Moral Discourse Boosts Confidence in Moral Judgments

Nora Heinzelmann*†, Benedikt Höltgen†, Viet Tran‡

Faculty of Philosophy, Ludwig Maximilian University, Munich, Germany

*nora.heinzelmann@fau.de, Institute for Philosophy, Friedrich-Alexander University of Erlangen-Nuremberg, Germany; ORCID: 0000-0003-4083-3739

† Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich, Germany

‡ Department of Statistics, Ludwig Maximilian University of Munich, Germany

Abstract. *The so-called "conciliatory" norm in epistemology and meta-ethics requires that an agent, upon encountering peer disagreement with her judgment, lower her confidence about that judgment. But whether agents actually abide by this norm is unclear. Although confidence is excessively researched in the empirical sciences, possible effects of disagreement on confidence have been understudied. Here, we target this lacuna, reporting a study that measured confidence about moral beliefs before and after exposure to moral discourse about a controversial issue. Our findings indicate that participants do not abide by the conciliatory norm. Neither do they conform to a rival "steadfast" norm that demands their confidence to remain the same. Instead, moral discourse seems to boost confidence. Interestingly, we also find a confidence*

boost for factual beliefs, and a correlation between the extremity of moral views and confidence. One possible explanation of our findings is that when engaging in moral discourse participants become more extreme in their opinions, which leads them to become more confident about them, or vice versa: they become more confident and in turn more extreme. Although our work provides initial evidence for the former mechanism, further research is needed for a better understanding of confidence and moral discourse.

Biographical notes

Nora Heinzelmann is a junior member of the Faculty of Philosophy at the University of Erlangen, Germany. She works in moral philosophy and related fields, specializing in empirically informed ethics and moral psychology.

Introduction

The debate about moral disagreement concerns the question whether or not the mere fact that an epistemic peer disagrees with one's moral judgment provides *prima facie* evidence against one's belief at issue (Frances and Matheson 2019). According to some recent suggestions (cf. Christensen 2009, Dorst 2019, Kelly 2010, p. 200), peer disagreement may affect not the agent's belief, that is, her first-order mental state. Instead, it may affect a second-order mental state, for instance, how certain she is about her moral belief.

Such meta-cognitive states are a hot research topic in the contemporary cognitive and behavioral sciences and psychology. For, they have been shown to play a crucial role in decision-making, learning, social interaction, and other important human activities. However, to date, relatively little is known about the role of confidence about *moral* beliefs. Conversely, whilst ethicists and epistemologists are primarily concerned with the concepts and norms of confidence, its psychology has remained understudied.

Here, we aim to take a first step to bridge the two strands of research in an attempt to contribute in both directions. On the one hand, we examine the psychological mechanism of confidence in the moral domain. On the other hand, we study how confidence changes when agents are exposed to moral discourse with epistemic peers about a contested topic. We focus on moral beliefs concerning genetic technologies, a highly controversial topic currently widely debated in academia as well as the general public.

In the first part of our paper, we rely on the philosophical literature to develop three empirically testable hypotheses about how moral discourse may affect an agent's confidence. According to the *Steadfast Hypothesis*, moral discourse does not affect an agent's confidence about her first-order moral belief. According to the *Conciliatory Hypothesis*, it lowers her confidence. According to the *Encouraging Hypothesis*, it boosts her confidence. Next, we outline our measures. We developed a survey covering a range of ethical issues taken from the bioethics literature, and validated it in two pilot studies using different designs. We then describe our main study, where we used the survey to measure confidence about moral beliefs before and after discourse with epistemic peers. We argue that our data supports the *Encouraging Hypothesis*: moral discourse seems to boost confidence about one's moral beliefs. We also report a confidence boost for factual beliefs, even if they are wrong. However, as we highlight in our discussion, our study faces several limitations and cannot be taken as more than preliminary evidence. We therefore conclude by outlining directions and questions for future work. Conceptual and empirical researchers can and should collaborate for mutual benefit to understand the influence of moral discourse on second-order mental states such as confidence.

Confidence and moral disagreement

Suppose you discuss with a colleague whether it is morally permissible to edit the genome of human embryos in order to protect them from a fatal disease such as AIDS (Lander et al. 2019): you agree, she disagrees. Both of you provide reasons to justify your views. You believe that, provided worries about risk and abuse are addressed, the expected benefits of this procedure outweigh its potential harms. She thinks that it is in principle morally wrong to interfere with the sanctity of gestation and the human

germ line, benefit or no benefit. You respect your colleague's epistemological and ethical competency but you think that she is mistaken on this particular point. You find the reasons and arguments she provides for her moral belief deeply flawed.

We have described an example for a moral discourse. In the present paper, we follow philosophical custom in using "moral discourse" to refer to a (typically verbal) conversation between two or more agents about a moral issue (Tosi and Warmke 2016, p. 197, Stokke 2016, p. 84, Stalnaker 1999, p. 96). "Moral disagreement" is a similar but importantly different term: it typically refers to a moral discourse whose participants *disagree* about the issue at hand (Willer 2013, Smith 1994), that is, one of them asserts that *p* ("it is morally permissible to edit the genome of embryos") whilst the other asserts that *not-p* ("it is not the case that it is morally permissible to edit the genome of embryos"). The moral discourse in our example is a moral disagreement in this sense. However, besides referring to this discursive *activity* "moral disagreement" may also denote a *state* of disagreement: here, one agent believes that *p* and the other believes that *not-p* but they need not interact (Khoo and Knobe 2016, Cappelen and Hawthorne 2009, McFarlane 2007). In this paper, we shall use "disagreement" in the latter sense. In our example, you and your colleague are in such a state of disagreement, yet you are also having a conversation about it, so you are *in addition* in the process of disagreeing with each other, and you are engaged in a moral discourse.

What is going to happen? What ought to happen? There are different options for discourse between two parties in disagreement. A popular one is to lower one's *credence* in the contested proposition. Even though you take issue with the evidence and reasons your colleague provides, *the mere fact that she disagrees with you* might

cause you to revise. Perhaps, after talking to your colleague, you believe to a lesser degree that genome editing is morally permissible. Maybe you suspend judgment about it altogether. At least, perhaps this is what you ought to do, anyways. In both cases, your first-order mental state changes. Conversely, it seems entirely conceivable that you do not alter your moral belief at all, and that there is nothing wrong with that.

In contrast, other possible reactions concern the second-order mental state of the agent, such as beliefs about one's beliefs, conviction, confidence, etc. To illustrate the difference between first- and second-order mental states, imagine you are asked your credence that a fair coin, when thrown, will come up heads. Your answer "0.5" concerns your first-order mental state. Imagine you are also asked how confident or certain you are about that answer. Given that the coin is fair, you might be highly confident (perhaps even 100% confident). The second question concerns your second-order mental state: a mental state about a mental state, here your confidence about your credence.

