to 4.00	**< 0.001**
Semester 2-1	0.26	0.01 to 0.52	**0.042**	0.04	– 0.25 to 0.32	0.794	0.25	– 0.02 to 0.53	0.071
Semester 3-2	– 0.06	– 0.33 to 0.20	0.634	– 0.12	– 0.42 to 0.18	0.424	– 0.08	– 0.37 to 0.20	0.574
Semester 4-3	– 0.03	– 0.31 to 0.26	0.853	0.01	– 0.31 to 0.32	0.969	0.04	– 0.27 to 0.34	0.812
Semester 5-4	– 0.14	– 0.42 to 0.15	0.343	– 0.04	– 0.36 to 0.28	0.814	– 0.12	– 0.43 to 0.19	0.452
Semester 6-5	0.13	– 0.16 to 0.42	0.373	0.12	– 0.21 to 0.44	0.487	0.02	– 0.30 to 0.33	0.907
Semester 7-6	– 0.13	– 0.44 to 0.18	0.423	– 0.06	– 0.41 to 0.28	0.723	0.03	– 0.30 to 0.37	0.854
Group [Experimental]	0.07	– 0.12 to 0.26	0.449	– 0.39	– 0.59 to – 0.19	**< 0.001**	– 0.21	– 0.42 to – 0.00	**0.046**
Semester 2-1: GroupExperimental	– 0.22	– 0.57 to 0.13	0.209	– 0.20	– 0.59 to 0.19	0.306	– 0.21	– 0.59 to 0.17	0.282
Semester 3-2: GroupExperimental	0.25	– 0.13 to 0.64	0.197	0.30	– 0.13 to 0.73	0.171	0.28	– 0.13 to 0.70	0.183
Semester 4-3: GroupExperimental	0.12	– 0.31 to 0.55	0.592	– 0.03	– 0.51 to 0.46	0.915	– 0.17	– 0.64 to 0.30	0.471
Semester 5-4: GroupExperimental	0.04	– 0.42 to 0.49	0.869	– 0.26	– 0.77 to 0.25	0.317	0.15	– 0.34 to 0.64	0.557
Semester 6-5: GroupExperimental	– 0.08	– 0.54 to 0.37	0.722	– 0.01	– 0.52 to 0.50	0.975	– 0.00	– 0.50 to 0.49	0.985

Table 13 (continued)

Predictors	Gettier (confidence)			Fake Barns (confidence)			Truetemp (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Semester 7-6: GroupExperimental	0.18	− 0.29 to 0.65	0.461	0.22	− 0.31 to 0.75	0.413	0.07	− 0.44 to 0.58	0.781
Random Effects									
σ^2	0.62			0.79			0.73		
$\tau_{00\ \mathrm{ident}}$	0.28			0.28			0.33		
ICC	0.31			0.26			0.31		
N_{ident}	226			226			226		
Observations	851			851			849		
Marginal R^2/Conditional R^2	0.014/0.322			0.040/0.291			0.020/0.323		

Predictors	Knobe (confidence)			Twin Earth (confidence)			Gödel/Schmidt (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	3.99	3.85 to 4.14	*< 0.001*	4.40	4.27 to 4.53	**< 0.001**	4.05	3.91 to 4.20	**< 0.001**
Semester 2-1	− 0.11	− 0.39 to 0.17	0.443	− 0.38	− 0.63 to − 0.13	**0.003**	0.13	− 0.17 to 0.42	0.395
Semester 3-2	0.01	− 0.28 to 0.30	0.958	0.02	− 0.24 to 0.28	0.870	− 0.07	− 0.38 to 0.23	0.649
Semester 4-3	0.18	− 0.13 to 0.49	0.259	0.08	− 0.20 to 0.36	0.563	0.08	− 0.24 to 0.41	0.623

Table 13 (continued)

Predictors	Knobe (confidence)			Twin Earth (confidence)			Gödel/Schmidt (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Semester 5-4	− 0.02	− 0.34 to 0.29	0.880	− 0.22	− 0.50 to 0.06	0.130	− 0.02	− 0.35 to 0.31	0.928
Semester 6-5	0.04	− 0.28 to 0.36	0.809	0.09	− 0.20 to 0.37	0.559	0.01	− 0.33 to 0.34	0.973
Semester 7-6	− 0.26	− 0.60 to 0.07	0.126	0.10	− 0.21 to 0.40	0.533	0.11	− 0.25 to 0.47	0.545
Group [Experimental]	− 0.39	− 0.60 to − 0.19	**< 0.001**	− 0.35	− 0.54 to − 0.16	**< 0.001**	− 0.49	− 0.70 to − 0.28	**< 0.001**
Semester 2-1: GroupExperimental	− 0.06	− 0.44 to 0.32	0.756	− 0.06	− 0.40 to 0.28	0.728	− 0.33	− 0.74 to 0.07	0.102
Semester 3-2: GroupExperimental	0.12	− 0.30 to 0.54	0.575	0.09	− 0.29 to 0.47	0.630	0.30	− 0.15 to 0.74	0.189
Semester 4-3: GroupExperimental	− 0.10	− 0.57 to 0.38	0.688	− 0.11	− 0.54 to 0.32	0.611	− 0.31	− 0.81 to 0.19	0.223
Semester 5-4: GroupExperimental	− 0.02	− 0.52 to 0.48	0.936	0.30	− 0.15 to 0.74	0.195	0.09	− 0.44 to 0.61	0.742
Semester 6-5: GroupExperimental	− 0.15	− 0.64 to 0.35	0.564	− 0.17	− 0.62 to 0.28	0.462	0.24	− 0.29 to 0.76	0.371
Semester 7-6: GroupExperimental	0.65	0.14 to 1.17	**0.013**	− 0.07	− 0.53 to 0.39	0.766	0.04	− 0.51 to 0.58	0.894
Random Effects									
σ^2	0.74			0.61			0.83		

Table 13 (continued)

Predictors	Knobe (confidence)			Twin Earth (confidence)			Gödel/Schmidt (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
$\tau_{00\ \text{ident}}$	0.34			0.28			0.31		
ICC	0.32			0.32			0.27		
N_{ident}	226			226			225		
Observations	850			851			850		
Marginal R^2/Conditional R^2	0.048/0.348			0.057/0.359			0.061/0.315		
Predictors	Experience Machine (confidence)			Violinist (confidence)			Frankfurt—Possible..? (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	4.10	3.90 to 4.29	**< 0.001**	3.88	3.69 to 4.06	**< 0.001**	4.19	4.04 to 4.33	**< 0.001**
Semester 2-1	− 0.05	− 0.35 to 0.25	0.751	− 0.10	− 0.40 to 0.20	0.516	0.04	− 0.24 to 0.32	0.790
Semester 3-2	0.05	− 0.26 to 0.36	0.753	0.03	− 0.28 to 0.35	0.844	− 0.13	− 0.42 to 0.16	0.373
Semester 4-3	0.09	− 0.24 to 0.42	0.579	0.21	− 0.13 to 0.54	0.226	− 0.12	− 0.43 to 0.19	0.446
Semester 5-4	− 0.13	− 0.47 to 0.20	0.429	0.02	− 0.32 to 0.36	0.925	0.09	− 0.22 to 0.40	0.565

Table 13 (continued)

Predictors	Experience Machine (confidence)			Violinist (confidence)			Frankfurt—Possible..? (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Semester 6-5	− 0.04	− 0.38 to 0.30	0.809	0.19	− 0.16 to 0.53	0.281	− 0.03	− 0.35 to 0.28	0.843
Semester 7-6	− 0.08	− 0.44 to 0.28	0.665	− 0.41	− 0.77 to − 0.04	**0.031**	0.28	− 0.06 to 0.62	0.105
Group [Experimental]	− 0.16	− 0.44 to 0.11	0.252	− 0.21	− 0.48 to 0.06	0.128	− 0.43	− 0.64 to − 0.21	**< 0.001**
Semester 2-1: GroupExperimental	0.17	− 0.25 to 0.58	0.429	− 0.08	− 0.50 to 0.34	0.716	− 0.05	− 0.43 to 0.33	0.794
Semester 3-2: GroupExperimental	0.13	− 0.32 to 0.59	0.564	− 0.01	− 0.47 to 0.45	0.976	0.45	0.03 to 0.87	**0.037**
Semester 4-3: GroupExperimental	− 0.26	− 0.77 to 0.24	0.310	0.13	− 0.38 to 0.65	0.615	− 0.05	− 0.53 to 0.42	0.830
Semester 5-4: GroupExperimental	− 0.15	− 0.68 to 0.38	0.569	− 0.08	− 0.62 to 0.46	0.771	− 0.09	− 0.58 to 0.41	0.729
Semester 6-5: GroupExperimental	0.51	− 0.02 to 1.04	0.058	− 0.10	− 0.64 to 0.44	0.723	0.08	− 0.41 to 0.58	0.742
Semester 7-6: GroupExperimental	0.11	− 0.44 to 0.66	0.699	0.43	− 0.13 to 0.98	0.136	− 0.17	− 0.68 to 0.35	0.518
Random Effects									
σ^2	0.84			0.87			0.74		
$\tau_{00\ \mathrm{ident}}$	0.74			0.67			0.38		
ICC	0.47			0.44			0.34		
N_{ident}	226			226			226		
Observations	848			850			850		

Table 13 (continued)

Predictors	Experience Machine (confidence)			Violinist (confidence)			Frankfurt—Possible..? (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Marginal R²/Conditional R²	0.014/0.476			0.023/0.449			0.056/0.377		

Predictors	Frankfurt—Responsible? (confidence)			Frankfurt—Blameworthy? (confidence)			Teleportation (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	*4.32*	*4.17 to 4.46*	*< 0.001*	4.16	4.00 to 4.31	**< 0.001**	3.58	3.41 to 3.76	**< 0.001**
Semester 2-1	− 0.13	− 0.40 to 0.13	0.322	− 0.16	− 0.45 to 0.12	0.258	− 0.16	− 0.49 to 0.16	0.324
Semester 3-2	0.11	− 0.17 to 0.38	0.434	0.01	− 0.29 to 0.30	0.953	0.24	− 0.10 to 0.58	0.160
Semester 4-3	− 0.13	− 0.42 to 0.17	0.400	− 0.04	− 0.35 to 0.27	0.803	− 0.02	− 0.38 to 0.34	0.930
Semester 5-4	− 0.06	− 0.35 to 0.24	0.706	− 0.08	− 0.40 to 0.24	0.636	− 0.20	− 0.57 to 0.17	0.286
Semester 6-5	− 0.12	− 0.42 to 0.18	0.443	0.24	− 0.08 to 0.56	0.144	− 0.11	− 0.48 to 0.27	0.576
Semester 7-6	0.49	0.17 to 0.81	**0.003**	0.14	− 0.20 to 0.49	0.413	0.45	0.05 to 0.84	**0.026**
Group [Experimental]	− 0.32	− 0.53 to − 0.11	**0.003**	− 0.31	− 0.53 to − 0.08	**0.008**	− 0.34	− 0.60 to − 0.09	**0.008**
Semester 2-1: Group-Experimental	0.13	− 0.23 to 0.49	0.484	0.33	− 0.06 to 0.72	0.100	0.10	− 0.35 to 0.54	0.674

Table 13 (continued)

Predictors	Frankfurt—Responsible? (confidence)			Frankfurt—Blameworthy? (confidence)			Teleportation (confidence)		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Semester 3-2: Group-Experimental	− 0.12	− 0.52 to 0.28	0.570	− 0.13	− 0.56 to 0.30	0.559	− 0.07	− 0.57 to 0.42	0.771
Semester 4-3: Group-Experimental	0.10	− 0.34 to 0.55	0.649	− 0.05	− 0.53 to 0.44	0.849	0.03	− 0.52 to 0.59	0.912
Semester 5-4: Group-Experimental	0.22	− 0.25 to 0.69	0.367	0.36	− 0.14 to 0.87	0.160	0.13	− 0.45 to 0.71	0.656
Semester 6-5: Group-Experimental	0.15	− 0.33 to 0.62	0.543	− 0.33	− 0.84 to 0.18	0.202	− 0.03	− 0.62 to 0.55	0.910
Semester 7-6: Group-Experimental	− 0.38	− 0.87 to 0.11	0.126	− 0.01	− 0.53 to 0.52	0.982	− 0.01	− 0.61 to 0.59	0.971
Random Effects									
σ^2	0.67			0.77			1.02		
$\tau_{00\ ident}$	0.37			0.44			0.52		
ICC	0.36			0.36			0.34		
N_{ident}	226			225			225		
Observations	847			848			849		
Marginal R^2/Conditional R^2	0.040/0.384			0.031/0.383			0.031/0.359		

Bold is used to indicate statistically significant predictors

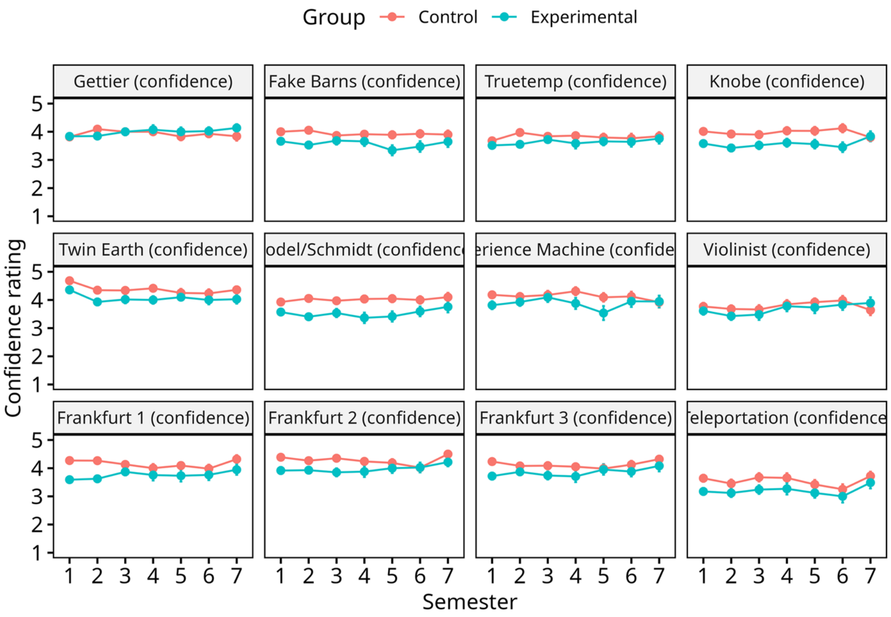

Fig. 11 Confidence levels for two groups of subjects for all questions. Frankfurt 1 refers to "Possible not to kill?" question, Frankfurt 2 to "Responsible?" question and Frankfurt 3 to "Blameworthy?" question

responses between the first and second semesters, but in the re-run analysis no effects reached statistical significance. For the *Blameworthy?* question about the Frankfurt case, in the original analysis, we found change in intuitions in both groups but in the opposite direction between the first and second semesters. This pattern is still present in the reduced dataset, but in a much weaker form that did not reach statistical significance ($b = -0.71, p = 0.135$; interaction: $b = 0.94, p = 0.199$; see Table 36; Fig. 31).

The original analysis of answers to the Teleportation Case did not reveal any statistically significant predictors. However, in the analysis on the reduced dataset, we found two statistically significant effects. First, we observed a tendency to go towards philosophical orthodoxy between the third and fourth semesters in the control group ($b = 1.06$, $p = 0.046$) and a trend in the opposite direction in the group of philosophy students, which was indicated by a statistically significant interaction ($b = -1.89$, $p = 0.023$, see Table 37; Fig. 32).

Overall, additional analyses on the reduced dataset support the robustness of the findings of the main analyses. The main findings and the general pattern of responses remained largely the same.