Some philosophical approaches to moral discourse focus on confidence. For instance, instead of changing your belief that genome editing is morally permissible, you may or should change how confident or certain you are about it. On what we may call the "*conciliatory view*", following Christensen (2009), moral discourse leads an agent to lower her confidence. Perhaps your confidence will or ought to drop. Assume you still believe that it is morally permissible to edit the genome of human embryos. But whilst you were nearly 100% certain about that before talking to your colleague, you are only 80% certain afterwards. Conversely, maybe peer disagreement does not and should not affect how confident you are. On the *"steadfast view"*, your confidence need not change. You remain as certain about your belief as you were before talking to your

colleague. In addition to the steadfast and the conciliatory views, there is a range of middle positions in between them.

Epistemological accounts like the conciliatory or steadfast views apply to non-moral as well as moral beliefs. That is, the conciliatory view advises agents to lower their confidence about, say, the factual belief that Tbilisi is the capital of Georgia when they learn that a peer disagrees with them. In contrast, the steadfast view allows them to retain their confidence in the face of disagreement. From this perspective, there is no crucial difference between moral and non-moral beliefs, moral and non-moral disagreement, or moral and non-moral discourse.

Whilst a philosophical debate is raging over what the normatively adequate position is on the continuum between conciliatory and steadfast views, philosophers have to date paid relatively little attention to empirical findings concerning the conditions and factors that affect an agent's confidence. In fact, so far, there seems to be no empirical support for either of the two views. This is surprising given that we may not be able to change our confidence at will, and thus our normative requirements may be constrained (Griffin 2010): if we *cannot* alter our confidence about a belief, then it is plausibly not the case that we *ought* to do so (Hume, *Treatise* 3.1.1). Therefore, understanding the empirical mechanism of confidence might advance conceptual research in philosophy.

Confidence understood as a second-order (or "metacognitive") mental state has recently been the object of extensive research in the cognitive and psychological sciences (see Fleming 2014 for an overview). Mounting evidence suggests that it plays a key role in decision-making and social interaction. For instance, it has been found that the degree

of confidence in one's perceptual belief correlates with actual accuracy (Bahrami et al. 2010). Confidence about factual and preferential beliefs seems to reflect the degree of consensus with one's peers (Koriat 2012, Krosnick et al. 1993). Agents with greater confidence are more persuasive in face-to-face interactions (Schwardman and Weele 2019).

Moreover, there is ample research on belief updating, both moral and non-moral (Holyoak and Powell 2016, Horne et al. 2013). There may or may not be differences between moral and non-moral cognition (Cushman and Young 2011, Hauser and Young 2008), yet we need to bracket this debate here. Relevant for our current study, a key role in belief-updating has emerged for social influences. For instance, Haidt (2001) famously suggested that intuitions and judgments are crucially determined by social influences. The main purpose of reasoning may be to develop arguments that persuade others (Mercier and Sperber 2011), and Bayesian models justify social influences on belief updating (Bovens and Hartmann 2003). Lastly, it has long been known that group discussions lead to polarization of opinions (Isenberg 1986, Myers and Lamm 1975).

However, empirical research into confidence has so far paid little attention to what philosophers call "*moral discourse*", i.e., discourse about moral beliefs. Ethicists and epistemologists alike have been particularly interested in this kind of conversation because of its impact on choice and action. For instance, relatively little will hinge on whether you or your brother is right about the name of your elementary school teacher. But how we resolve debates about moral questions concerning, e.g., future generations or the treatment of animals, may impact our political and private lives. What is more,

even if we suspend judgment about an action or merely partially endorse it: eventually we either do it, or we do not do it (Feldman 2006).

There are thus converging interests of empirical and philosophical research into confidence: on the one hand, scientists and psychologists might wish to include moral discourse amongst the factors potentially influencing metacognition. On the other hand, the psychological mechanism that determines one's confidence about moral beliefs may directly constrain the epistemological norms concerning moral discourse and disagreement.

The present paper aims to take a first step to connect the two field, and to contribute some initial data. Specifically, we report findings from an investigation into whether and how moral discourse affects one's confidence about a moral belief. We measured confidence about moral judgments in two socioeconomically similar groups before and after exposure to moral discourse during a two-day workshop. In the target or experimental group, participants engaged specifically with the content of the moral beliefs. In the control group, participants were preoccupied with an unrelated, non-moral topic. As a benchmark, we also asked participants to indicate their confidence about a range of factual beliefs that were either related to the workshop, or unrelated.

If human psychology conforms to the conciliatory view, our participants in the target group ought to report lower levels of confidence about their first-order views after moral discourse than before. This should be the case both moral beliefs and factual beliefs. This hypothesis enjoys some support from previous research which indicates that people conform to the conciliatory view under certain circumstances. For example,

McGarty et al. (1993) report that participants lower their confidence about a visual task when they receive incongruous evidence. In another study, participants were less certain about a judgement when they believed that someone they considered similar to themselves expressed disagreement with them (Orive 1988).

Conversely, if their confidence abides by the steadfast view, one can expect that participants in the experimental group do not alter their confidence. It has been shown that we regard ethical statements almost as objective as states of affairs (Heiphetz and Young 2016) and scientific facts (Goodwin and Darley 2012), and that we believe we need not be wrong when disagreeing with someone else in conversation (Khoo and Knobe 2016). Consequently, encountering opposition or challenges in moral discourse might not motivate us to sway in our confidence.

Furthermore, evidence suggests that we regard attitudes we are highly convinced of as more objective, that we are less likely to compromise on them, and that they are less susceptible to social influence (Skitka et al. 2015, Skitka 2010, Skitka et al. 2005). Moral conviction is the second-order belief that a first-order belief is based on one's core moral beliefs. Confidence about a moral belief is thus somewhat different from moral conviction. Not all attitudes about which a person is highly confident amount to attitudes high in moral conviction. For instance, although Linda is highly confident that eating factory-farmed beef is morally worse than eating organically grown spinach, this moral belief might not be based on her core beliefs. Nevertheless, one might expect a certain degree of overlap between moral conviction and moral confidence. One would thus expect, in-line with the steadfast view, that participants' confidence levels about moral beliefs remain unchanged when they are exposed to social influence.

A third option is also conceivable: perhaps participants in the target group report higher levels of confidence after moral discourse. As they were given the opportunity to interact and exchange information, they might seek out evidence supporting their own view. This phenomenon has become known as the "confirmation bias" (Nickerson 1989). In the context of the present paper, it should be noted that this bias has also been observed in children and students (Schauble 1990, Greenhoot et al. 2004). In addition, as experts have long been known for overconfidence in their judgment in many cases (Tetlock 2017, Kahneman 2011, Meehl 1954), participating in the workshop might lead the target group to believe they gained expertise on the topic, and in turn make them feel more confident. As a consequence, both mechanisms might encourage and boost participants' confidence about first-order beliefs which would now be better justified than before.

We can thus formulate three rival hypotheses:

1. **Conciliatory Hypothesis.** Confidence is lower after exposure to moral discourse than before.

2. **Steadfast Hypothesis.** Confidence is about the same after exposure to moral discourse than before.

3. **Encouraging Hypothesis.** Confidence is higher after exposure to moral discourse than before.

In our study, we collect data that can speak to the debate about moral discourse by providing evidence for any of the three hypotheses. The next section explains our design and measures.

Study

Rationale and material

We explored the role of moral discourse and social interaction for the circumscribed domain of bioethics. Specifically, we focused on ethical issues arising from novel genetic technologies, such as genome editing and genetic testing. These technologies are still relatively new and developing rapidly, opening up new possibilities for medical treatment and beyond. We investigated to what extent, if any, individuals change their minds on bioethical questions when they are exposed to others' opinions and group discussions. Our studies were approved by the Ethics Committee of the University of Konstanz (approval no. 33/2018).

We surveyed high school students before and after they attended a two-day workshop on ethical issues concerning novel genetic technologies. The event exposed them to a range of ethical views and arguments through instruction and peer interaction. We measured and compared moral beliefs and confidence about these beliefs before and after the workshop, respectively. Additionally, we compared these measures to those of a control group who attended a workshop on an unrelated topic. The online questionnaire we used had been previously developed and validated in two pilot studies.

Pilot studies

For the first pilot, we developed an online questionnaire based on the bioethical literature, identifying eight categories of moral issues surrounding novel genetic technologies: genetic testing of humans, genetic testing of non-humans, genome editing, moral responsibility for events partially determined by one's genetic makeup,

protection of genetic data, social justice concerning, e.g., the development of genetic technologies, personalized medicine (which often targets a tumor's or patients' genetic makeup), and reproductive technologies involving genetic profiling. For each category, we devised two vignettes and two statements, using real-world examples and cases. For instance, one statement read:

> "*Genetic tests are ethically impermissible even if a hereditary disease runs in a family.*"

The matched vignette read:

> "Jennifer is planning to conceive a child. She knows that severe hereditary diseases run in her family. *Jennifer is ethically required to perform a genetic test prior to conception.*"

Our final questionnaire thus comprised 32 items in random order. For each, participants indicated on a 6-point Likert scale to what extent they agreed or disagreed with the italicized moral claim.

We recruited participants through Amazon's online platform Mechanical Turk (MTurk). Excluding those who failed multiple attention checks, we analyzed data from 79 participants (27 females, mean age 35.2, SD 9.9 years). In addition to gender and age, we also collected information about participants' religiosity, education, political affiliation, and prior exposure to genetic testing or cancer (where personalized medical procedures are widely used). The questionnaire and data are available on the Open Science Framework (OSF): https://osf.io/xntm9/

Our main findings were consistent with the literature. We found that, on average, participants endorsed genetic technologies, but moral permissibility ratings varied widely with context of application and individual traits (Gaskell et al. 2017, Scheufele et al. 2017, Weisberg et al. 2017, McCaughey et al. 2016, Blendon et al. 2016, Morren et al. 2007). For instance, we found that women were more opposed to genetic testing of animals than men, which is in line with earlier research indicating that women show more empathy towards non-human animals (Lehmann et al. 2013). Strikingly, our data revealed that participants were quite divided over ethical issues (Fig. 1, 1st pilot), indicating large moral disagreement in our sample.

Our analysis did not reveal any differences between responses to vignettes and responses to statements. For this reason, we only used statements in subsequent studies. We also decided to drop the category "personalized medicine" because we realized that most participants did not know much about this topic and had hardly any ethical opinions about it.

For the second pilot study, we recruited 70 un-paid participants from the academic community (37 females, mean age 34.7, SD 10.4 years), partially at a summer school in Athens and partially through social media like Facebook. The questionnaire consisted of the statements from seven categories used in the previous pilot study. To increase statistical power, we added three new items per category, so that the survey comprised 35 items overall. Again, we collected responses on the extent to which participants agreed or disagreed with an ethical claim on a 6-point Likert scale.

We employed a contrastive design, which has been widely used in experimental philosophy (e.g., Fitz et al. 2014, Knobe 2003, Burstin et al. 1980). This design controls for information experimenters are not interested it, and for coincidental features like wording. In our study, half of the participants received the above-mentioned (target) statements about genetic technologies. The other half received contrast statements which were identical to the target statements, except that they did not concern genetic but comparable technologies. For instance, one of our target statements read:

> "It is ethically permissible to perform genetic tests on a breeding horse before offering it at an auction."

The corresponding contrast statement read:

> "It is ethically permissible to perform radiological screenings on a breeding horse before offering it at an auction."

Both techniques are actually used for horse auctions, but only one of them employs a genetic technology. The contrastive design thus allowed us to specifically attribute any difference in responses between target and control groups to the difference in technology.

Importantly, the (target) group that received statements about genetic technologies and the (control) group that assessed claims about non-genetic technologies were sufficiently similar to each other: average age was 32.9 in the target and 36.6 in the control group, ($t(54) = -1.46, p = .15$); there were 18 females in the target and 19 in the control group, ($X^2(1) = .15, p = .70$). Hence, any difference in responses was presumably due to the difference in statements and not due to a difference between the two groups.

This allowed us to conclude, for instance, that participants have substantial ethical concerns about the protection of genetic data. For, ethical ratings by the target group were significantly higher than those by the control group, and the former was concerned with protection of genetic data whilst the latter was asked about protection of non-genetic data.

In addition, we again found that ethical views varied with the context of application, thus confirming our findings from the first pilot study and reports in the empirical literature (McCaughey et al. 2016, Gaskell et al. 2017, Scheufele et al. 2017). As before, participants were divided in their first-order views (Fig. 1, 2nd pilot), indicating substantial moral disagreement in our sample. [Figure 1 around here]

Participants

For our main study, we recruited teenagers attending an extracurricular training program run by the German Cancer Research Center in Heidelberg, Germany. We also obtained written consent from their parents. Methods and materials for this study are available on the OSF (https://osf.io/23ajd/).

Participants registered for and attended either of two workshops, thereby assigning themselves to the treatment or the control group. As this procedure falls short of a randomized controlled trial, it naturally leads to worries about the validity of our results. In particular, participants who signed up for the different workshops might have done so because of different prior knowledge or relevant ethical convictions. To alleviate such worries as much as possible, we checked for initial differences in ethical views and factual knowledge as well as confidence therein. We did not find any significant

differences except for participants' confidence regarding ethical items like the claim that it is ethically permissible to use public funding on expensive gene therapies. Here, participants indicated slightly higher confidence (p = .045, t(1064.8) = 2.00) in the treatment group (M = 71.1, SD = 24.53), as compared to the control group (M = 68.4, SD = 20.73). As this difference is quite small, especially in comparison to the pronounced results we report below, we believe it unlikely to have played a major role. Nevertheless, we cannot exclude the possibility that the selection procedure has influenced our results.