4.3 Discussion

We found statistically significant changes in the intuitions of philosophy students in six out of the ten thought experiments we tested. Some of these changes can be

Table 14 Combined scores at the first measurement point for participants who completed the third semester (*Continued*) and those who dropped out of the study (*Dropped*)

Scenario	Continued			Dropped			t	df	p
	n	M	SD	n	M	SD			
Gettier	46	– 1.11	3.50	52	– 2.33	3.49	1.72	94.46	0.089
Fake Barns	46	0.93	3.70	52	0.92	3.75	0.02	94.81	0.988
Truetemp	46	0.38	3.63	52	– 0.38	3.75	1.02	93.82	0.312
Knobe	46	– 0.63	3.46	52	0.19	3.98	– 1.09	95.97	0.270
Twin Earth	46	– 3.63	2.43	52	– 4.19	1.74	1.30	80.32	0.197
Göodel/Schmidt	46	– 0.78	3.54	52	– 1.77	3.51	1.38	94.36	0.170
Experience Machine	46	3.20	2.38	52	3.39	2.43	– 0.39	92.92	0.697
Violinist	46	– 0.85	3.81	52	– 0.94	3.78	0.12	94.39	0.902
Frankfurt—Responsible?	46	3.05	2.56	52	2.92	3.13	0.21	92.99	0.833
Frankfurt—Blameworthy?	46	2.77	2.60	52	2.71	3.01	0.11	93.94	0.915
Frankfurt—Possible…?	46	– 1.02	3.52	52	– 0.10	3.94	– 1.23	95.99	0.223
Teleportation	46	0.20	3.15	52	– 0.31	3.67	0.73	95.92	0.466

directly connected to the classes the students were required to take at a particular time. First of all and most strikingly, we observed a massive change in intuitions regarding the Gettier case and the Fake Barns case after the first semester—which is to say, after both were covered extensively in epistemology. The changes were in the direction of philosophical orthodoxy. Moreover, a brief survey among lecturers in epistemology has confirmed that, in their opinion, philosophy students' judgments about knowledge as true justified belief change after discussing Gettier's examples. Less pronounced changes occurred in participants' responses to the Violinist case after obligatory courses in ethics, but they did not persist. In the philosophy group, we observed an increase in non-Thomsonian intuitions between the third and the fourth measurements, which then bounced back to the previous level at the beginning of the fifth semester. Although the change took an unexpected direction, we think it can be linked to the ethics course. The second change that can be related to the participants' taking of an ethics course was a slight change in responses to the Frankfurt case. Philosophy students after this course tended to be more confident that it was impossible for Frank not to kill Furt. At the same time, they tended to agree more decisively that he was responsible for Furt's death. This result is in line with the original interpretation of this case given by Frankfurt. These findings suggest that, at least in some cases, professional training has an influence on philosophical intuitions but in certain cases the change does not last.

We found little to no changes in judgments about cases that were not directly discussed in class. Importantly, the pattern of responses did not depend on the subject being taught. For example, taking epistemology did not affect the students' intuitions about a wide range of thought experiments concerning knowledge. This is best illustrated by contrasting intuitions about the Fake Barns case, which is discussed in the epistemology course, and the Truetemp case, which is not. In the Fake Barns case, we observed a large change in intuitions in the philosophy students, whereas in the Truetemp case the change was very small compared to the Gettier and Fake Barns cases, and a statistically significant effect of interaction could not be straightforwardly attributed to the philosophy students' correction of intuitions towards orthodoxy. This suggests a limited carryover of the effect of philosophical training even if we consider cases from the same philosophical subdiscipline.

With regard to two cases we do not have clear explanations, but we would like to offer tentative ones. First is the Gödel/Schmidt case, where we observed a change in intuitions in the direction of philosophical orthodoxy earlier than expected. Recall that the first time that our participants could encounter this case was during ontology classes, taken in the third and the fourth semesters. Unexpectedly, the most significant change in the philosophy group was observed after the first two semesters. Indeed, while the participants were, on average, descriptivists (mean score below midpoint) at the second measurement point, they were pretty firmly in the Kripkean camp at third. Note that, in the control group, this change occurred after the fourth semester, which is perfectly consistent with the change occurring as the result of the students' taking philosophy of language in the second year. We think that this result could be related to the fact that this example was discussed during the logic course, which the students took in the first two semesters. Unfortunately our data on exposure to cases starts from the fourth semester and because of that we were unable to confirm this hypothesis. The second problematic case was the Twin Earth scenario. We observed a weak but statistically significant trend toward agreement with the statement that XYZ is water. However, a closer look at the data reveals that participants' judgment about the case did not shift—what changed was their confidence in the answers. Because the effect was present in both groups, we think that it might be a reflection of the general critical attitude and cautiousness developed during a ternary education. However, it is difficult to square this explanation with the fact that a similar drop in confidence between the first and second measurement points did not occur in judgments associated with any other scenario.

In two cases (Knobe harm and Fake Barns), we found a considerable initial difference in intuitions between philosophy and cognitive science students. In the Fake Barns case, this difference increased at the second measurement point, but in the Knobe case, it remained stable. This finding indicates that there may be some peculiarities regarding the cognitive profile of people who decided to study philosophy at an academic level. It is interesting to note that, for the Knobe harm case, the side-effect effect was stronger for philosophy students than for students of cognitive science—they were more likely to attribute intentionality of the side effect to the protagonist's action. One possible explanation of this finding is that we used a variation of the harm vignette that was not extensively tested.[8] That may have resulted in an unexpected pattern of responses.

Another interesting finding of our study is that philosophy students display generally lower confidence ratings compared to cognitive science students. Two possible explanations should be considered. First, individuals who choose the philosophy program exhibit a different cognitive profile compared to the general population. They might be more cautious and intellectually humble, which is why they have more doubts about their judgments on tested cases. Second, for philosophy students, the stories and the questions *matter* because they concern problems relevant to the field of their study; by contrast, students of cognitive science may regard the scenarios and the probes as irrelevant puzzles.

4.4 Objections and limitations

The presented study has several limitations. First, our control group was not perfect. As we have mentioned, students of cognitive science at University of Warsaw do have some exposure to philosophy. They are, inter alia, required to take courses in philosophy of language and philosophy of mind (which add up to a total of 150 h of compulsory philosophy classes during their studies). Another problem is that at least some of those students may have developed an interest in philosophy in general—the Program of Cognitive Sciences at the University of Warsaw is generally considered in the Polish academic community to be rather philosophy-heavy. Nonetheless, we think that their responses provide a reasonable baseline for the analysis of how the intuitions of philosophy students changed over time.

A second limitation of the study is that we had no direct control over which cases were discussed during classes by different instructors and how they were discussed. Many things might depend on the teaching style of an individual instructor and on the subject-matter of the course. Some instructors may have encouraged students to challenge the textbook consensus whereas others may have been more focused on explaining the thought experiments in a way that promoted intuitions associated with mainstream views. We feel that the former approach might be more widely adopted in courses in ethics and the latter in epistemology, where there is strong community-wide agreement about certain thought experiments. After the study, we conducted an informal survey with the instructors about this matter. Epistemology lecturers unanimously declared that they took pains to make sure that the students understood the Gettier and False Barn cases. Ontology lecturers made a similar declaration about Putnam's Twin Earth thought experiment. As to the other relevant courses, the lecturers reported that, while they had discussed the thought experiments in class, they did not expect the students to acquire a thorough, in-depth understanding of them.

[8] Although Bochyńska (2021) used exactly the same scenario as we did, her respondents were asked to ascribe intentionality using a different Polish equivalent of the word 'intentionally'. In Polish, there are several non-equivalent possible translations of the English 'intentionality'. The adjective we used ('umyślnie') is negatively valenced (cf. Kuś & Maćkiewicz, 2021a; Kuś & Maćkiewicz, 2021b), so perhaps it works well in Knobe's original scenarios, but is too strong to describe actions such as making an aunt feel bad because of not being able to attend a surprise birthday party.

One may also raise concerns about the representativeness of our sample. The question is how representative of philosophical training in general is the training offered at the University of Warsaw. The structure of the undergraduate program is typical of European universities with the focus divided between contemporary analytic and continental philosophy, on the one hand, and history of philosophy, on the other. That being said, it is possible that a different educational approach with more electives, characteristic of American and British universities, might yield different results. However, we suspect that given the narrow scope of typical elective courses, the pattern of response that we obtained—namely, that most changes in case judgments are course-driven—would still hold. Nevertheless, given the pioneering nature of the present study, the generalizability of its findings cannot be assessed right away and requires further empirical investigation.

5 Philosophical implications

Our study addresses the question of how case judgments made by philosophy students evolve over time compared to the judgments made by subjects from the control group. This means that, although we can attribute observed differences in case judgments to differences in the curriculum, we cannot establish the further claim that those differences in case judgments are the reflection of a developing philosophical expertise. Instead, we have to introduce this claim as a working assumption. Accordingly, in the first part of what follows, we assume that philosophical studies give rise to the kind of cognitive skills required by the expertise defense. This will allow us to evaluate the three models of philosophical expertise described in Sect. 2.2, but our conclusions will only be conditional: if the expertise assumption is true, then our data support some hypotheses about the influence of expertise on case judgments while undermining others. But, naturally, we will still be left with an answered question about the truth of the expertise assumption. Although it would be impossible, at this stage, to address it in a fully satisfactory manner, we will provide a provisional answer to it based on data from both our study and existing cross-sectional research.

Assuming that formal training in philosophy leads to the development of cognitive skills that improve the ability to make credible case judgments, the results of our study speak against two out of the three models of philosophical expertise sketched in the introduction. According to the Method Model, the student masters a general method of philosophical thought experimentation applicable to any area of philosophy. This model predicts that increased proficiency at philosophical thought experimentation informs all philosophical case judgments regardless of subfield. We found no such pattern in our data. In fact, all observed changes in case judgments were restricted to specific areas of philosophy. This would seem to support the Subfield Model of philosophical expertise. However, the Subfield Model also predicts that changes in discipline-related cognitive skills affect all case judgments in the relevant subfield and our data indicate otherwise. For example, in the domain of moral philosophy, we found significant changes in responses to the Violinist and the Frankfurt case during and after the second year, when the students were required to take a two-semester course in ethics, but we observed no changes in judgments regarding the Experience

Machine. Thus, the only model that comports with our data is the Restricted Expertise Model, which predicts that cognitive skills involved in making case judgments are highly specific, affecting only some of the judgments relevant to a particular subfield or even concept.

This model is very weak, however. It would be supported even if each case judgment turned out to be affected by a separate cognitive skill. This is a problem because, intuitively, the ability to consistently make a single kind of case judgment hardly deserves the name of a skill. Since we found no persuasive evidence for a robust carryover effect—meaning that no cognitive skill acquired within a particular period seems to have had a significant impact on judgments about cases not discussed in class—we have to consider the possibility that formal training in philosophy does not improve the ability to make such judgments.

A natural strategy to handle this difficulty would be to argue for an interpretation of the data that moves beyond the simplified picture of the three models of philosophical expertise we have been assuming. This is fairly easy to do. Given how little is known about the determinants of philosophical case judgments, there are indefinitely many such interpretations to choose from. While we cannot discuss all the possibilities here, it is worth noting three kinds of moves that can be made in this connection. First, one can maintain that the cognitive skills involved in making case judgments do not correspond with the subfields of philosophy, so there may exist carryover effects that do not respect traditional subdisciplinary boundaries. For example, based on our data, one can hypothesize a causal link between a putative cognitive skillset acquired in the first year of philosophical studies and case judgments relevant to epistemology and philosophy of language: this is the period in which we observed statistically significant changes in the judgments made by philosophy students about the Gettier case, Fake Barns and also the Gödel/Schmidt case. Second, perhaps some philosophically relevant case judgments, such as those elicited by the Gödel/Schmidt and Fake Barns scenarios, are affected by expertise whereas others are not (e.g., the Knobe harm case, Teleportation and the Experience Machine). Third, it is possible that many, perhaps all, case judgments can be improved by expertise, but some of the relevant cognitive skills develop later than others, so we would observe more training-related changes if we had followed the participants of our study for a longer period of time.

Although impossible to exclude, these more complex accounts are open to the charge of being ad hoc. While we admit that one of them may eventually turn out to be true, we would argue that, at present, they all lack sufficient theoretical motivation. It would be difficult to explain why the Gödel/Schmidt intuition should be informed by cognitive abilities affecting the Gettier case and Fake Barns, but not Twin Earth. Likewise, we have no idea why only some intuitions should be affected by expertise—what could be the relevant difference between the Twin Earth and the Gödel/Schmidt scenarios, for example? The hypothesis positing delays in the development of selected cognitive skills faces similar problems.

To recapitulate, under the expertise assumption, our data undermine all but the weakest model of philosophical expertise—a model, on which specific cognitive skills developed in the course of an undergraduate program in philosophy each affect a very narrow set of case judgments. In fact, the only way to square our data with the existence of a carryover effect relies on introducing implausible hypotheses about the nature of philosophical expertise.

The weaknesses of an expertise-based account of our findings suggest that perhaps a better explanation of the data is possible that does not appeal to the assumption that variation in the curriculum influences case judgments via acquired cognitive skills. We believe that there is such an alternative explanation that fits well with the data, though it cannot account for all our observations. This alternative explanation says that most of the changes we have observed did not result from the students' deploying new cognitive skills, but from the fact that they simply adopted specific beliefs endorsed by their teachers. This is not to say that philosophy instructors are bent on preserving textbook consensus or indoctrinating their students. Rather, in many classes, the student is required to know the canonical analysis and interpretation of certain thought experiments, and, having learned what they are, may simply adopt the corresponding beliefs without much deliberation. In sum, in light of our data, it is a plausible supposition that, when it comes to making case judgments elicited by philosophical thought experiments, professional philosophers do not have any special skills distinguishing them from laypeople. The significant difference between the two populations is that philosophers have accepted the "standard" interpretation of a number of philosophical thought experiments whereas the folk have not.

Besides accounting well for our data, this hypothesis seems simpler and more conservative than its expertise-based competition. It is not ad hoc, since the mechanism it invokes is familiar and well-established in psychology and social science. Furthermore, as things stand now, it meshes well with the findings of existing cross-sectional research on philosophical expertise. As we saw in Sect. 3, available cross-sectional studies indicate that professional philosophers asked to make case judgments are susceptible to many of the same biases as the folk. This suggests that, different though they may be, case judgments made by professional philosophers are by no means superior to those made by ordinary people.

Acknowledgements We would like to thank Ethan Landes and Remigiusz Kawala for insightful ideas at the initial stage of study preparation, and two anonymous reviewers at *Synthese* for comments and criticisms on earlier draft. We also appreciate helpful feedback on this work from the participants of European Online XPhi Conference 2020 and XIV Philosophers' Rally 2018. Special thanks go to all the philosophy and cognitive science students at the University of Warsaw, without whom our research would not have been possible.

Funding The Project has been financed by a Grant (2018/29/N/HS1/02863) from the National Science Centre, Poland.

Declarations

Conflict of interest The author declares that there is no conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval Ethical clearance for all procedures used in the study was obtained from the Rector's Committee for the Ethics of Research Involving Human Participants, University of Warsaw (Decision No. 21/2017).

Informed consent Written and freely-given informed consent was obtained from all individual participants included in the study. The participants' data were anonymized.