We excluded data from two participants in the treatment group because they consistently clicked only on the far left or on the far right of the rating scales. This criterion was not pre-registered. To check whether including these participants would affect the results, we re-ran all analyses with these participants included. All significant results remained significant (p<.05), and all highly significant results remained highly significant (p<0.001). We thus analyzed data from 17 participants (12 females; mean age: 15.7 years) in the treatment group and from 16 in the control group (10 females; mean age: 16.6 years). We did not find evidence for differences between the two groups with regard to age (t(31) = -1.91, p = .07), gender (X^2(1) = 0.02, p = .90), or socioeconomic status (Z = 3.86, p = .70). Therefore, any differences in ethical judgments or confidence is more likely due to the different workshops that the two groups attended, and not due to differences in age, gender, or socioeconomic status.

Design and procedure

The workshops took place during weekends (Friday through Sunday). They covered normative issues surrounding genetic technologies (treatment group) or project

management (control group). All participants engaged in small group and plenary discussions and peer-to-peer presentations. For the treatment group in particular, the stated goal was to develop a differentiated and informed view on ethical issues arising from novel genetic technologies, which they put in writing. During the event, participants were first provided with written material and expert presentations on different aspects of the topic: scientific research, medical research and treatment, big data, moral psychology, and ethical theory. Afterwards, participants split into smaller groups for discussion. They then came up with arguments in favor and against genetic technologies. Based on these, participants wrote short evaluations and recommendations for legal, medical, policy, and scientific guidelines. These were finally presented to and discussed by the whole group.

To measure bioethical opinion and confidence, we administered an online survey. All participants completed the survey twice: first on the Friday before and then during the week after the workshop. Each participant assigned themselves a pseudonym so we could connect the pre- with the post-workshop responses.

The survey was part of a larger questionnaire that also measured identity goal activation and self-complexity. The part relevant to the present paper consisted of 35 ethical statements and 12 factual statements in randomized order, followed by questions about age, gender and parents' education. The ethical statements were identical to the ones administered in the second pilot study and covered seven categories like genome editing, social justice or data protection. For each of the statements, participants were asked to indicate their agreement on a 6-point Likert scale ranging from "strongly agree" to "strongly disagree". They were also asked to indicate their level of confidence

on a continuous rating scale from 0 to 100 (Fig. 2)[Figure 2 around here]. The orientation of agreement and confidence scales was randomized between participants in order to control for any systematic biases.

Analyses and results

We conducted a paired t-test for both the factual and ethical items, and for each group of participants (experimental and control) to detect any systematic changes in judgments and confidence before vs. after the workshop. For some of the categories, the responses were not normally distributed, so we performed Wilcoxon signed rank tests for these, and t-tests for the others.

First, we investigated whether our sample was in a state of moral disagreement about the issues raised by the questions of our survey. Fig. 1 depicts the spread of first-order responses before the workshop, collapsed for both treatment and control group. Our data indicate substantial moral disagreement, replicating the findings of our two pilot studies.

Second, to directly test the three main hypotheses outlined above, we conducted paired t-tests on the pre- and post-workshop confidence data within the experimental and the control group, respectively. For these and all further paired t-tests, one data point consisted of one response given by one participant before the workshop and another response by this participant to the same question afterwards. For instance, each participant in the control group (N=16) provided one pair of ratings for the 35 ethical items, yielding 560 data points overall.

For moral beliefs of participants in the experimental group, we found a significant confidence boost (Fig. 3)[Figure 3 around here]. While the paired t-test in the control group revealed no significant difference (t(559) = -1.48, p = .14) before (M = 68.4%, SD = 24.5%) versus after (M = 69.9%, SD = 23.5%) the workshop, the same test was highly significant in the treatment group (t(513) = -8.81, p < .001) before (M = 71.1%, SD = 20.7%) versus after the workshop (M = 80.1%, SD = 17.3%). These results indicate that participants' confidence levels rise when they are exposed to peer disagreement. Our data thus support the Encouraging Hypothesis and challenge the two rival Conciliatory and Steadfast Hypotheses.

A similar pattern emerged for confidence about factual beliefs (Fig. 4a), which concerned claims like "side-effects of an action are defined as the ethically irrelevant effects of an action". Again, we found no significant difference in confidence levels for the control group before (M = 50.4%, SD = 33.3%) compared to after (M = 49.8%, SD = 33.1%) moral discourse (t(191) = 0.26, p = .79). However, participants in the treatment group showed a highly significant (t(171) = -9.37, p < .001) confidence boost, with higher levels of confidence after (M = 73.7%, SD = 26.8%) than before (M = 51.3%, SD = 27.9%) moral discourse. Interestingly, this was also true for factual statements about which participants had false beliefs (t(44) = -4.65, p < .001, Fig. 4b[Figure 4 around here]): mean confidence levels were at 42.6% (SD = 28.9%) before and 68.5% (SD = 27.8%) after moral discourse.

On a side note, participants' self-reported levels of confidence were relatively high for all domains and conditions (the smallest average confidence was around 50% in the factual and 70% in the ethical domain). Confidence about ethical judgements was

higher than confidence about factual ones. In the control group, ethical confidence (M = 68.4, SD = 24.5) was significantly higher (t(265.56) = 6.87, p < .001) than factual confidence (M = 50.4, SD = 33.3) before the workshop. Those values remained virtually unchanged after the workshop. Similarly, in the treatment group, ethical confidence (M = 71.1, SD = 20.8) was significantly higher (t(232.7) = 8.32, p < .001) than factual confidence (M = 51.3, SD = 28.9) before treatment. Here, the difference between ethical (M = 80.1, SD = 17.3) and factual confidence (M = 73.7, SD = 26.8) was smaller after treatment but confidence remained significantly higher for ethical judgments than for factual ones (t(220.8) = 2.94, p<.01).

Not only confidence, but also the extremity of participants' first-order responses increased during the workshop. Before treatment, 42% of all ethical responses were moderate (in the middle of, i.e. 3 or 4 on, the 6-point Likert scale), 39% were intermediate (2 or 5), and 19% were extreme (1 or 6). Afterwards, moral beliefs were more extreme, with a distribution of 37%, 38%, and 26%, respectively. This results in a significant (paired) Wilcoxon signed rank test (p < .01, V = 21085).

We also find this shift towards more extreme responses in the factual domain (p < .001, V = 1350), with a distribution shift from (53%, 33%, 15%) to (24%, 39%, 37%). We did not observe this effect in the control group. Within limits, our data might be relevant to the literature about the backfire effect, the finding that exposure to a factually incorrect claim followed by corrective information results in less accurate belief as compared to exposure to the incorrect claim only (Wood and Porter 2019, Nyhan and Reifler 2010). In our case, a backfire effect would have been a finding that only in the experimental group that was exposed to corrective information, false factual beliefs

increased relative to the control group. We do not find such an effect, in-line with recent failures to replicate the backfire effect studies (Wood and Porter 2019, Haglin 2017). In fact, we find a significant difference in the reduction of false factual beliefs between the treatment group and to the control group ($X^2(1)=7.87$, p=.005).

We also examined the relationship between first-order beliefs about factual or ethical statements and confidence. These concerned claims like "personalized cancer therapies always require sequencing the entire genome of cancer cells". Interestingly, we found a highly significant relationship between these two across all conditions and groups: extremity of first-order belief, positive as well as negative, was positively correlated with level of confidence (Fig. 5). In both the treatment ($r_s = .62$, p < .001) and the control group ($r_s = .57$, p < .001), the Spearman rank order correlation coefficient indicated a moderate or even strong correlation both before and after the workshop, and for both factual and ethical beliefs.

Discussion

Our study investigates the effect of moral discourse and social interaction on bioethical judgments and confidence. We measured these by administering a questionnaire before and after a workshop about bioethics to a group of teenagers, as well as to a similar control group who participated in a workshop not related to bioethics.

Our research has three main findings. First, it confirms the *Encouraging Hypothesis*: moral discourse did boost confidence in the target group. The control group, which did not discuss moral issues, does not show this effect. This indicates that participants neither conform to the conciliatory nor to the steadfast view. Whether they *could* have

lowered their confidence or have left it unchanged is a question our data do not speak to. However, they indicate that the psychological mechanism of confidence change may deviate from the normatively recommended path: it neither decreases nor remains the same but increases when an agent engages in moral discourse with peers.

Our second finding was that confidence about both correct and mistaken factual beliefs was higher after exposure to countervailing evidence. One possible explanation for this effect might be that bioethical beliefs have a moral as well as a factual component, and that when discussing moral beliefs both these components are discussed. For instance, the bioethical belief that editing the genome of human embryos is morally bad may rest on the moral belief that inflicting pain on embryos is morally wrong as well as the factual belief that genome editing inflicts pain on embryos. A discussion about the bioethical belief may concern the moral issue of whether inflicting pain is morally wrong or whether it might be permissible in some circumstances (say, when it prevents greater harm in the future). Yet it might also concern the factual issue of whether genome editing does inflict pain on embryos or not. It is thus highly likely that participants engaged with facts relevant to the present bioethical issues, and that this in turn encouraged them to raise their confidence about factual beliefs as well. Furthermore, even for those answers to factual questions that participants answered *incorrectly*, confidence levels rose. This may warrant caution about hopes that education or fact-based discussions could mitigate an agent's overconfidence about her false views.

The third result was unexpected. We found a striking relationship between the first-order moral beliefs and confidence across all groups and conditions (Fig. 5) [Figure 5

around here]: the more extreme a participant's moral belief (in either direction), the stronger was her confidence. Conversely, for moderate views, reported confidence was significantly lower. This resonates with the commonplace claim that extremists have the loudest voices. However, this relationship between extremity of moral belief and reported confidence has to date not been reported in the empirical literature. We therefore advice to treat this finding with due caution and explore it further in unpublished research.

Naturally, the pioneering and thus highly circumscribed nature of our research allows merely for circumscribed conclusions. In particular, our study has at least the following three limitations. First, because the number of attendees for both workshops was restricted, our sample size was small (in total, we analyzed data from 33 participants). In addition, we studied a highly selective sample, namely high-school students from an affluent and well-educated background. Relatedly, the age difference between the samples of our main study and the pilot studies might have influenced our results. Follow-up studies with larger and more representative samples are thus needed to examine whether the effects we find generalize. Our data should thus be seen as preliminary evidence for an effect of peer interaction on confidence about ethical judgments; the psychological mechanisms and generalizability of this effect is to be explored in further research.

A second limitation of our research is its lack of specificity: whilst we have good reasons to assume that participants in our treatment group were exposed to controversial discussion with peers, we did not actively manipulate discourse as an independent variable. Breaking down our small groups further into different conditions would very

likely have reduced our statistical power to a useless minimum. In consequence, we cannot know whether it was, say, the amount of new evidence, the strength of (counter)arguments, or the proportion of supporting and contradicting arguments or evidence, that determined the confidence boost. All these questions must be left as opportunities for future and more specific research.

An alternative explanation to a direct effect of moral discourse could be that exposure to discussion and new information may have led to a polarization effect, as we found that participants' views only in the treatment group became more extreme. As we also found that extremity correlated with confidence, polarization might have ultimately caused the confidence boost. A third possible mechanism is also conceivable, as our participants were exposed to arguments in favor and against genetic technologies. It is thus possible that information supporting participants' initial views contributed to the confidence boost.

To gain some insight into the relationship between the boost in extremity and the boost in confidence, we conducted four exploratory regression analyses[1]. If extremity of belief determines confidence boost, we would expect significant results when confidence is modelled as a function of belief. If confidence boost determines extremity of belief, we would expect significant results when belief is modelled as a function of confidence. Therefore, for ethical and factual beliefs of each participant, we separately modelled belief and confidence at t2 (after the workshop) as a function of

- belief at t1 (before the workshop),
- extremity at t1 (extremity_t1),

---

[1] We thank Madeline G. Reinecke for this suggestion.

- extremity at t2 (extremity_t2),

- the absolute difference between belief before and after the workshop (belief_diff),

- gender, and

- age.

We then ran the following four models:

Model 1 for moral beliefs: confidence_t2 ~ confidence_t1 + belief_t1 + extremity_t1 + extremity_t2 + belief_diff + gender + age

Model 2 for factual beliefs: confidence_t2 ~ confidence_t1 + belief_t1 + extremity_t1 + extremity_t2 + belief_diff + gender + age

Model 3 for moral beliefs: belief_t2 ~ confidence_t1 + belief_t1 + extremity_t1 + extremity_t2 + belief_diff + gender + age

Model 4 for factual beliefs: belief_t2 ~ confidence_t1 + belief_t1 + extremity_t1 + extremity_t2 + belief_diff + gender + age

The results are reported in Table 1 [Table 1 around here], and we provide further details on the OSF website[2].

We find statistically significant effects of extremity at t2 on confidence at t2 for both factual and moral beliefs (models 1 and 2). We also find an effect of moral belief at t1 on confidence at t2 (model 1). But we do not find effects of confidence at t1 on belief at t2 for either the moral or the factual domain (models 3 and 4). Neither do we find any significant effect of belief difference in any model.

---

[2] https://osf.io/23ajd/

These results indicate that extremity does determine confidence but not vice versa because we find effects of extremity and belief on confidence yet no evidence for effects of confidence on belief. There are two further though not sufficient reasons for the directionality from extremity to confidence. First, in the temporal order of our task, we asked about belief first and then about confidence. Thus it would seem plausible that changes in belief affect confidence but not vice versa. Second, the literature on polarization documents that discourse increases extremity (Isenberg 1986, Myers and Lamm 1975) but has not reported a boost in confidence, indicating that one may observe the former without the latter.