Appendix 1: Vignettes

Table 15 Vignettes used in the study

1. Amerykański samochód Bartek ma znajomą Julię. Pamięta, że Julia przez wiele lat jeździła samochodem amerykańskiej marki Buick. Sądzi więc, że Julia jeździ amerykańskim samochodem. Nie wie jednak, że Julia niedawno sprzedała ten samochód i kupiała samochód marki Pontiac. Pontiac to amerykańska marka Czy zgadzasz się ze stwierdzeniem *Bartek wie, że Julia jeździ amerykańskim samochodem*? Tak/Nie	**1. American car** Bob has a friend, Jill. He remembers that she has driven a Buick, which is an American-made car, for many years. Bob therefore thinks that Jill drives an American car. He does not know, however, that Jill has recently sold her Buick and bought a Pontiac. Pontiac is an American car brand Do you agree with the claim: *Bob knows that Jill drives an American-made car*? Yes/No
2. Stodoły Zuza wygląda przez okno samochodu i widzi niedaleko drogi stodołę. Sądzi więc, że przy drodze jest stodoła. Zuza nie zdaje sobie jednak sprawy z tego, że właśnie przejeżdża przez okolicę, w której kręcą film, i filmowcy zbudowali w tym miejscu wiele makiet stodół, które z drogi wyglądają jak prawdziwe stodoły. Ta stodoła, na którą patrzy Zuza, nie jest akurat makietą, jest jedyną prawdziwą stodołą w okolicy Czy zgadzasz się ze stwierdzeniem *Zuza wie, że przy drodze, którą jedzie, jest stodoła?* Tak/Nie	**2. Barns** Suzy looks out the car window and sees a barn not far from the road. Thus she thinks there is a barn by the road. What Suzy doesn't realize, however, is that she is just driving through an area where a movie is being shot, and the filmmakers have built many barn façades in the area, which look like real barns from the road. The barn that Suzy is looking at is not a façade, it is the only real barn in the area Do you agree with the claim: *Suzy knows that there is a barn by the road he is driving on?* Yes/No
3. John-termometr Pewnego dnia John został ogłuszony i porwany przez grupę działających w słusznej sprawie naukowców wysłanych przez starszyznę społeczności, do której należał John. Mózg Johna został przeprogramowany w taki sposób, by John zawsze prawidłowo szacował temperaturę otoczenia. John nie miał jednak zielonego pojęcia, że ktokolwiek coś zmienił w jego mózgu. Kilka tygodni później przeprogramowanie mózgu sprawiło, że John zaczął uważać, że w jego pokoju jest 21 stopni Celsjusza. Nie miał na to żadnych innych dowodów prócz swojego silnego przekonania. Skądinąd w pokoju rzeczywiście było 21 stopni Celsjusza Czy zgadzasz się ze stwierdzeniem *John wie, że w jego pokoju jest 21 stopni Celsjusza?* Tak/Nie	**3. John-thermometer** One day John is knocked out and captured by a team of well-meaning scientists sent by the elders of his community. John's brain is re-wired so that he is always absolutely right whenever he estimates the temperature where he is. John is completely unaware that his brain has been altered in this way. A few weeks later, this brain re-wiring leads John to believe that it is 21 °C in his room. He has no evidence for this other than his strong belief. In fact, it is 21 °C in his room Do you agree with the claim: *John knows that it is 21 °C in his room*? Yes/No

Table 15 (continued)

4. Teleportacja Jest 2450 rok. Cywilizacja ludzka dokonała ledwo wyobrażalnego dla nas skoku technologicznego. Derek mieszka na Ziemi, jego żona jest na Wenus, a matka na Marsie. Derek, mając dość samotności, wchodzi do teleportera kwantowego, który ma na Ziemi blisko domu, i mówi: „Chcę spotkać się i z żoną, i z mamą”. Naciska guzik. Po chwili całe jego ciało jest już przeskanowane, a informacja o strukturze komórek i stanie umysłowym zostaje wysłana tunelem czasoprzestrzennym na Wenus i na Marsa, gdzie zostaje bezbłędnie odwzorowana w formie fizycznej. Po chwili na Wenus w apartamencie żony Dereka w teleporterze pojawia się postać. Kobieta obejmuje ją z radością w oczach, mówiąc: „Mój drogi, jak miło cię widzieć!” Oboje są szczęśliwi. W tym samym czasie również na Marsie u matki Dereka z teleportera wychodzi postać. Kobieta obejmuje ją z radością w oczach, mówiąc: „Mój drogi, jak miło cię widzieć!” Dom Dereka na Ziemi jest teraz pusty Która możliwość najlepiej opisuje to, co się stało? – Derek objął swoją żonę, ale matkę objął ktoś inny – Derek objął swoją matkę, ale żonę objął ktoś inny – Ktoś inny objął żonę, ktoś inny matkę, żadnej nie obejmował Derek – Derek objął zarówno żonę, jak i matkę	**4. Teleportation** The year is 2450. Human civilization has advanced technologically so far that we could barely comprehend it. Derek is currently on Earth, his wife is on Venus and his mother is on Mars. Feeling somewhat lonely, Derek enters the quantum teleporter he has on Earth close to home, and says: “I want to visit both my wife and my mother.” Then he presses the button. In an instant, his entire body is scanned, and information about his cell structure and mental state is sent through an information wormhole to Venus and Mars, where it is perfectly reconstituted. Instantly, a figure appears in the teleporter in Derek’s wife’s apartment on Venus. The woman embraces him with joy in her eyes, saying: “My dear! How nice to see you!” They are both happy. At the same time, a figure emerges also from the teleporter at Derek’s mother’s apartment on Mars. The woman embraces him with joy in her eyes, saying: “My Dear! How nice to see you!” Derek’s home on Earth is now empty Which option best describes what happens in the story? – Derek embraced his wife but someone else embraced his mother – Derek embraced his mother but someone else embraced his wife – Someone else embraced his wife, someone else embraced his mother, none of them embraced Derek – Derek embraced his wife and Derek embraced his mother

Table 15 (continued)

5. Urodziny Ania zastanawia się z mamą nad przyjęciem niespodzianką urodzinową dla taty. Mówi: „Mamo, jeśli zorganizujemy w tę sobotę przyjęcie urodzinowe dla taty, zrobimy mu ogromną przyjemność. Tylko cioci Lusi będzie przykro, bo nie będzie mogła wtedy przyjść". Mama odpowiada: „Zupełnie nie obchodzi mnie to, czy cioci będzie przykro. Chcę jedynie zrobić przyjemną niespodziankę twojemu tacie. Zorganizujmy przyjęcie urodzinowe w tę sobotę". Przyjęcie urodzinowe zostało zorganizowane w sobotę. Zgodnie z przewidywaniami, cioci było przykro, że nie mogła przyjść Czy zgadzasz się ze stwierdzeniem: *Mama umyślnie zrobiła przykrość cioci Lusi?* Tak/Nie	**5. Birthday party** Anne is thinking with her mother about a surprise birthday party for her dad. She says: "Mom, if we organize a birthday party for dad this Saturday, we will make him very happy. Only Aunt Lucy will feel bad, because she won't be able to come then." Mom replies: "I don't care at all if Auntie will feel bad. I just want to make a pleasant surprise for your dad. Let's organize a birthday party this Saturday." The birthday party was organized on Saturday. As expected, the aunt felt bad that she couldn't come Do you agree with the claim: *Mom intentionally made Aunt Lucy feel bad?* Yes/No
6. Planeta B297A Planeta B297A do złudzenia przypomina Ziemię. Jest jednak pewna różnica – bezbarwna i bezwonna substancja, którą można znaleźć w jeziorach i rzekach i którą można pić, by ugasić pragnienie, nie składa się z dwóch atomów wodoru i jednego atomu tlenu, lecz ma inną strukturę chemiczną – XYZ Czy zgadzasz się ze stwierdzeniem: *XYZ to woda*? Tak/Nie	**6. Planet B297A** The planet B297A is strikingly similar to Earth. However, there is a difference—the colorless and odorless substance that can be found in lakes and rivers, and that can be drunk to quench thirst, does not consist of two hydrogen atoms and one oxygen atom, but has a different chemical structure—XYZ Do you agree with the claim: *XYZ is water*? Yes/No

Table 15 (continued)

7. Niezupełność arytmetyki Wyobraźmy sobie, że Jan na studiach dowiedział się, że to Gödel udowodnił jedno z najważniejszych twierdzeń matematyki, tak zwane twierdzenie o niezupełności. Jan, będąc niezgorszym matematykiem, potrafi zrekonstruować dowód, a za jego autora uważa oczywiście Gödla. Nie wie jednak nic więcej o Gödlu. A teraz wyobraźmy sobie, że Gödel nie jest w rzeczywistości autorem tego twierdzenia. Jako pierwszy dowód przeprowadził człowiek o nazwisku „Schmidt", którego ciało zostało znalezione w tajemniczych okolicznościach wiele lat temu w Wiedniu. Jego przyjaciel Gödel przywłaszczył sobie manuskrypt i przypisał sobie autorstwo dowodu. W ten sposób Gödel zapisał się na kartach historii jako twórca dowodu niezupełności arytmetyki. Większość ludzi, którzy kiedykolwiek słyszeli nazwisko „Gödel", jest jak Jan: jedyne, co słyszeli o Gödlu, to to, że jako pierwszy przeprowadził dowód twierdzenia o niezupełności O kim mówi John, gdy używa nazwiska „Gödel"? – O człowieku, który w rzeczywistości odkrył twierdzenie o niezupełności – O człowieku, który ukradł manuskrypt i przypisał sobie jego autorstwo	**7. Incompleteness of arithmetic** Suppose that John has learned in college that it was Gödel who proved one of the most important theorems of mathematics, called the incompleteness theorem. John, being quite good at mathematics, is able to reconstruct the proof, and of course considers Gödel to be its author. However, he knows nothing more about Gödel. Now let's suppose that Gödel is not the author of this theorem. The proof was first constructed by a man named Schmidt, whose body was found under mysterious circumstances many years ago in Vienna. His friend Gödel got hold of the manuscript and claimed credit for the proof. In this way, Gödel went down in history as the creator of the proof of the incompleteness of arithmetic. Most people who have ever heard the name "Gödel" are like John: all they have heard of Gödel is that he discovered the incompleteness theorem When John uses the name "Gödel," is he talking about – The person who really discovered the incompleteness of arithmetic – The person who got hold of the manuscript and claimed credit for the work?

Table 15 (continued)

8. Frank i Furt Frank ma powody, by nienawidzić Furta. Zastanawia się, czy go nie zastrzelić. Frank jednak nie wie, że ma wszczepione urządzenie, które pozwala pewnemu neuronaukowcowi monitorować pracę jego mózgu i na nią wpływać. Gdyby neuronaukowiec zaczął podejrzewać, że Frank nie zdecyduje się zabić Furta (neuronaukowiec potrafi takie rzeczy odczytać z mózgu Franka), to odpowiednio wpływając na procesy mózgowe Franka, wymusiłby na nim podjęcie decyzji o zabiciu Furta. Frank sam podejmuje jednak decyzję o tym, że zastrzeli Furta I udaje mu się to zrobić Czy zgadzasz się ze stwierdzeniem: *Frank mógł nie zabić Furta*? Tak/Nie Czy zgadzasz się ze stwierdzeniem: *Frank jest odpowiedzialny za śmierć Furta*? Tak/Nie Czy zgadzasz się ze stwierdzeniem: *Frank jest winny zabicia Furta*? Tak/Nie	**8. Frank and Furt** Frank has reason to hate Furt. He contemplates shooting him. What Frank doesn't know, however, is that he has an implanted device that allows a certain neuroscientist to monitor and influence his brain function. If the neuroscientist began to suspect that Frank decided not to kill Furt (the neuroscientist can read such things off of Frank's brain), then by appropriately influencing Frank's brain processes, he would force Frank to decide to kill Furt. Frank, however, makes the decision himself to shoot Furt. And he succeeds in doing so Do you agree with the claim: *It was possible for Frank not to kill Furt*? Yes/No Do you agree with the claim: *Frank is responsible for Furt's death*? Yes/No Do you agree with the claim: *Frank is blameworthy for killing Furt*? Yes/No
9. Skrzypek Któregoś październikowego ranka budzisz się i odkrywasz, że leżysz na szpitalnym łóżku podłączony do światowej sławy skrzypka. Skrzypek ma uszkodzone nerki, a zgodnie ze wszystkimi danymi lekarskimi jesteś jedyną osobą, która ma odpowiedni typ krwi i przeciwciał, by móc mu pomóc. Zeszłej nocy krwioobieg skrzypka został połączony z twoim tak, by twoje nerki mogły również oczyszczać jego krew. Ordynator szpitala mówi ci: „Jeśli odłączymy teraz od ciebie skrzypka, to umrze. Żeby przeżyć potrzebuje być podłączony do ciebie przez trzy kwartały. W czerwcu, kiedy skrzypek się wzmocni, będziemy mogli bezpiecznie go od ciebie odłączyć" Czy zgadzasz się ze stwierdzeniem: *Moim obowiązkiem moralnym jest pozostać podłączonym do skrzypka*? Tak Nie	**9. Violinist** You wake up one October morning and discover that you are lying on a hospital bed connected to a world-famous violinist. The violinist has kidney failure, and according to all the medical records, you are the only person who has the right blood type and antibodies to be able to help him. Last night, the violinist's circulatory system was plugged into yours, so that your kidneys could be used to extract poisons from his blood as well as your own. The director of the hospital tells you: "If we unplug the violinist from you now, he will die. To survive, he needs to be connected to you for three quarters. In June, when the violinist has recovered, we will be able to safely unplug him from you" Do you agree with the claim: *It is my moral duty to stay connected to the violinist*? Yes/No

Table 15 (continued)

10. Maszyna przyjemności Któregoś dnia słyszysz dzwonek do drzwi. W progu stoi wysoki mężczyzna w czarnym płaszczu i okularach przeciwsłonecznych. Przedstawia się jako Smith. Twierdzi, że ma dla ciebie bardzo ważną propozycję. Trochę zaniepokojony, ale ciekawy zapraszasz go do środka. „Zostałeś wskazany przez nasz system jako idealny kandydat" mówi Smith. "Możemy podłączyć twój mózg do stworzonej przez nasz zespół neuronaukowców maszyny symulującej doświadczenie. Podczas gdy twoje ciało będzie w maszynie, będziemy mogli stymulować twój mózg tak, by to, czego doświadczysz, nie różniło się jakościowo od myśli, przeżyć ani uczuć, które możesz mieć w świecie rzeczywistym. Możesz powiedzieć nam dokładnie, co chciałbyś przeżyć i osiągnąć, i zmienić swoje życie w satysfakcjonujące pasmo przyjemności. My odpowiednio zaprogramujemy maszynę, byś miał wybrane przez siebie wrażenia i przeżycia. Będziesz ze wszech miar szczęśliwy. W maszynie zapomnisz, że kiedykolwiek do niej wchodziłeś, wszystko będzie wydawało ci się rzeczywiste. Nie musisz martwić się też o rodzinę i bliskich im też zaproponujemy wejście do maszyn, będą więc mogli zadecydować o swoim życiu i szczęściu" Co byś wybrał? podłączyć się do maszyny/pozostać w świecie rzeczywistym	**10. Pleasure machine** One day you hear the doorbell ring. In the doorway stands a tall man wearing a black coat and sunglasses. He introduces himself as Smith. He claims to have an important proposal for you. Mildly troubled but still curious, you let him in. "You have been identified by our system as an ideal candidate," says Smith. "We can connect your brain to a machine created by our team of neuroscientists that simulates experience. While your body is in the machine, we will be able to stimulate your brain so that what you experience is not qualitatively different from the thoughts, experiences and feelings you might have in the real world. You can tell us exactly what you would like to experience and achieve, and change your life into one that is full of satisfying pleasure. We will program the machine accordingly so that you will have the sensations and experiences of your choice. You will certainly be very happy. In the machine, you will forget that you ever entered it; everything will seem real to you. Nor do you have to worry about your family and loved ones. We will also offer them to enter the machines, so they will be able to decide their life and happiness" What would you choose? be connected to the machine/stay in the real world

Appendix 2: Exposure to cases included in the study

See Table 16.