We find effects of confidence at t1 on confidence at t2 for factual beliefs only (model 2), and of belief at t1 on belief at t2 (models 3 and 4). Extremity at t1 and t2 determines belief at t2 for the moral and factual domain, respectively (models 3 and 4). These findings plausibly reflects consistency of participants' confidence and beliefs over time.

Interestingly, we find gender effects for confidence at t2 about factual beliefs (model 2) and for ethical beliefs at t2 (model 3). We think that this may be due to the gender imbalance in our samples, as 71% and 63% of participants self-identified as female in the target and control groups, respectively. Therefore, we caution against taking our results as evidence of a gender bias.

Overall, then, whilst our data fail to support the Conciliatory and Steadfast Hypotheses, further research beyond our exploratory study is needed to challenge them. In particular, we aim to directly test for and compare agreement and disagreement conditions in a future study.

A third caveat concerns our measures for first- and second-order judgments. Although we explicitly separated the two in our instructions, it is possible that participants already indicated their level of confidence when giving their first-order ratings, perhaps even unintentionally so. This raises the worry that our measures were not completely separate, casting doubt on the finding that confidence correlates with extremeness of first-order view. A deeper underlying issue might be that first-order judgments are partially influenced by second-order confidence, or vice versa (but see Rollwage et al. 2020). That is, a boost in extremity of first-order judgments might have caused a boost in confidence, or the other way around. Either way, at this stage we can neither rule out nor confirm this interdependence. Hence, we merely raise it here as a hypothesis to be tested explicitly in future research.

In sum, the work we report here aims to uncover rather than decide between several possible avenues for future research. In particular, it tentatively suggests that exposure to moral discourse boosts confidence about one's moral beliefs. It also hopes to demonstrate that interdisciplinary investigations combining conceptual approaches from social epistemology with experimental ones from the cognitive sciences may provide insights useful to both disciplines.

Conclusion

How should agents adjust their views when engaging in moral discourse? According to a recent suggestion in epistemology, they ought to lower their confidence about the moral belief at issue when encountering disagreement. However, the moral psychology of confidence has so far received little attention in the philosophical community,

although it has recently become a hot topic in empirical research under the label of "metacognition". Nevertheless, empirical research into the effect of moral discourse on metacognition has so far been underdeveloped.

Connecting these two strands of research, our paper aims to take a first step towards understanding the metacognition of moral discourse. Specifically, we measured confidence about moral beliefs (e.g., concerning the claim that it is ethically permissible to request genetic tests from self-proclaimed heirs as evidence of descent) before and after participants were engaged in moral discourse with peers, and in a control group. Strikingly, we found that confidence about moral beliefs was higher after than before discussion. Importantly, we do not find this effect in the control group, which indicates that it is due to the treatment and not to coincidental aspects like repetition or temporal delay.

Our findings thus indicate that agents do not follow the epistemological norm that requires them to lower their confidence when encountering moral disagreement in discussion. Neither do they abide by a norm that demands their confidence to remain the same. Instead, interaction with others seems to encourage their previously held ethical beliefs: these become more extreme, and participants become more confident about them. This, in turn, could be explained by participants' impression to have become more knowledgeable about the topic and therefore overconfident.

These suggestions about a possible psychological mechanism of the impact of social interaction are speculative and need to be tested explicitly in future research. The pioneering nature of our study comes at a cost of further limitations. For one thing, our

samples were small and not representative. Our work thus cannot be understood as anything but a first step towards a full understanding of the role of confidence and social interaction, and a pointer to directions for further investigation. Nevertheless, we hope to have done not just that but also demonstrated that joint experimental work into moral discourse and confidence about moral beliefs holds promise for philosophy and the behavioral sciences alike.

Acknowledgments

Declaration of interests

We declare that we have no potential competing interest.

Bibliography

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010): "Optimally interacting minds", *Science* 329(5995): 1081–5.

Blendon, R. J., Gorski, M.T., and Benson, J.M. (2016): "The public and the gene-editing revolution", *New England Journal of Medicine* 374(15): 1406-1411.

Bovens, L., and Hartmann, S. (2003): *Bayesian Epistemology*. Oxford: Oxford University Press.

Burstin, K., Doughtie, E.B., and Raphaeli, A. (1980): "Contrastive vignette technique: An indirect methodology designed to address reactive social attitude measurement", *Journal of Applied Social Psychology* 10(2): 147-165.

Cappelen, H., and Hawthorne, J. (2009): *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Christensen, D. (2009): "Disagreement as evidence: the epistemology of controversy", *Philosophy Compass* 4(5): 756–67.

Cushman, F., and Young, L. (2011): "Patterns of moral judgment derive from nonmoral psychological representations", *Cognitive Science* 35(6): 1052-1075.

Dorst, K. (forthcoming) "Higher-Order Uncertainty", In: M. Skipper Rasmussen and A. Steglich-Peterson (eds.), *Higher-Order Evidence: New Essays*. Oxford University Press.

Feldman, R. (2006): "Clifford's Principle and James' Options", Social Epistemology 20: 19–33.

Fitz, N., Nadler, R., Manogaran, P., Chong, E.W.J., Reine, P.B. (2014): "Public attitudes toward cognitive enhancement", *Neuroethics* 7(2): 173–188.

Fleming, S. (2014): "The power of reflection", *Scientific American Mind* 25(5): 31–37.

Frances, B., and Matheson, J. (2019): "Disagreement", in: E. Zalta (ed.): *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition)

Gaskell, G., Bard, I., Allansdottir, A., da Cunha, R.V., Eduard, P., Hampel, J., Hildt, E., Hofmaier, C., Kronberger, N., Laursen, S., Meijknecht, A., Nordal, S., Quintanilha, A., Revuelta, G., Saladié, N., Sándor, J., Santos, J.B., Seyringer, S., Singh, I., Somsen, H., Toonders, W., Torgersen, H., Torre, V., Varju, M., and Zwart, H. (2017): "Public views on gene editing and its uses", *Nature Biotechnology* 35(11): 1021.

Goodwin, G. P., and Darley, J. M. (2012): "Why are some moral beliefs perceived to be more objective than others?", *Journal of Experimental Social Psychology* 48: 250–256.

Greenhoot, A.F., Semb, G., Colombo, J., and Schreiber, T. (2004) "Prior beliefs and methodological concepts in scientific reasoning", *Applied Cognitive Psychology* 18.2: 203-221.

Griffin, J. (2010): "'Ought' implies 'can'", *The Lindley Lecture*. University of Kansas.

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

Hauser M., and Young L. (2008): "Modules, Minds and Morality", In: Pfaff D., Kordon C., Chanson P., Christen Y. (eds): *Hormones and Social Behaviour. Research and Perspectives in Endocrine Interactions*. Heidelberg: Springer.

Heiphetz, L., and Young, L. (2017): "Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs", *Cognition* 167: 78-90.

Holyoak, K. J., and Powell, D. (2016): "Deontological coherence: A framework for commonsense moral reasoning", *Psychological Bulletin* 142(11): 1179.

Horne, Z., Powell, D., and Spino, J. (2013): Belief updating in moral dilemmas. *Review of Philosophy and Psychology*, 4(4): 705-714.

Hume, D. (1740 [2000]): *A Treatise of Human Nature*. D. Norton and M. Norton (eds.), Oxford: Oxford University Press.

Isenberg, D. J. (1986): "Group polarization: A critical review and meta-analysis", *Journal of Personality and Social Psychology* 50(6): 1141.

Kahneman, D. (2011): *Thinking, fast and slow*. Penguin.

Kelly, T. (2010): "Peer disagreement and higher order evidence", in: A. Goldman and D. Whitcomb (eds.), *Social Epistemology: Essential Readings*. Oxford: Oxford University Press, 183-217.

Khoo, J., and Knobe, J. (2018): "Moral disagreement and moral semantics", *Noûs*, 52(1), 109-143.

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

Koriat, A. (2012): "The self-consistency model of subjective confidence", *Psychological Review* 119(1): 80.

Krosnick, J.A., Boninger, D.S., Chuang, Y.C., Berent, M.K., and Carnot, C.G. (1993): "Attitude strength: One construct or many related constructs?", *Journal of Personality and Social Psychology* 65(6): 1132–1151.

Lander, E.S., Baylis, F., Zhang, F., Charpentier, E., Berg, P., Bourgain, B., Friedrich, B., Joung, J.K., Li, J., Liu, D., Naldini, L., Nie, J., Qiu, R., Schoene-Seifert, B., Shao, F., Terry, S., Wei, W., and Winnacker, E. (2019): "Adopt a moratorium on heritable genome editing", *Nature* 567: 165–168.

Lehmann, V., Huis, E.M., and Vingerhoets, A.J. (2013): "The human and animal baby schema effect: Correlates of individual differences", *Behavioural processes* 94: 99-108.

McCaughey, T., Sanfilippo, P.G., Gooden, G.E.C., Budden, D.M., Fan, L., Fenwick, E., Rees, G, MacGregor, C., Si, L., Chen, C., Liang, H.H., Baldwin, T., Pébay, A., and Hewitt, A.W. (2016): "A global social media survey of attitudes to human genome editing", *Cell stem cell* 18(5): 569-572.

MacFarlane, J. (2007): "Relativism and Disagreement", *Philosophical Studies* 132: 17–31.

McGarty, C., Turner, J.C., Oakes, P.J., Haslam, S. A. (1988): "The creation of uncertainty in the influence process: The roles of stimulus information and disagreement with similar others." *European Journal of Social Psychology* 23(1): 17-38.

Meehl, P. E. (1954): *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Mercier, H., and Sperber, D. (2011): "Why do humans reason? Arguments for an argumentative theory". *Behavioral and Brain* 34: 57-111.

Morren, M., Rijken, M., Baanders, A.N., and Bensinga, J. (2007): "Perceived genetic knowledge, attitudes towards genetic testing, and the relationship between these among patients with a chronic disease", *Patient education and counselling* 65(2): 197.

Myers, D. G., & Lamm, H. (1976): "The group polarization phenomenon", *Psychological bulletin* 83(4): 602.

Nickerson, R. (1998): "Confirmation bias: a ubiquitous phenomenon in many guises", *Review of general psychology* 2(2): 175-220.

Orive, R. (1988): "Social projection and social comparison of opinions", *Journal of Personality and Social Psychology* 54(6): 953.

Rollwage, M., Loosen, A., Hauser, T., Moran, R., Dolan R., and Fleming, S. (2020): "Confidence drives a neural confirmation bias", *Nature Communications* 11: 2634-

Schauble, L. (1990) "Belief revision in children: The role of prior knowledge and strategies for generating evidence", *Journal of experimental child psychology* 49.1: 31-57.

Scheufele, D.A., Xenos, M.A., Howell, E.L., Rose, K.M., Brossard, D., and Hardy, B.W. (2017): "US attitudes on human genome editing", *Science* 357(6351): 553–554.

Schwardmann, P., and Van der Weele, J. (2019): "Deception and self-deception", *Nature Human Behaviour*: 1-7.

Skitka, L.J., Washburn, A.N., and Carsel, T.S. (2015): "The psychological foundations and consequences of moral conviction", *Current Opinion in Psychology* 6: 41-44.

Skitka, L.J. (2010): "The psychology of moral conviction", *Social and Personality Psychology Compas*s 4(4): 267-81.

Skitka, L.J., Bauman, C.W., and Sargis, E.G. (2005): "Moral conviction: another contributor to attitude strength or something more?", *Journal of Personality and Social Psychology* 88: 895-917.

Stalnaker, R. (1999): *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stokke, A. (2016): Lying and Misleading in Discourse. *Philosophical Review* 125 (1): 83-134.

Tetlock, P. E. (2017): *Expert political judgment: how good is it? How can we know? - New edition*. Princeton University Press.

Tosi, J., and Warmke, B. (2016): Moral Grandstanding. *Philosophy and Public Affairs* 44 (3): 197-217.

Willer, M. (2013): Dynamics of Epistemic Modality. *Philosophical Review* 122 (1): 45-92.

Table 1. Regression models of confidence and belief change. In models 1 and 2, the dependent variable (DV) is confidence after moral discourse (at t2), in models 3 and 4 the DV is belief at t2. Models 1 and 3 concern the moral domain, models 2 and 4 the factual domain. p < '***' 0.001 '**' 0.01 '*' 0.05 '·' 0.1

| IVs	confidence at t1	belief at t1	extremity at t1	extremity at t2	belief difference	gender	age	R^2	N
DV: confidence at t2									
model 1	.062 (.037) .	-.917 (.434) *	1.991 (1.095) .	10.470 (0.929) ***	-1.258 (.771)	2.207 (1.457)	.020 (.471)	.30	507
model 2	.209 (.068) **	-.050 (1.575)	-2.583 (3.075)	20.959 (2.421) ***	2.306 (1.738)	9.858 (3.364) **	.068 (1.101)	.49	165
DV: belief at t2									
model 3	.003 (.003)	.769 (.033) **	-.236 (.083) **	.046 (.071)	-.001 (.059)	.221 (.111) *	-0.027 (.036)	.52	507
model 4	.002 (.005)	.559 (.113) ***	-.176 (.221)	.465 (.174) **	-.132 (.125)	-.200 (.241)	.095 (.079)	.25	165

Figure 1
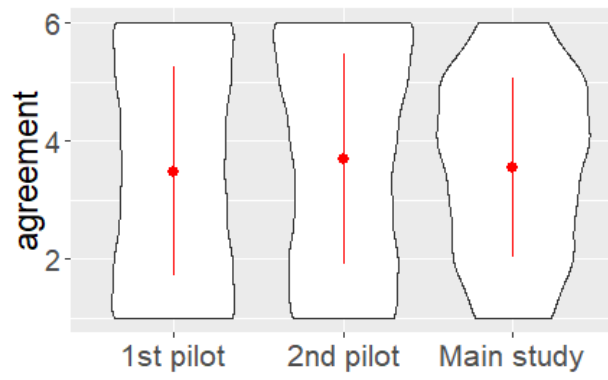
Figure 2

a) | It is ethically permissible to access data from genome databases for the purpose of investigating a crime.