Table 16 Percentages of participants that declared that they had discussed the cases in class

Case	Gettier		Fake Barns		Truetemp		Knobe		Twin Earth	
Semester	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)
4	34.5	100	37.9	100	24.1	82.9	19	29.3	79.3	82.9
5	42.2	95.1	40.6	97.6	26.6	65.9	23.4	39	75	82.9
6	48.2	100	48.2	100	28.6	78.6	46.4	31	80.4	88.1
7	48	97.3	50	100	38	83.8	36	37.8	82	91.9
Case	Gödel/Schmidt		Frankfurt		Violinist		Experience Machine		Teleportation	
Semester	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)
4	79.3	29.3	17.2	9.8	10.3	19.5	19	34.1	48.3	34.1
5	75	63.4	20.3	53.7	20.3	48.8	25	46.3	56.2	53.7
6	82.1	73.8	30.4	50	39.3	61.9	42.9	54.8	48.2	59.5
7	76	81.1	36	56.8	36	78.4	42	64.9	56	67.6

“C” refers to the control group (cognitive science students) and “P” to the experimental group (philosophy students)

Appendix 3: Results of fitting a generalized linear mixed-effects model with logit as a link function

Gettier case

See Table 17 and Fig. 12.

Table 17 Logistic mixed-effects model for binary answers in the Gettier case

Predictors	Binary forced-choice answers (Gettier case)		
	Odds ratios	CI	p
(Intercept)	0.24	0.15 to 0.39	**< 0.001**
Semester 2-1	1.93	0.81 to 4.59	0.138
Semester 3-2	0.36	0.14 to 0.92	**0.032**
Semester 4-3	0.94	0.32 to 2.75	0.906
Semester 5-4	2.24	0.78 to 6.37	0.132
Semester 6-5	1.05	0.40 to 2.72	0.926
Semester 7-6	0.91	0.32 to 2.56	0.857
Group [Experimental]	0.20	0.09 to 0.44	**< 0.001**
Semester 2-1: GroupExperimental	0.07	0.02 to 0.27	**< 0.001**
Semester 3-2: GroupExperimental	4.88	1.00 to 23.89	0.050
Semester 4-3: GroupExperimental	1.18	0.20 to 6.79	0.856
Semester 5-4: GroupExperimental	0.30	0.05 to 1.92	0.202
Semester 6-5: GroupExperimental	0.19	0.01 to 2.46	0.202
Semester 7-6: GroupExperimental	7.81	0.60 to 102.26	0.117
Random Effects			
σ^2	3.29		
τ^{00} ident	2.81		
ICC	0.46		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.145/0.539		

Bold is used to indicate statistically significant predictors

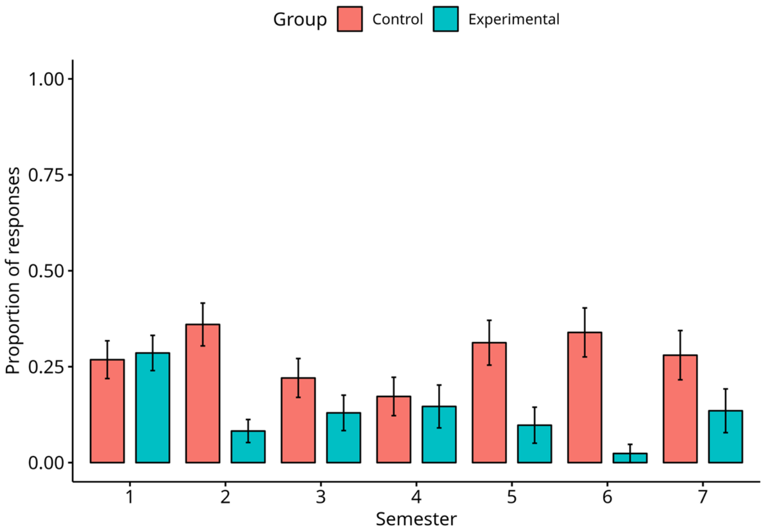

Fig. 12 Changes in intuitions about the Gettier case over time. The height of the bars represents the proportion of participants who attributed knowledge to the protagonist of the story. Error bars correspond to the standard error

Fake Barns

See Table 18 and Fig. 13.

Table 18 Logistic mixed-effects model for binary answers in the Fake Barns case

Predictors	Binary forced-choice answers (Fake Barns)		
	Odds ratios	CI	p
(Intercept)	3.76	2.42 to 5.86	**< 0.001**
Semester 2-1	1.09	0.45 to 2.63	0.843
Semester 3-2	0.76	0.31 to 1.88	0.554
Semester 4-3	0.85	0.34 to 2.17	0.740
Semester 5-4	0.72	0.28 to 1.84	0.498
Semester 6-5	2.18	0.81 to 5.86	0.123
Semester 7-6	0.75	0.26 to 2.20	0.602
Group [Experimental]	0.18	0.10 to 0.33	**< 0.001**
Semester 2-1: GroupExperimental	0.19	0.06 to 0.60	**0.005**
Semester 3-2: GroupExperimental	0.84	0.23 to 3.01	0.784
Semester 4-3: GroupExperimental	1.54	0.37 to 6.38	0.551
Semester 5-4: GroupExperimental	2.29	0.54 to 9.79	0.263
Semester 6-5: GroupExperimental	0.49	0.11 to 2.13	0.341
Semester 7-6: GroupExperimental	1.21	0.26 to 5.63	0.811
Random Effects			
σ^2	3.29		
τ^{00} ident	2.43		
ICC	0.42		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.139/0.505		

Bold is used to indicate statistically significant predictors

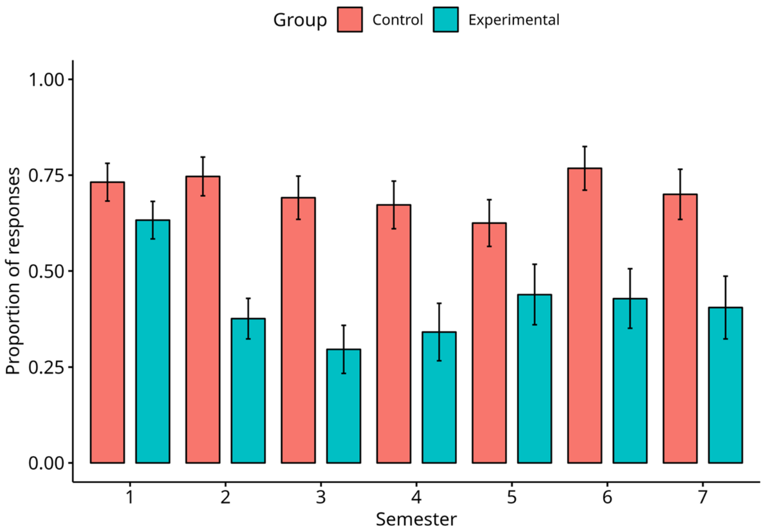

Fig. 13 Changes in intuitions about the Fake Barns case over time. The height of the bars represents the proportion of participants who attributed knowledge to the protagonist of the story. Error bars correspond to the standard error

Truetemp

See Table 19 and Fig. 14.

Table 19 Logistic mixed-effects model for binary answers in the Truetemp case

Predictors	Binary forced-choice answers (Truetemp)		
	Odds ratios	CI	p
(Intercept)	1.13	0.78 to 1.62	0.518
Semester 2-1	3.33	1.53 to 7.27	**0.003**
Semester 3-2	0.77	0.35 to 1.68	0.510
Semester 4-3	1.27	0.55 to 2.90	0.576
Semester 5-4	0.88	0.38 to 2.03	0.761
Semester 6-5	1.31	0.56 to 3.07	0.532
Semester 7-6	0.82	0.33 to 2.05	0.677
Group [Experimental]	0.69	0.40 to 1.17	0.167
Semester 2-1: GroupExperimental	0.24	0.08 to 0.68	**0.007**
Semester 3-2: GroupExperimental	0.89	0.28 to 2.80	0.841
Semester 4-3: GroupExperimental	1.25	0.35 to 4.48	0.737
Semester 5-4: GroupExperimental	2.22	0.58 to 8.47	0.244
Semester 6-5: GroupExperimental	0.33	0.08 to 1.26	0.103
Semester 7-6: GroupExperimental	1.44	0.36 to 5.79	0.609
Random Effects			
σ^2	3.29		
τ^{00} ident	1.96		
ICC	0.37		
N_{ident}	226		
Observations	849		
Marginal R^2/Conditional R^2	0.032/0.393		

Bold is used to indicate statistically significant predictors

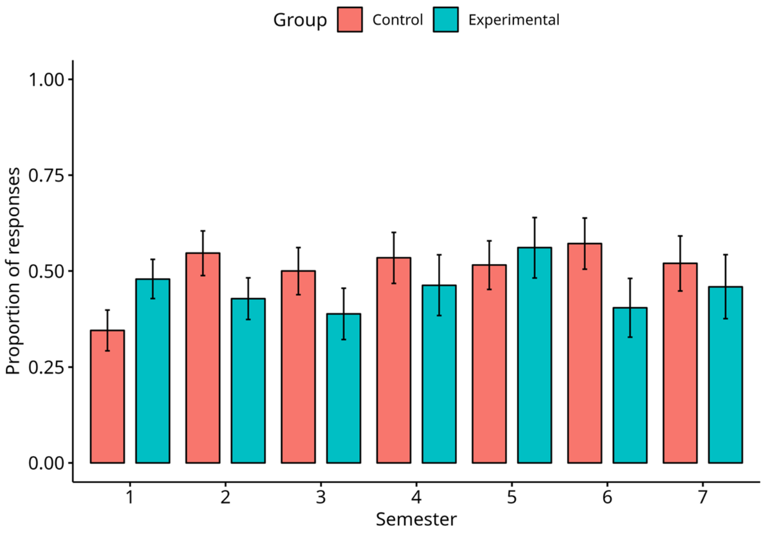

Fig. 14 Changes in intuitions about the Truetemp case over time. The height of the bars represents the proportion of participants who attributed knowledge to the protagonist of the story. Error bars correspond to the standard error

Knobe harm case

See Table 20 and Fig. 15.

Table 20 Logistic mixed-effects model for binary answers in the Knobe harm case

Predictors	Binary forced-choice answers (Knobe)		
	Odds ratios	CI	p
(Intercept)	0.19	0.11 to 0.33	**< 0.001**
Semester 2-1	1.06	0.42 to 2.67	0.905
Semester 3-2	0.82	0.31 to 2.15	0.683
Semester 4-3	1.64	0.59 to 4.54	0.344
Semester 5-4	0.98	0.35 to 2.69	0.964
Semester 6-5	0.62	0.21 to 1.82	0.385
Semester 7-6	0.95	0.30 to 3.02	0.927
Group [Experimental]	2.98	1.46 to 6.08	0.003
Semester 2-1: GroupExperimental	0.75	0.22 to 2.55	0.650
Semester 3-2: GroupExperimental	1.30	0.34 to 4.97	0.700
Semester 4-3: GroupExperimental	0.31	0.07 to 1.42	0.132
Semester 5-4: GroupExperimental	3.45	0.72 to 16.59	0.122

Table 20 (continued)

Predictors	Binary forced-choice answers (Knobe)		
	Odds ratios	CI	p
Semester 6-5: GroupExperimental	0.56	0.11 to 2.74	0.470
Semester 7-6: GroupExperimental	1.06	0.20 to 5.68	0.946
Random Effects			
σ^2	3.29		
τ^{00} ident	3.69		
ICC	0.53		
N_{ident}	226		
Observations	850		
Marginal R^2/Conditional R^2	0.056/0.555		

Bold is used to indicate statistically significant predictors

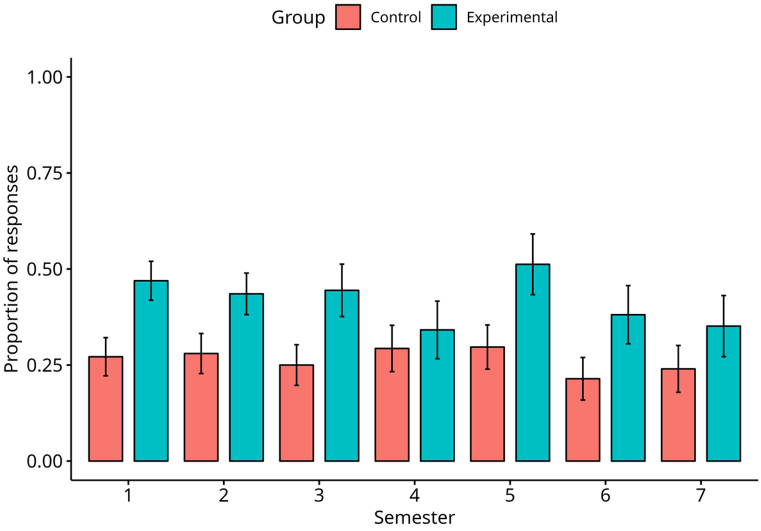

Fig. 15 Changes in intuitions about the Knobe harm case over time. The height of the bars represents the proportion of participants who attributed knowledge to the protagonist of the story. Error bars correspond to the standard error

Twin Earth

See Table 21 and Fig. 16.

Table 21 Logistic mixed-effects model for binary answers in the Twin Earth case

Predictors	Binary forced-choice answers (Twin Earth)		
	Odds ratios	CI	p
(Intercept)	0.02	0.01 to 0.07	**< 0.001**
Semester 2-1	6.48	0.98 to 42.83	0.052
Semester 3-2	1.38	0.36 to 5.23	0.639
Semester 4-3	0.90	0.23 to 3.56	0.880
Semester 5-4	0.58	0.13 to 2.63	0.479
Semester 6-5	2.81	0.65 to 12.10	0.166
Semester 7-6	1.43	0.37 to 5.49	0.599
Group [Experimental]	1.73	0.66 to 4.58	0.267
Semester 2-1: GroupExperimental	0.66	0.07 to 6.28	0.716
Semester 3-2: GroupExperimental	0.59	0.09 to 3.66	0.568
Semester 4-3: GroupExperimental	0.91	0.12 to 6.89	0.925
Semester 5-4: GroupExperimental	1.86	0.21 to 16.83	0.580
Semester 6-5: GroupExperimental	0.68	0.08 to 5.55	0.716
Semester 7-6: GroupExperimental	0.57	0.08 to 4.18	0.581
Random Effects			
σ^2	3.29		
τ^{00} ident	5.59		
ICC	0.63		
N_{ident}	226		
Observations	851		
Marginal R^2/Conditional R^2	0.073/0.657		

Bold is used to indicate statistically significant predictors

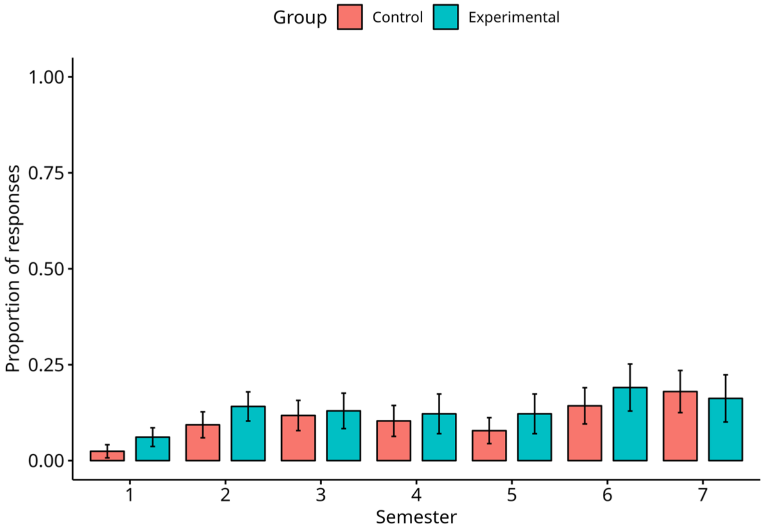

Fig. 16 Changes in intuitions about the Twin Earth case over time. The height of the bars represents the proportion of participants who agreed that XYZ is water. Error bars correspond to the standard error

Gödel/Schmidt

See Table 22 and Fig. 17.

Table 22 Logistic mixed-effects model for binary answers in the Twin Earth case

Predictors	Binary forced-choice answers (Gödel/Schmidt)		
	Odds ratios	CI	p
(Intercept)	0.57	0.39 to 0.86	**0.007**
Semester 2-1	0.66	0.29 to 1.51	0.329
Semester 3-2	1.77	0.76 to 4.13	0.183
Semester 4-3	0.82	0.34 to 1.99	0.667
Semester 5-4	1.68	0.69 to 4.12	0.253
Semester 6-5	1.72	0.71 to 4.18	0.228
Semester 7-6	0.32	0.12 to 0.84	**0.020**
Group [Experimental]	2.14	1.19 to 3.85	**0.011**
Semester 2-1: GroupExperimental	2.77	0.89 to 8.58	0.077
Semester 3-2: GroupExperimental	1.99	0.58 to 6.89	0.275
Semester 4-3: GroupExperimental	1.07	0.27 to 4.24	0.926
Semester 5-4: GroupExperimental	0.44	0.11 to 1.85	0.265
Semester 6-5: GroupExperimental	0.75	0.18 to 3.07	0.684
Semester 7-6: GroupExperimental	3.24	0.72 to 14.49	0.124
Random Effects			
σ^2	3.29		
τ^{00} ident	2.41		
ICC	0.42		
N_{ident}	225		
Observations	850		
Marginal R^2/Conditional R^2	0.070/0.463		

Bold is used to indicate statistically significant predictors

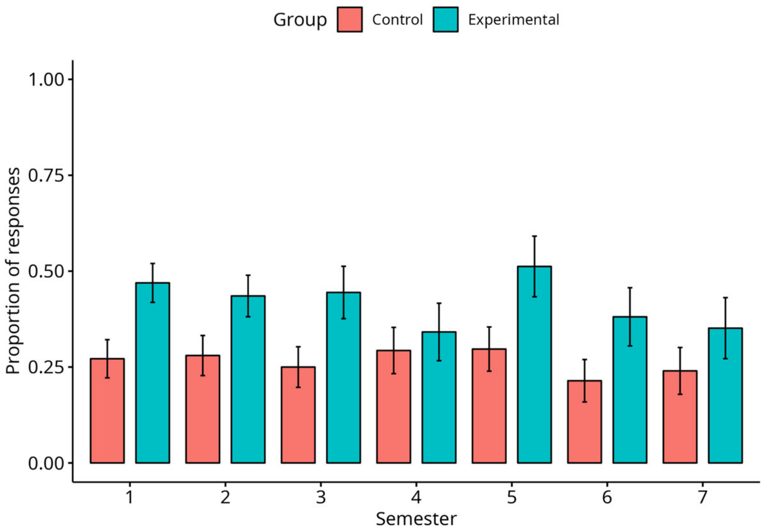

Fig. 17 Changes in intuitions about the Gödel/Schmidt case over time. The height of the bars represents the proportion of participants who agreed that the name "Gödel" refers to the fraud. Error bars correspond to the standard error

Experience Machine

See Table 23 and Fig. 18.

Table 23 Logistic mixed-effects model for binary answers in the Experience Machine case

Predictors	Binary answers (Experience Machine)		
	Odds ratios	CI	p
(Intercept)	2185.63	324.48 to 14,721.97	**< 0.001**
Semester 2-1	2.26	0.47 to 10.75	0.307
Semester 3-2	5.13	0.73 to 36.22	0.101
Semester 4-3	0.64	0.07 to 6.29	0.702
Semester 5-4	0.64	0.08 to 5.15	0.673
Semester 6-5	0.15	0.02 to 0.87	**0.035**
Semester 7-6	2.22	0.41 to 12.10	0.358
Group [Experimental]	0.83	0.15 to 4.53	0.834
Semester 2-1: GroupExperimental	0.05	0.00 to 0.55	0.014
Semester 3-2: GroupExperimental	0.35	0.03 to 4.46	0.417
Semester 4-3: GroupExperimental	2.13	0.11 to 41.39	0.617
Semester 5-4: GroupExperimental	1.03	0.06 to 18.66	0.984
Semester 6-5: GroupExperimental	6.96	0.47 to 102.06	0.157
Semester 7-6: GroupExperimental	1.01	0.07 to 14.80	0.994
Random Effects			
σ^2	3.29		
τ^{00} ident	58.62		
ICC	0.95		
N_{ident}	225		
Observations	847		
Marginal R^2/Conditional R^2	0.013/0.948		

Bold is used to indicate statistically significant predictors

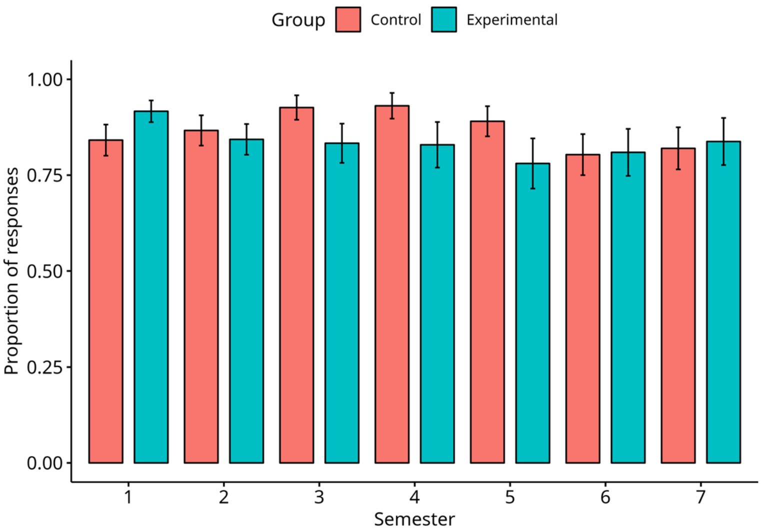

Fig. 18 Changes in intuitions about the Experience Machine case over time. The height of the bars represents the proportion of participants who would decide to stay in the real world. Error bars correspond to the standard error

Violinist

See Table 24 and Fig. 19.

Table 24 Logistic mixed-effects model for binary answers in the Violinist case

Predictors	Binary forced-choice answers (Violinist)		
	Odds ratios	CI	p
(Intercept)	0.30	0.16 to 0.56	**< 0.001**
Semester 2-1	1.41	0.56 to 3.55	0.471
Semester 3-2	0.74	0.28 to 1.91	0.531
Semester 4-3	0.36	0.13 to 1.01	0.053
Semester 5-4	1.44	0.49 to 4.25	0.508
Semester 6-5	0.68	0.23 to 2.00	0.480
Semester 7-6	1.73	0.56 to 5.32	0.342
Group [Experimental]	0.85	0.36 to 1.99	0.710
Semester 2-1:GroupExperimental	0.50	0.14 to 1.80	0.289
Semester 3-2:GroupExperimental	0.76	0.18 to 3.21	0.712
Semester 4-3:GroupExperimental	12.90	2.51 to 66.20	**0.002**
Semester 5-4:GroupExperimental	0.12	0.02 to 0.69	**0.018**
Semester 6-5:GroupExperimental	4.87	0.86 to 27.51	0.073
Semester 7-6:GroupExperimental	0.23	0.04 to 1.31	0.098
Random Effects			
σ^2	3.29		
τ^{00} ident	5.93		
ICC	0.64		
N_{ident}	226		
Observations	849		
Marginal R^2/Conditional R^2	0.028/0.653		

Bold is used to indicate statistically significant predictors

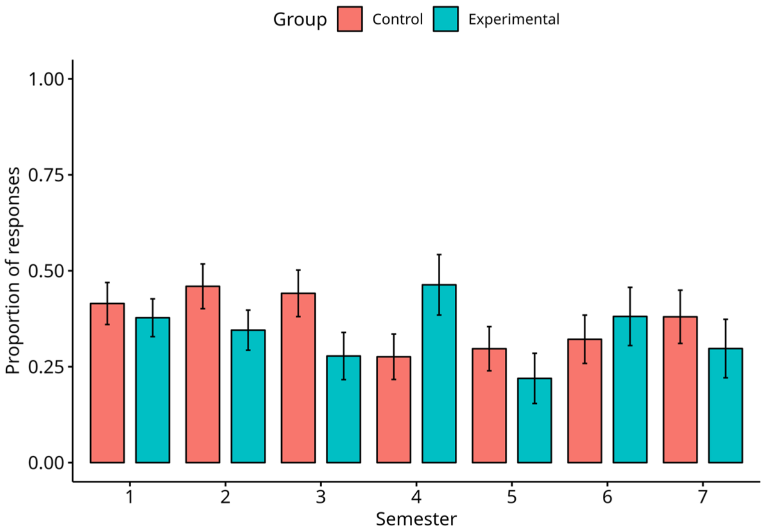

Fig. 19 Changes in intuitions about the Violinist case over time. The height of the bars represents the proportion of participants who agreed that they have a moral duty to stay connected to the violinist. Error bars correspond to the standard error

Frankfurt case

See Table 25 and Fig. 20.

Table 25 Logistic mixed-effects model for binary answers to three questions in Frankfurt case

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Odds ratios	CI	p	Odds ratios	CI	p	Odds ratios	CI	p
(Intercept)	0.32	0.21 to 0.49	**< 0.001**	46.24	14.03 to 152.36	**< 0.001**	25.97	11.55 to 58.37	**< 0.001**
Semester 2-1	1.27	0.56 to 2.89	0.571	0.53	0.15 to 1.92	**0.332**	0.37	0.11 to 1.26	**0.112**
Semester 3-2	0.71	0.30 to 1.69	0.442	2.45	0.60 to 9.99	0.211	1.48	0.46 to 4.70	0.509
Semester 4-3	0.89	0.35 to 2.28	0.804	0.49	0.11 to 2.22	0.352	0.95	0.26 to 3.46	0.939
Semester 5-4	1.59	0.62 to 4.09	0.334	0.32	0.08 to 1.22	0.094	0.76	0.21 to 2.74	0.671
Semester 6-5	0.45	0.17 to 1.19	0.108	3.57	0.88 to 14.41	0.074	0.73	0.21 to 2.49	0.611
Semester 7-6	1.09	0.38 to 3.12	0.869	0.61	0.13 to 2.87	0.530	6.04	1.01 to 36.17	**0.049**
Group [Experimental]	1.30	0.71 to 2.35	0.394	0.72	0.27 to 1.91	0.505	0.95	0.40 to 2.25	0.905
Semester 2-1: GroupExperimental	0.46	0.15 to 1.39	0.166	2.44	0.45 to 13.13	0.299	4.34	0.89 to 21.04	**0.069**
Semester 3-2: GroupExperimental	1.90	0.56 to 6.52	0.306	0.45	0.07 to 3.14	0.422	3.39	0.41 to 28.23	0.260
Semester 4-3: GroupExperimental	0.88	0.21 to 3.57	0.854	0.94	0.11 to 8.14	0.955	0.12	0.01 to 1.17	0.068
Semester 5-4: GroupExperimental	0.73	0.17 to 3.18	0.679	5.72	0.66 to 49.62	0.114	2.55	0.34 to 19.05	0.361

Table 25 (continued)

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Odds ratios	CI	p	Odds ratios	CI	p	Odds ratios	CI	p
Semester 6-5: GroupExperimental	1.89	0.44 to 8.19	0.395	0.16	0.02 to 1.38	0.095	1.04	0.14 to 7.87	0.967
Semester 7-6: GroupExperimental	0.28	0.05 to 1.41	0.123	16.93	1.13 to 254.42	**0.041**	1.30	0.06 to 26.45	0.866
Random Effects									
σ^2	3.29			3.29			3.29		
$\tau_{00\ \mathrm{ident}}$	2.34			6.10			3.29		
ICC	0.42			0.65			0.50		
N_{ident}	226			226			226		
Observations	851			850			851		
Marginal R^2/Conditional R^2	0.028/0.432			0.040/0.664			0.072/0.536		

Bold is used to indicate statistically significant predictors

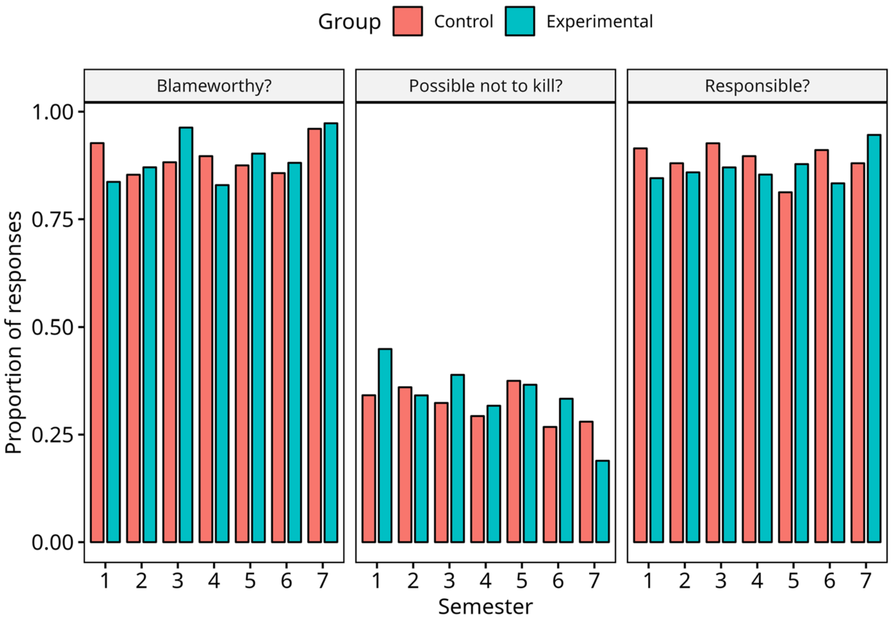

Fig. 20 Changes in intuitions about Frankfurt case over time. The height of the bars represents the proportion of participants who believe that neither copy was the original. Error bars correspond to the standard error

Teleportation

See Table 26 and Fig. 21.

Table 26 Logistic mixed-effects model for binary answers in the Teleportation case

Predictors	Binary forced-choice answers (Teleportation)		
	Odds ratios	CI	p
(Intercept)	0.97	0.60 to 1.56	0.887
Semester 2-1	1.21	0.53 to 2.77	0.651
Semester 3-2	0.60	0.25 to 1.41	0.240
Semester 4-3	0.74	0.30 to 1.83	0.518
Semester 5-4	1.30	0.52 to 3.25	0.579
Semester 6-5	0.65	0.26 to 1.63	0.357
Semester 7-6	1.12	0.42 to 3.00	0.825
Group [Experimental]	0.86	0.43 to 1.72	0.675
Semester 2-1: GroupExperimental	0.44	0.14 to 1.37	0.156
Semester 3-2: GroupExperimental	1.74	0.51 to 6.01	0.378

Table 26 (continued)

Predictors	Binary forced-choice answers (Teleportation)		
	Odds ratios	CI	p
Semester 4-3: GroupExperimental	2.29	0.57 to 9.13	0.242
Semester 5-4: GroupExperimental	0.48	0.11 to 2.07	0.326
Semester 6-5: GroupExperimental	2.35	0.55 to 10.08	0.249
Semester 7-6: GroupExperimental	0.52	0.12 to 2.35	0.398
Random Effects			
σ2	3.29		
τ^{00} ident	4.01		
ICC	0.55		
N_{ident}	226		
Observations	850		
Marginal R^2/Conditional R^2	0.014/0.556		

Bold is used to indicate statistically significant predictors

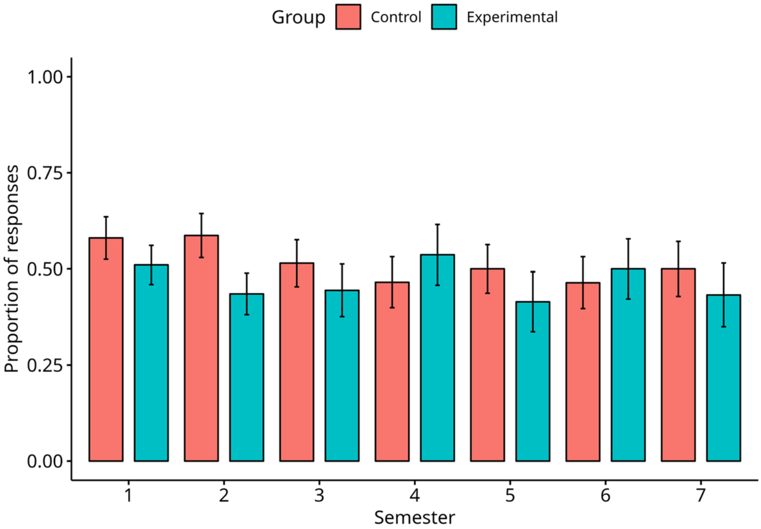

Fig. 21 Changes in intuitions about the Teleportation case over time. The height of the bars represents the proportion of participants who believe that neither copy was the original. Error bars correspond to the standard error

Appendix 4: Detailed breakdown of the answers for the Teleportation case

See Table 27 and Fig. 22.