Strongly disagree	Disagree	Rather disagree	Rather agree	Agree	Strongly agree
○	○	○	○	○	○

b) | Personalized cancer therapies always require sequencing the cancer cells' entire genome.

Strongly disagree	Disagree	Rather disagree	Rather agree	Agree	Strongly agree
○	○	○	○	○	○

c) | How confident are you regarding your answer to the last question?

0%   100%
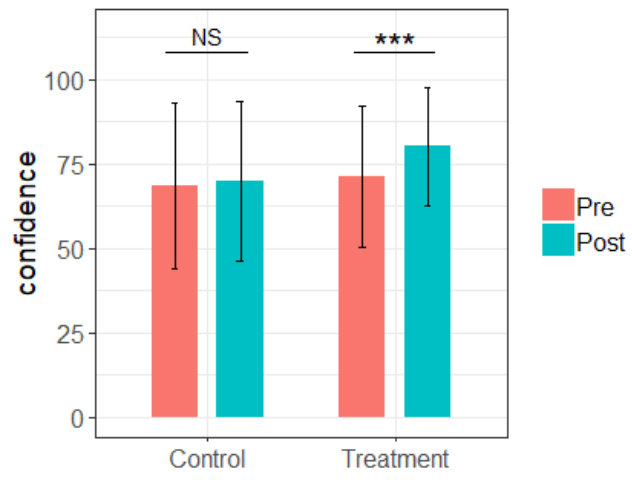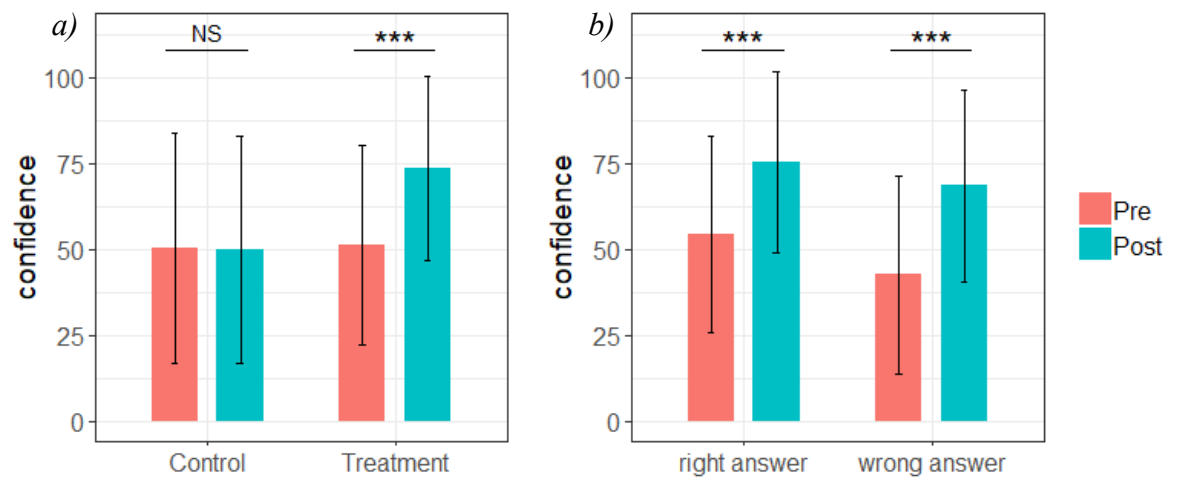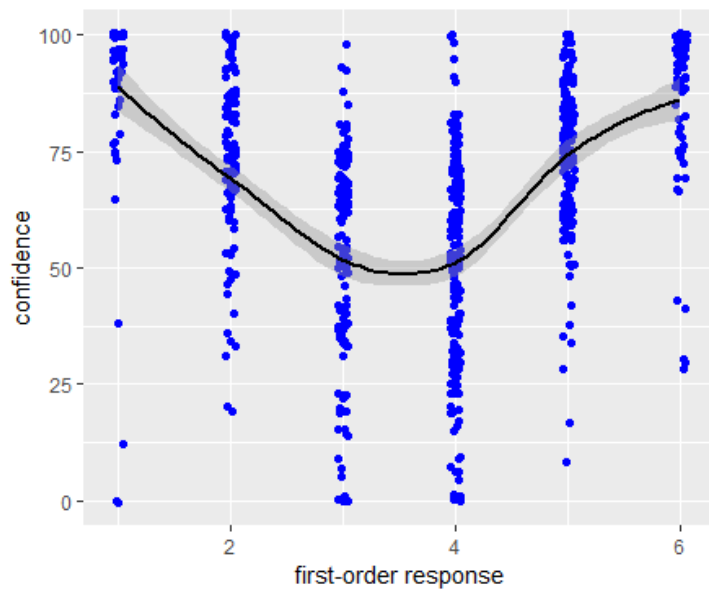
Figure 3

Figure 4

Figure 5

Figure captions

Figure 1. **Moral disagreement.** *First-order ratings of agreement or disagreement with ethical statements in our first and second pilot samples, and the main sample. Opinions are widely spread in all three cases, with SDs of 1.78, 1.78, and 1.51, respectively.*

Figure 2. **Examples for questionnaire items.** *Participants first indicated the extent to which they agreed or disagreed with either (a) an ethical claim (here concerning data protection) or (b) a factual claim (here concerning personalized medicine) on a discrete 6-point Likert scale. In both cases, they subsequently indicated (c) their degree of confidence about their first-order judgment on a continuous rating scale.*

Figure 3. Fig. 3: **Exposure to moral discourse boosts confidence of ethical judgments.** *There was no significant change in confidence on ethical judgments before versus after a workshop on an unrelated topic (control group). In contrast, confidence significantly (p < .001) increased for participants in the treatment group who attended a workshop where they engaged with the topic. Error bars indicate standard deviations.*

Figure 4. **Workshop on bioethics boosts confidence about factual belief irrespective of truth.**
a) Confidence about factual items was not affected by a workshop on an unrelated topic (control group). However, confidence significantly (p < .001) increased for participants of the treatment group who attended a workshop where they engaged with the topic.

b) This increase in the treatment group was significant irrespective of whether the given answer was correct or not (each p < .001). Error bars indicate standard deviation from the mean.

Figure 5. **Correlation between first-order responses and confidence.** *More extreme first-order views (stronger agreement or disagreement) were correlated with higher confidence. This was true for both groups, before as well as after the workshops, and in both the factual and the ethical domain. Figure shows all responses by the treatment group (r = .62, p < .001).*