Table 27 Breakdown of the answers in the Teleportation case

Semester	Control								Experimental							
	Both		Mother		Neither		Wife		Both		Mother		Neither		Wife	
	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n
1	52.44	43	1.22	1	41.46	34	3.66	3	50.00	49	0.00	0	48.98	48	1.02	1
2	54.67	41	0.00	0	41.33	31	4.00	3	43.53	37	0.00	0	56.47	48	0.00	0
3	50.00	34	0.00	0	48.53	33	1.47	1	42.59	23	1.85	1	55.56	30	0.00	0
4	44.83	26	0.00	0	53.45	31	1.72	1	51.22	21	0.00	0	46.34	19	2.44	1
5	46.88	30	0.00	0	50.00	32	3.12	2	39.02	16	0.00	0	58.54	24	2.44	1
6	44.64	25	0.00	0	53.57	30	1.79	1	50.00	21	0.00	0	50.00	21	0.00	0
7	50.00	25	0.00	0	50.00	25	0.00	0	43.24	16	0.00	0	56.76	21	0.00	0

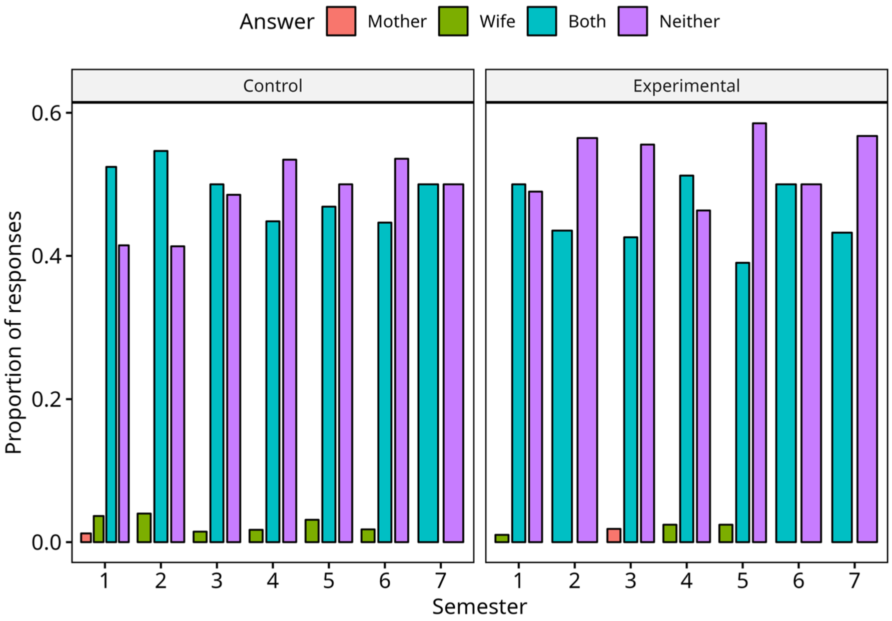

Fig. 22 Breakdown of the answers in the Teleportation case. Missing bars represent a count of 0

Appendix 5: Results of fitting a linear mixed-effects model to the reduced data set

In order to check whether the selection bias affected our results in a significant way, we decided to re-run the main analysis on a subset of the original data. In these analyses we included only the participants for which we have (a) all seven measurement points; (b) six measurement points if only sixth or seventh measurement point is missing.

Gettier case

See Table 28 and Fig. 23.

Table 28 Linear mixed-effects model for the combined scores in the Gettier case

Predictors	Combined score (Gettier case)		
	Estimates	CI	p
(Intercept)	− 2.26	− 2.84 to − 1.67	**< 0.001**
Semester 2-1	0.28	− 0.86 to 1.42	0.628
Semester 3-2	− 0.63	− 1.78 to 0.51	0.279
Semester 4-3	− 0.43	− 1.57 to 0.71	0.457

Table 28 (continued)

Predictors	Combined score (Gettier case)		
	Estimates	CI	p
Semester 5-4	1.25	0.12 to 2.39	**0.031**
Semester 6-5	0.31	– 0.86 to 1.48	0.605
Semester 7-6	– 0.98	– 2.19 to 0.23	0.113
Group [Experimental]	– 0.89	– 1.79 to 0.00	0.051
Semester 2-1: GroupExperimental	– 2.28	– 4.02 to – 0.54	**0.010**
Semester 3-2: GroupExperimental	1.08	– 0.68 to 2.83	0.229
Semester 4-3: GroupExperimental	0.17	– 1.60 to 1.94	0.846
Semester 5-4: GroupExperimental	– 1.00	– 2.77 to 0.77	0.269
Semester 6-5: GroupExperimental	– 0.97	– 2.75 to 0.80	0.281
Semester7-6: GroupExperimental	1.84	0.03 to 3.65	**0.046**
Random Effects			
σ^2	6.24		
$\tau_{00\ ident}$	2.50		
ICC	0.29		
N_{ident}	68		
Observations	444		
Marginal R^2/Conditional R^2	0.057/0.327		

Bold is used to indicate statistically significant predictors

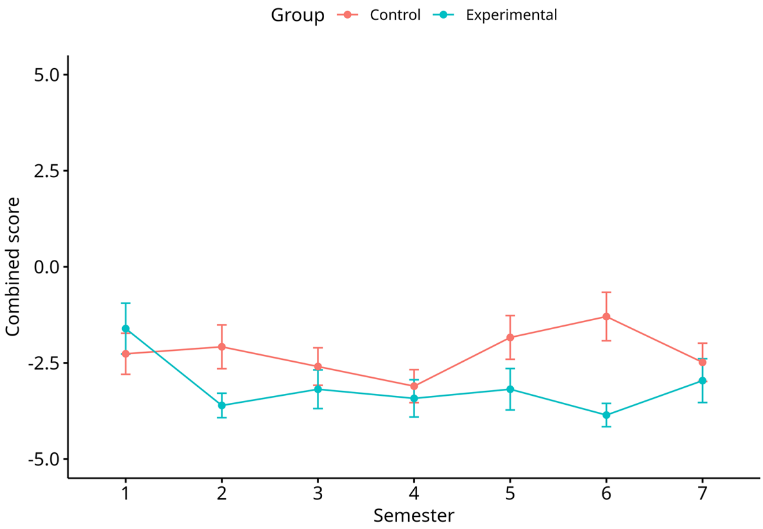

Fig. 23 Changes in intuitions about the Gettier case over time. The combined score of + 5 represents attribution of knowledge with maximum confidence and – 5 represents a denial of knowledge with maximum confidence. Error bars correspond to the standard error of the mean

Fake Barns

See Table 29 and Fig. 24.

Table 29 Linear mixed-effects model for the combined scores in the Fake Barns case

Predictors	Combined score (Fake Barns)		
	Estimates	CI	p
(Intercept)	1.61	0.80 to 2.42	**< 0.001**
Semester 2-1	0.19	− 1.02 to 1.40	0.760
Semester 3-2	− 0.60	− 1.82 to 0.62	0.334
Semester 4-3	0.11	− 1.11 to 1.32	0.862
Semester 5-4	− 0.53	− 1.74 to 0.68	0.391
Semester 6-5	1.13	− 0.13 to 2.38	0.078
Semester 7-6	− 0.03	− 1.32 to 1.27	0.966
Group [Experimental]	− 2.79	− 4.03 to − 1.55	**< 0.001**
Semester t2-1: GroupExperimental	− 2.36	− 4.21 to − 0.50	**0.013**
Semester 3-2: GroupExperimental	0.52	− 1.35 to 2.40	0.583
Semester 4-3: GroupExperimental	− 0.19	− 2.08 to 1.70	0.845
Semester 5-4: GroupExperimental	1.20	− 0.69 to 3.09	0.214
Semester 6-5: GroupExperimental	− 1.11	− 3.01 to 0.78	0.248
Semester 7-6: GroupExperimental	0.10	− 1.83 to 2.04	0.917
Random Effects			
σ^2	7.11		
$\tau_{00\ \mathrm{ident}}$	5.52		
ICC	0.44		
N_{ident}	68		
Observations	444		
Marginal R^2/Conditional R^2	0.148/0.521		

Bold is used to indicate statistically significant predictors

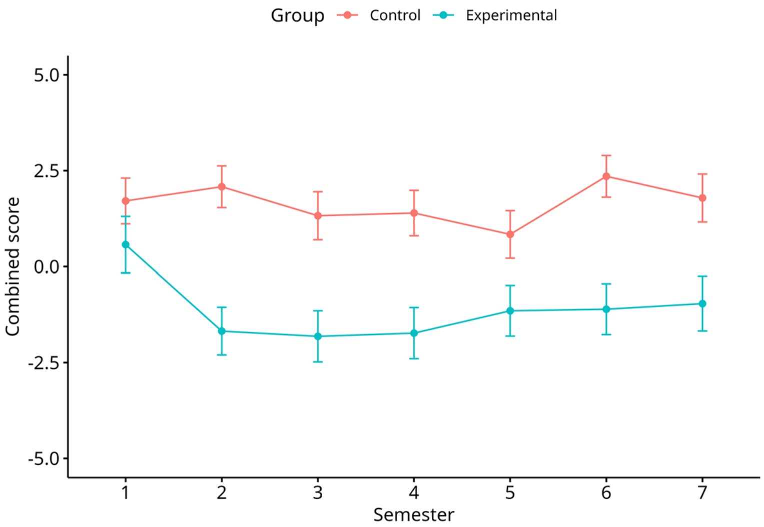

Fig. 24 Changes in intuitions about the Fake Barns case over time. The combined score of + 5 represents attribution of knowledge with maximal confidence and − 5 represents a denial of knowledge with maximal confidence. Error bars correspond to the standard error of the mean

Truetemp

See Table 30 and Fig. 25.

Table 30 Linear mixed-effects model for the combined scores in the Truetemp case

Predictors	Combined score (Truetemp)		
	Estimates	CI	p
(Intercept)	− 0.38	− 1.13 to 0.37	0.318
Semester 2-1	1.94	0.53 to 3.36	**0.007**
Semester 3-2	0.24	− 1.18 to 1.65	0.742
Semester 4-3	0.18	− 1.22 to 1.59	0.798
Semester 5-4	0.14	− 1.27 to 1.54	0.850
Semester 6-5	0.34	− 1.11 to 1.79	0.646
Semester 7-6	− 0.49	− 1.99 to 1.01	0.521
Group [Experimental]	− 0.30	− 1.45 to 0.84	0.601
Semester 2-1: GroupExperimental	− 3.18	− 5.35 to − 1.01	**0.004**
Semester 3-2: GroupExperimental	− 0.15	− 2.32 to 2.02	0.894

Table 30 (continued)

Predictors	Combined score (Truetemp)		
	Estimates	CI	p
Semester 4-3: GroupExperimental	0.18	− 2.01 to 2.37	0.871
Semester 5-4: GroupExperimental	0.93	− 1.26 to 3.12	0.405
Semester 6-5: GroupExperimental	− 1.12	− 3.31 to 1.08	0.317
Semester 7-6: GroupExperimental	0.56	− 1.67 to 2.80	0.620
Random Effects			
σ^2	9.54		
$\tau_{00\ \mathrm{ident}}$	4.17		
ICC	0.30		
N_{ident}	68		
Observations	442		
Marginal R^2/Conditional R^2	0.040/0.332		

Bold is used to indicate statistically significant predictors

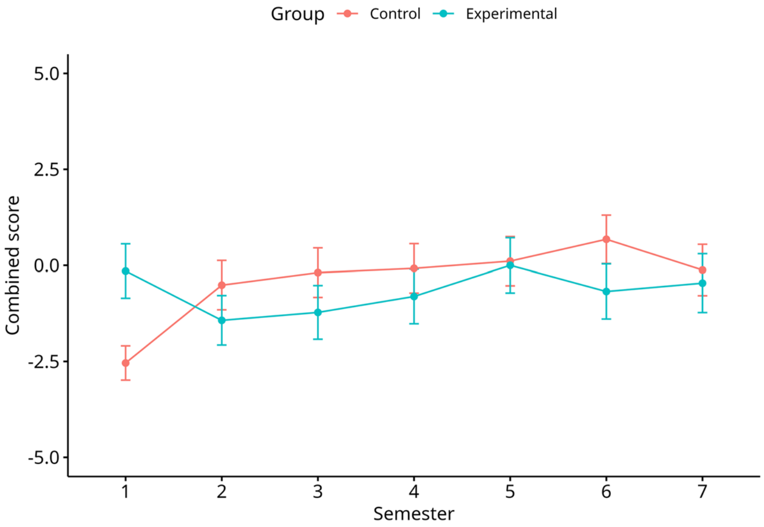

Fig. 25 Changes in intuitions about the Truetemp case over time. The combined score of + 5 represents attribution of knowledge with maximum confidence and − 5 represents a denial of knowledge with maximum confidence. Error bars correspond to the standard error of the mean

Knobe harm case

See Table 31 and Fig. 26.

Table 31 Linear mixed-effects model for the combined scores in the Knobe harm case

Predictors	Combined score (Knobe)		
	Estimates	CI	p
(Intercept)	− 2.38	− 3.21 to − 1.54	**< 0.001**
Semester 2-1	− 0.26	− 1.37 to 0.85	0.646
Semester 3-2	0.53	− 0.58 to 1.64	0.350
Semester 4-3	0.47	− 0.63 to 1.57	0.405
Semester 5-4	− 0.45	− 1.55 to 0.65	0.421
Semester 6-5	− 0.64	− 1.78 to 0.50	0.272
Semester 7-6	0.13	− 1.04 to 1.31	0.825
Group [Experimental]	1.59	0.31 to 2.87	**0.015**
Semester 2-1: GroupExperimental	0.51	− 1.18 to 2.20	0.553
Semester 3-2: GroupExperimental	− 0.37	− 2.07 to 1.33	0.672
Semester 4-3: GroupExperimental	− 1.50	− 3.22 to 0.21	0.086
Semester 5-4: GroupExperimental	1.57	− 0.15 to 3.28	0.073
Semester 6-5: GroupExperimental	− 0.09	− 1.81 to 1.63	0.920
Semester 7-6: GroupExperimental	− 0.19	− 1.94 to 1.57	0.833
Random Effects			
σ^2	5.86		
$\tau_{00\ \mathrm{ident}}$	6.14		
ICC	0.51		
N_{ident}	68		
Observations	443		
Marginal R^2/Conditional R^2	0.059/0.540		

Bold is used to indicate statistically significant predictors

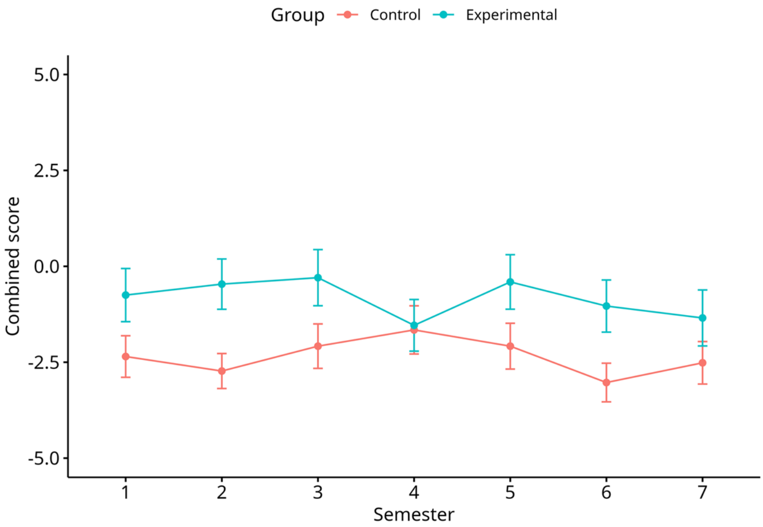

Fig. 26 Changes in intuitions about the Knobe harm case over time. The combined score of + 5 represents attribution of intentionality with maximum confidence and − 5 represents a denial of intentionality with maximum confidence. Error bars correspond to the standard error of the mean

Twin Earth

See Table 32 and Fig. 27.

Table 32 Linear mixed-effects model for the combined scores in the Twin Earth case

Predictors	Combined score (Twin Earth)		
	Estimates	CI	p
(Intercept)	− 3.47	− 4.03 to − 2.91	**< 0.001**
Semester 2-1	0.90	− 0.14 to 1.93	0.090
Semester 3-2	− 0.09	− 1.13 to 0.95	0.864
Semester 4-3	0.47	− 0.56 to 1.51	0.370
Semester 5-4	− 0.23	− 1.26 to 0.81	0.664
Semester 6-5	0.58	− 0.49 to 1.65	0.284
Semester 7-6	− 0.05	− 1.15 to 1.06	0.934
Group [Experimental]	0.49	− 0.37 to 1.34	0.262
Semester 2-1: GroupExperimental	0.30	− 1.29 to 1.88	0.714
Semester 3-2: GroupExperimental	− 0.94	− 2.54 to 0.65	0.246

Table 32 (continued)

Predictors	Combined score (Twin Earth)		
	Estimates	CI	p
Semester 4-3: GroupExperimental	− 0.14	− 1.75 to 1.47	0.863
Semester 5-4: GroupExperimental	0.32	− 1.29 to 1.94	0.693
Semester 6-5: GroupExperimental	− 0.44	− 2.05 to 1.18	0.593
Semester 7-6: GroupExperimental	0.54	− 1.11 to 2.18	0.523
Random Effects			
σ^2	5.18		
$\tau_{00\ ident}$	2.35		
ICC	0.31		
N_{ident}	68		
Observations	444		
Marginal R^2/Conditional R^2	0.038/0.338		

Bold is used to indicate statistically significant predictors

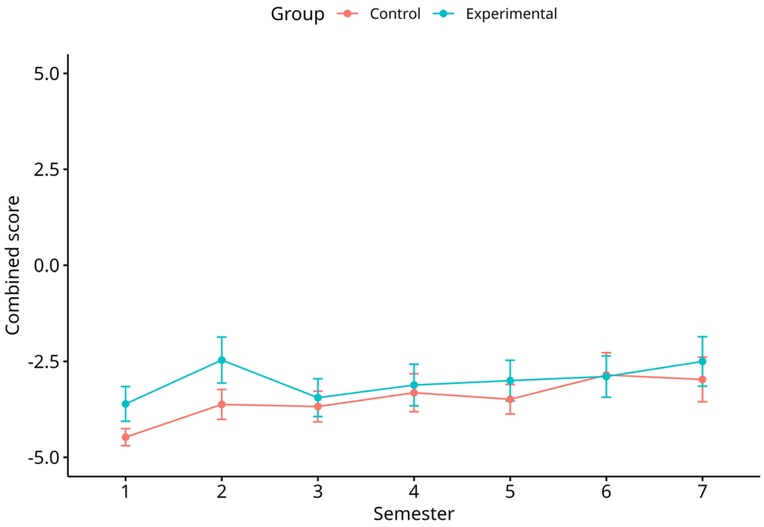

Fig. 27 Changes in intuitions about the Twin Earth case over time. The combined score of + 5 represents a belief in the statement that XYZ is water with maximum confidence and − 5 represents a denial of that statement with maximum confidence. Error bars correspond to the standard error of the mean

Gödel/Schmidt

See Table 33 and Fig. 28.

Table 33 Linear mixed-effects model for the combined scores in the Gödel/Schmidt case

Predictors	Combined score (Gödel/Schmidt)		
	Estimates	CI	p
(Intercept)	− 1.10	− 1.92 to − 0.27	**0.009**
Semester 2-1	− 0.16	− 1.50 to 1.17	0.810
Semester 3-2	− 0.45	− 1.79 to 0.90	0.515
Semester 4-3	0.41	− 0.92 to 1.75	0.546
Semester 5-4	0.35	− 0.99 to 1.68	0.608
Semester 6-5	0.63	− 0.74 to 2.01	0.366
Semester 7-6	− 1.70	− 3.12 to − 0.27	**0.020**
Group [Experimental]	1.51	0.25 to 2.77	**0.019**
Semester 2-1: GroupExperimental	0.49	− 1.55 to 2.53	0.638
Semester 3-2: GroupExperimental	2.29	0.23 to 4.35	**0.030**
Semester 4-3: GroupExperimental	− 0.91	− 2.99 to 1.16	0.388
Semester 5-4: GroupExperimental	− 0.42	− 2.50 to 1.66	0.692
Semester 6-5: GroupExperimental	− 0.84	− 2.93 to 1.24	0.426
Semester 7-6: GroupExperimental	1.92	− 0.21 to 4.04	0.077
Random Effects			
σ^2	8.60		
$\tau_{00\ \mathrm{ident}}$	5.52		
ICC	0.39		
N_{ident}	68		
Observations	444		
Marginal R^2/Conditional R^2	0.061/0.428		

Bold is used to indicate statistically significant predictors

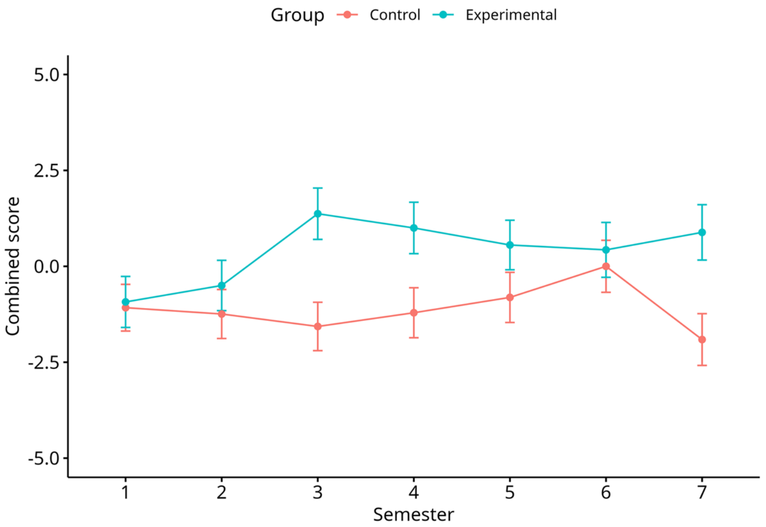

Fig. 28 Changes in intuitions about the Gödel/Schmidt case over time. The combined score of + 5 represents a belief in the statement that the name "Gödel" refers to the fraud with maximum confidence and − 5 represents a belief in the statement that it refers to the author of the proof with maximum confidence. Error bars correspond to the standard error of the mean

Experience Machine

See Table 34 and Fig. 29.

Table 34 Linear mixed-effects model for the combined scores in the Experience Machine case

Predictors	Combined score (Experience Machine)		
	Estimates	CI	p
(Intercept)	3.34	2.54 to 4.15	**< 0.001**
Semester 2-1	− 0.35	− 1.12 to 0.42	0.374
Semester 3-2	0.48	− 0.29 to 1.26	0.220
Semester 4-3	− 0.06	− 0.83 to 0.70	0.869
Semester 5-4	0.04	− 0.73 to 0.81	0.915
Semester 6-5	− 0.78	− 1.57 to 0.02	0.055
Semester 7-6	0.18	− 0.64 to 1.00	0.664
Group [Experimental]	− 1.35	− 2.58 to − 0.11	**0.033**
Semester 2-1: GroupExperimental	− 0.55	− 1.74 to 0.63	0.359
Semester 3-2: GroupExperimental	− 0.10	− 1.29 to 1.10	0.871

Table 34 (continued)

Predictors	Combined score (Experience Machine)		
	Estimates	CI	p
Semester 4-3: GroupExperimental	0.43	− 0.77 to 1.63	0.478
Semester 5-4: GroupExperimental	− 0.19	− 1.39 to 1.01	0.756
Semester 6-5: GroupExperimental	0.84	− 0.36 to 2.03	0.172
Semester 7-6: GroupExperimental	0.10	− 1.13 to 1.32	0.873
Random Effects			
σ^2	2.85		
$\tau_{00\ \text{ident}}$	6.14		
ICC	0.68		
N_{ident}	68		
Observations	442		
Marginal R^2/Conditional R^2	0.057/0.701		

Bold is used to indicate statistically significant predictors

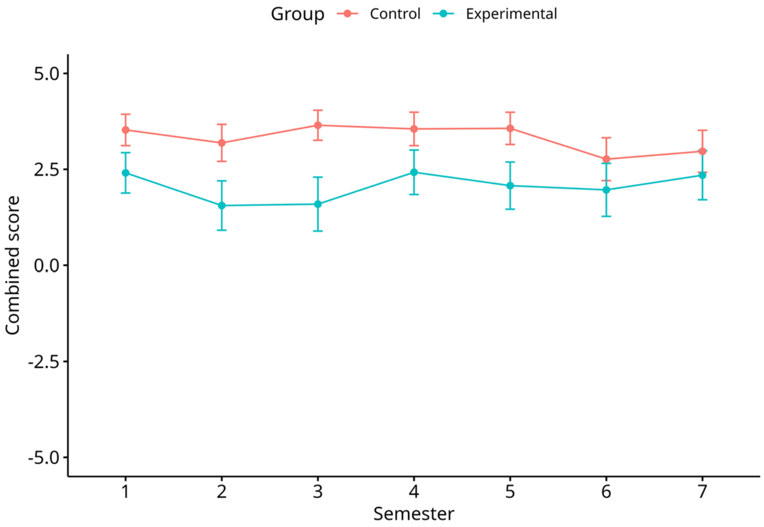

Fig. 29 Changes in intuitions about the Experience Machine case over time. The combined score of + 5 represents a preference for staging in the real world with maximum confidence and − 5 represents a preference for connecting to the machine with maximum confidence. Error bars correspond to the standard error of the mean

Violinist

See Table 35 and Fig. 30.

Table 35 Linear mixed-effects model for the combined scores in the Violinist case

Predictors	Combined score (Violinist)		
	Estimates	CI	p
(Intercept)	− 1.41	− 2.31 to − 0.52	**0.002**
Semester 2-1	− 0.08	− 1.16 to 1.00	0.881
Semester 3-2	− 0.33	− 1.42 to 0.76	0.551
Semester 4-3	− 0.76	− 1.84 to 0.32	0.165
Semester 5-4	0.14	− 0.94 to 1.22	0.799
Semester 6-5	− 0.15	− 1.26 to 0.96	0.792
Semester 7-6	0.35	− 0.80 to 1.50	0.548
Group [Experimental]	0.09	− 1.28 to 1.47	0.892
Semester 2-1: GroupExperimental	− 0.50	− 2.14 to 1.15	0.553
Semester 3-2: GroupExperimental	− 0.72	− 2.38 to 0.95	0.397
Semester 4-3: GroupExperimental	2.48	0.81 to 4.16	**0.004**
Semester 5-4: GroupExperimental	− 1.01	− 2.69 to 0.67	0.236
Semester 6-5: GroupExperimental	0.15	− 1.53 to 1.83	0.859
Semester 7-6: GroupExperimental	− 0.91	− 2.63 to 0.80	0.296
Random Effects			
σ^2	5.60		
$\tau_{00\ \mathrm{ident}}$	7.25		
ICC	0.56		
N_{ident}	68		
Observations	444		
Marginal R^2/Conditional R^2	0.021/0.573		

Bold is used to indicate statistically significant predictors

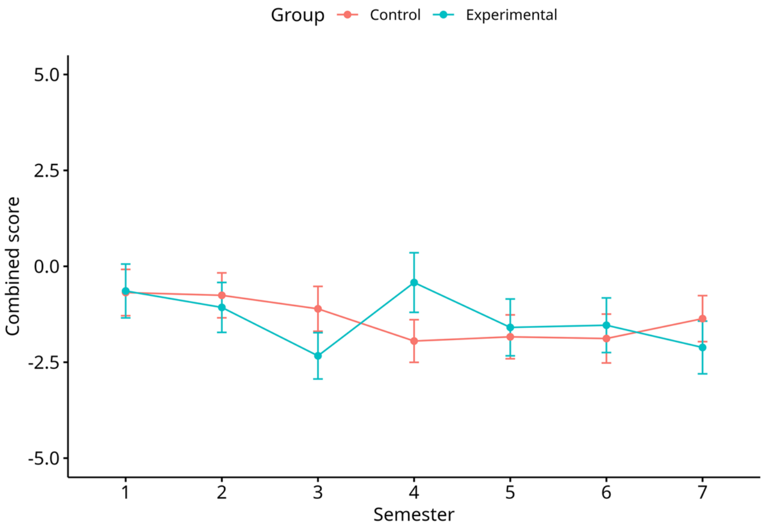

Fig. 30 Changes in intuitions about the Violinist case over time. The combined score of + 5 represents a belief that one has a moral duty to stay connected to the violinist with maximum confidence and − 5 represents a denial of the existence of this duty with maximum confidence. Error bars correspond to the standard error of the mean

Frankfurt case

See Table 36 and Fig. 31.

Table 36 Linear mixed-effects models for the combined scores in three questions about the Frankfurt case

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	− 1.72	− 2.64 to − 0.80	**< 0.001**	3.34	2.79 to 3.90	**< 0.001**	3.41	2.99 to 3.83	**< 0.001**
Semester 2-1	1.55	0.30 to 2.80	**0.015**	− 0.81	− 1.78 to 0.16	0.103	− 0.71	− 1.64 to 0.22	0.135
Semester 3-2	− 1.12	− 2.37 to 0.13	0.078	0.59	− 0.39 to 1.57	0.236	0.20	− 0.74 to 1.13	0.683
Semester 4-3	− 0.02	− 1.26 to 1.22	0.974	− 0.76	− 1.73 to 0.22	0.127	− 0.40	− 1.33 to 0.53	0.397
Semester 5-4	0.45	− 0.79 to 1.69	0.477	− 0.76	− 1.73 to 0.22	0.127	− 0.37	− 1.31 to 0.56	0.429
Semester 6-5	− 1.03	− 2.31 to 0.25	0.114	0.57	− 0.44 to 1.57	0.266	0.41	− 0.55 to 1.37	0.397
Semester 7-6	− 0.04	− 1.36 to 1.28	0.956	0.18	− 0.86 to 1.22	0.731	0.59	− 0.40 to 1.58	0.241
Group [Experimental]	0.37	− 1.03 to 1.78	0.601	0.03	− 0.82 to 0.88	0.938	0.17	− 0.47 to 0.81	0.601
Semester 2-1: Group-Experimental	− 1.70	− 3.60 to 0.20	0.079	0.83	− 0.66 to 2.31	0.275	0.94	− 0.50 to 2.37	0.199
Semester 3-2: Group-Experimental	1.23	− 0.68 to 3.14	0.206	− 0.77	− 2.27 to 0.73	0.314	− 0.14	− 1.58 to 1.29	0.847
Semester 4-3: Group-Experimental	− 0.72	− 2.65 to 1.21	0.465	0.58	− 0.93 to 2.10	0.449	0.18	− 1.27 to 1.63	0.809
Semester 5-4: Group-Experimental	− 0.04	− 1.97 to 1.89	0.966	1.15	− 0.36 to 2.67	0.135	0.36	− 1.08 to 1.81	0.621

Table 36 (continued)

Predictors	Possible not to kill?			Responsible?			Blameworthy?		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
Semester 6-5: Group-Experimental	1.54	− 0.39 to 3.48	0.117	− 0.61	− 2.13 to 0.90	0.426	− 0.55	− 2.00 to 0.90	0.454
Semester 7-6: Group-Experimental	− 1.94	− 3.91 to 0.03	0.054	0.30	− 1.25 to 1.85	0.704	3.41	2.99 to 3.83	**< 0.001**
Random Effects									
σ^2	7.40			4.57			4.19		
$\tau_{00\ \mathrm{ident}}$	7.36			2.41			1.13		
ICC	0.50			0.35			0.21		
N_{ident}	68			68			68		
Observations	443			444			443		
Marginal R^2/Conditional R^2	0.026/0.512			0.027/0.363			0.022/0.229		

Bold is used to indicate statistically significant predictors

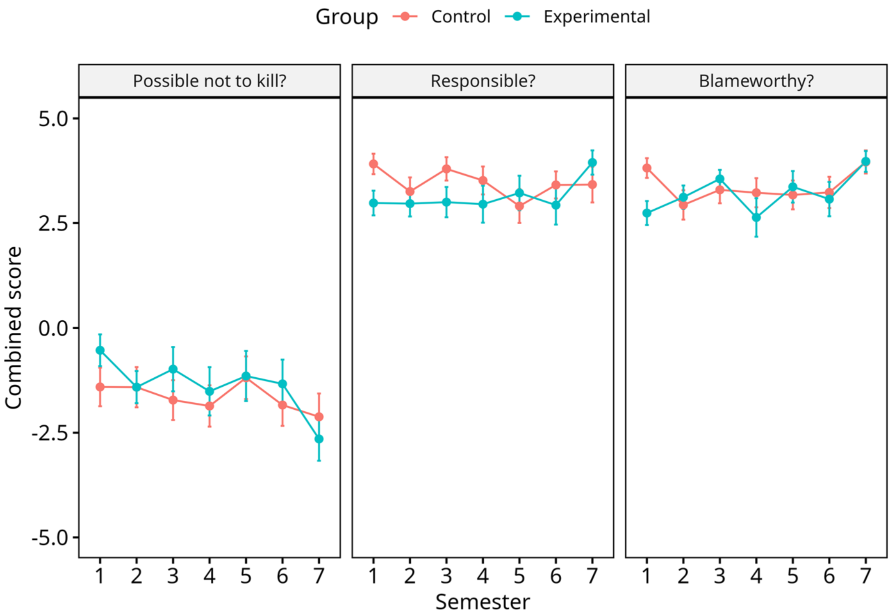

Fig. 31 Changes in intuitions about the Frankfurt case over time. Three panes of the plot correspond to the three questions that were asked. The combined score of + 5 represents a positive answer to a given question with maximum confidence and − 5 represents a negative answer with maximum confidence. Error bars correspond to the standard error of the mean

Teleportation

See Table 37 and Fig. 32.

Table 37 Linear mixed-effects model for the combined scores in the Teleportation case

Predictors	Combined score (Teleportation)		
	Estimates	CI	p
(Intercept)	− 0.52	− 1.42 to 0.38	0.258
Semester 2-1	− 0.20	− 1.25 to 0.85	0.706
Semester 3-2	0.16	− 0.89 to 1.21	0.765
Semester 4-3	1.06	0.02 to 2.11	**0.046**
Semester 5-4	− 0.31	− 1.35 to 0.74	0.564
Semester 6-5	0.27	− 0.81 to 1.34	0.629
Semester 7-6	− 0.17	− 1.28 to 0.95	0.770
Group [Experimental]	0.97	− 0.41 to 2.35	0.167
Semester 2-1: GroupExperimental	0.98	− 0.62 to 2.58	0.230
Semester 3-2: GroupExperimental	− 0.34	− 1.95 to 1.27	0.675
Semester 4-3: GroupExperimental	− 1.89	− 3.51 to − 0.26	**0.023**
Semester 5-4: GroupExperimental	1.23	− 0.40 to 2.85	0.139
Semester 6-5: GroupExperimental	− 0.73	− 2.36 to 0.90	0.380
Semester 7-6: GroupExperimental	− 0.17	− 1.83 to 1.49	0.842
Random Effects			
σ^2	5.25		
$\tau_{00\ \mathrm{ident}}$	7.40		
ICC	0.58		
N_{ident}	68		
Observations	443		
Marginal R^2/Conditional R^2	0.033/0.599		

Bold is used to indicate statistically significant predictors

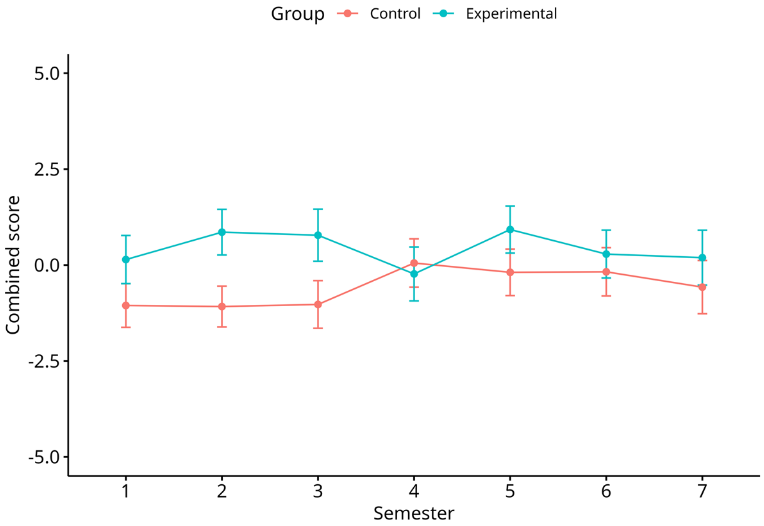

Fig. 32 Changes in intuitions about the Teleportation case over time. The combined score of + 5 represents an answer that neither copy was the original with maximal confidence and − 5 represents the opposite view (other answers) with maximal confidence. Error bars correspond to the standard error of the mean

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beebe, J. R., & Undercoffer, R. J. (2015). Moral valence and semantic intuitions. *Erkenntnis, 80*(2), 445–466. https://doi.org/10.1007/s10670-014-9653-6

Beebe, J. R., & Undercoffer, R. J. (2016). Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture, 16*(3–4), 322–357. https://doi.org/10.1163/15685373-12342182

Bergenholtz, C., Busch, J., & Praëm, S. K. (2023). Further insights on fake-barn cases and intuition variation. *Episteme, 20*(1), 163–180. https://doi.org/10.1017/epi.2021.12

Bochyńska, A. (2021). Badanie eksperymentalne efektu Knobe'a dla języka polskiego. In *Efekt Knobe'a w świetle rozważań językoznawczych i metodologicznych. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera* (Vol. 2, pp. 75–95). Semper.

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme, 11*(2), 199–212. https://doi.org/10.1017/epi.2014.7

De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology, 23*(1), 43–57. https://doi.org/10.1080/09515080903532290

Feltz, A., & Cokely, E. T. (2016). Personality and philosophical bias. In *A companion to experimental philosophy* (pp. 578–589). Wiley. https://doi.org/10.1002/9781118661666.ch41

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

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(23), 829–839. https://doi.org/10.2307/2023833

Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology, 25*(5), 717–742. https://doi.org/10.1080/09515089.2011.627538

Gettier, E. (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

Hales, S. D. (2006). *Relativism and the foundations of philosophy*. MIT Press.

Hindriks, F., & Douven, I. (2018). Nozick's experience machine: An empirical study. *Philosophical Psychology, 31*(2), 278–298. https://doi.org/10.1080/09515089.2017.1406600

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

Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology, 22*(1), 37–60. https://doi.org/10.1080/09515080802703687

Kilov, D., & Hendy, C. (2022). Pundits and possibilities: Philosophers are not modal experts. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2022.2058034

Kim, N. S., Johnson, S. G. B., Ahn, W., & Knobe, J. (2017). The effect of abstract versus concrete framing on judgments of biological and psychological bases of behavior. *Cognitive Research: Principles and Implications, 2*(1), 17. https://doi.org/10.1186/s41235-017-0056-5

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(3), 190–194. https://doi.org/10.1111/1467-8284.00419

Knobe, J. (2019). Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology and Philosophy of Science, 56*(2), 29–36. https://doi.org/10.5840/eps201956225

Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.

Kuś, K., & Maćkiewicz, B. (2021a). Badania strukturalne w filozofii eksperymentalnej. Efekt Knobe'a jako studium przypadku. In *Efekt Knobe'a w świetle rozważań językoznawczych i metodologicznych. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera* (Vol. 2, pp. 166–195). Semper.

Kuś, K., & Maćkiewicz, B. (2021b). Efekt Knobe'a jako iluzja poznawcza. In *Efekt Knobe'a w świetle rozważań językoznawczych i metodologicznych. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera* (Vol. 2, pp. 196–226). Wydawnictwo Naukowe Semper.

Lehrer, K. (1990). *Theory of knowledge*. Routledge.

Löhr, G. (2019). The experience machine and the expertise defense. *Philosophical Psychology, 32*(2), 257–273. https://doi.org/10.1080/09515089.2018.1540775

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*(1), 128–159. https://doi.org/10.1111/j.1475-4975.2007.00160.x

Machery, E. (2012). Expertise and intuitions about reference. *Theoria. Revista De Teoría, Historia y Fundamentos De La Ciencia, 27*(1), 37–54. https://doi.org/10.1387/theoria.3482

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

Machery, E., Deutsch, M., Mallon, R., Nichols, S., Sytsma, J., & Stich, S. P. (2010). Semantic intuitions: Reply to Lam. *Cognition, 117*(3), 361–366. https://doi.org/10.1016/j.cognition.2010.08.016

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

Machery, E., Olivola, C., & De Blanc, M. (2009). Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis, 69*(4), 689–894. https://doi.org/10.1093/analys/anp095

Machery, E., Stich, S. P., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2017). Gettier across cultures. *Noûs, 51*(3), 645–664. https://doi.org/10.1111/nous.12110

Machery, E., Stich, S. P., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2018). Gettier was framed! In *Epistemology for the rest of the world* (pp. 123–148). Oxford University Press. https://doi.org/10.1093/oso/9780190865085.003.0007

Machery, E., Sytsma, J., & Deutsch, M. (2015). Speaker's reference and cross-cultural semantics. In A. Bianchi (Ed.), *On reference* (pp. 62–76). Oxford University Press.

Miller, J. S., & Feltz, A. (2011). Frankfurt and the folk: An experimental investigation of Frankfurt-style cases. *Consciousness and Cognition, 20*(2), 401–414. https://doi.org/10.1016/j.concog.2010.10.015

Mortensen, K., & Nagel, J. (2016). Armchair-friendly experimental philosophy. In *A companion to experimental philosophy* (pp. 53–70). Wiley. https://doi.org/10.1002/9781118661666.ch4

Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition, 129*(3), 652–661. https://doi.org/10.1016/j.cognition.2013.02.008

Nahmias, E., & Murray, D. (2011). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In *New waves in philosophy of action* (pp. 189–216). Palgrave Macmillan. https://doi.org/10.1057/9780230304253_10

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*(4), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind, 125*(497), 145–175. https://doi.org/10.1093/mind/fzv196

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Putnam, H. (1974). Meaning and reference. *The Journal of Philosophy, 70*(19), 699–711. https://doi.org/10.2307/2025079

Putnam, H. (1975). The meaning of "Meaning". In *Mind, language and reality*. Philosophical papers (Vol. 2, pp. 215–271). Cambridge University Press.

Schindler, S., & Saint-Germier, P. (2022). Philosophical expertise put to the test. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2022.2040553

Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition, 20*(4), 1722–1731. https://doi.org/10.1016/j.concog.2011.04.007

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language, 27*(2), 135–153. https://doi.org/10.1111/j.1468-0017.2012.01438.x

Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127–137. https://doi.org/10.1016/j.cognition.2015.04.015

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53*(2), 252–266. https://doi.org/10.1016/0749-5978(92)90064-E

Starmans, C., & Friedman, O. (2020). Expert or esoteric? Philosophers attribute knowledge differently than all other academics. *Cognitive Science, 44*(7), e12850. https://doi.org/10.1111/cogs.12850

Stich, S., & Machery, E. (2023). Demographic differences in philosophical intuition: A reply to Joshua Knobe. *Review of Philosophy and Psychology, 14*, 401–434. https://doi.org/10.1007/s13164-021-00609-7

Stich, S., & Tobia, K. P. (2016). Experimental philosophy and the philosophical tradition. In *A companion to experimental philosophy* (pp. 3–21). Wiley. https://doi.org/10.1002/9781118661666.ch1

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. https://doi.org/10.1111/j.1933-1592.2007.00118.x

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the Land of the Rising Sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology, 6*, 213–230. https://doi.org/10.1007/s13164-014-0206-3

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies, 151*(2), 299–327. https://doi.org/10.1007/s11098-009-9439-x

Thomson, J. J. (1971). A defense of abortion. *Philosophy and Public Affairs, 1*(1), 47–66.

Tobia, K., Buckwalter, W., & Stich, S. (2013a). Moral intuitions: Are philosophers experts? *Philosophical Psychology, 26*(5), 629–638. https://doi.org/10.1080/09515089.2012.696327

Tobia, K., Chapman, G., & Stich, S. (2013b). Cleanliness is next to morality, even for philosophers. *Journal of Consciousness Studies, 20*(11–12), 195–204.

Tobia, K. P., Newman, G. E., & Knobe, J. (2020). Water is and is not H_2O. *Mind and Language, 35*(2), 183–208. https://doi.org/10.1111/mila.12234

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

Turri, J. (2016a). A new paradigm for epistemology: From reliabilism to abilism. *Ergo, 3*(8), 189–231. https://doi.org/10.3998/ergo.12405314.0003.008

Turri, J. (2016b). Vision, knowledge, and assertion. *Consciousness and Cognition, 41*, 41–49. https://doi.org/10.1016/j.concog.2016

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

Turri, J., Buckwalter, W., & Blouw, P. (2015). Knowledge and luck. *Psychonomic Bulletin and Review, 22*, 378–390. https://doi.org/10.3758/s13423-014-0683-5 van Dongen, N., Colombo, M., Romero, F., & Sprenger, J. (2021). Intuitions about the reference of proper names: A meta-analysis. *Review of Philosophy and Psychology, 12*, 745–774. https://doi.org/10.1007/s13164-020-00503-8

Venables, W. M., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.https://doi.org/10.1007/978-0-387-21706-2

Weaver, S., & Turri, J. (2018). Personal identity and persisting as many. In *Oxford studies in experimental philosophy* (Vol. 2, pp. 213–242). Oxford University Press.

Weijers, D. (2014). Nozick's experience machine is dead, long live the experience machine! *Philosophical Psychology, 27*(4), 513–535. https://doi.org/10.1080/09515089.2012.757889

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology, 23*(3), 331–355. https://doi.org/10.1080/09515089.2010.490944

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29*(1/2), 429–460. https://doi.org/10.5840/philtopics2001291/217

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

Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy, 42*(3), 215–229. https://doi.org/10.1111/j.1467-9973.2011.01685.x

Wright, J. C. (2010). On intuitional stability: The clear, the Strong, and the Paradigmatic. *Cognition, 115*(3), 491–503. https://doi.org/10.1016/j.cognition.2010.02.003

Ziółkowski, A. (2021). The stability of philosophical intuitions: Failed replications of Swain et al. (2008). *Episteme, 18*(2), 328–346. https://doi.org/10.1017/epi.2019.20

Ziółkowski, A., Wiegmann, A., Horvath, J., & Machery, E. (2023). Truetemp cooled down: The stability of Truetemp intuitions. *Synthese, 201*, 108. https://doi.org/10.1007/s11229-023-04055-z