Extremists are more confident

Nora Heinzelmann* and Viet Tran†

accepted by Erkenntnis

Abstract

Metacognitive mental states are mental states about mental states. For example, I may be uncertain whether my belief is correct. In social discourse, an interlocutor's metacognitive certainty may constitute evidence about the reliability of their testimony. For example, if a speaker is certain that their belief is correct, then we may take this as evidence in favour of their belief, or its content. This paper argues that, if metacognitive certainty is genuine evidence, then it is disproportionate evidence for extreme beliefs. In support of the argument, we report findings from five studies with different participant samples, designs, and measures. These studies show that, the more extreme an agent's belief (positive or negative), the more certain they are about it, and vice versa. This relationship might contribute to moralism, virtue signalling, and polarisation, which in turn may be epistemically and morally problematic. Therefore, we caution against taking metacognitive certainty as genuine evidence.

Keywords metacognition; moral psychology; experimental philosophy; epistemology; moral beliefs; social deliberation

1 Introduction

> *The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts.*
>
> (Bertrand Russell, quoted in Garg 2002, p. 207)

Russell states, among other things, that "fanatics" are more certain of themselves than "wiser people". This could be read as a claim about the extremity of viewpoints. According to a widespread view exemplified by Aristotle (*Nicomachean Ethics*, book II), virtue is a mean between two extremes. For example, a virtuous person is moderate in courage, avoiding the lower extreme (cowardice) as well as the higher one (rashness). We may thus interpret the quote as saying that extremists tend to be more

*Institute of Philosophy, Friedrich Alexander University of Erlangen-Nuremberg, Bismarckstrasse 1, Erlangen 91054, Germany, nora.heinzelmann@fau.de

†Faculty of Statistics, Ludwig Maximilian University, Munich, Germany

certain of themselves than moderates. Applied to a range of possible views about a given issue, extremists may be regarded as those who occupy either of the two extreme views on the spectrum. From a Russellian perspective, these extremists are more certain than moderates. This claim will be our main focus. We consider whether it is accurate and, if so, whether this is problematic.

More specifically, we provide new experimental evidence for Russell's claim. Across five studies using different measures, stimuli, and participant samples, we find a quadratic relationship between first-order mental states and metacognitive certainty. This indicates that we tend to be less certain about our moderate mental states, and more certain about our extreme ones. In Russell's words, fanatics are indeed certain of themselves, and wiser people full of doubt.

However, is this problematic? The bulk of this paper is devoted to developing a tentative response to this question: we caution that, whilst perhaps not "the whole problem with the world", the fact that extreme views correlate with greater metacognitive certainty may have worrisome implications. This is because metacognitive certainty may be taken as higher-order evidence.

It is a matter of debate in epistemological research whether and how an agent ought to adjust their first-order mental state in light of new higher-order evidence (e.g., Christensen 2007, 2010a; Dorst 2019c; Feldman 2005; Kelly 2010; Salow 2018; Schoenfield 2018). For example, testimony may provide such evidence pertaining to, say, the speaker's expertise, trustworthiness, or metacognitive certainty. Imagine a speaker tells us that they believe that p, and then adds: "I am nearly certain that my belief is correct". Are we justified, rationally permitted or even required to take the second claim as evidence when we form our own belief?

Some authors have argued that evidence from testimony is not and ought not be taken as genuine evidence. For example, Tosi and Warmke (2016) worry that contributions to discourse may negatively affect social deliberation if they are made with the intention to enhance the speaker's social status. Others have disagreed. For instance, Levy (2020) suggests that when a speaker conveys metacognitive certainty in their first-order belief that p, this is positive evidence in favour of p.

We do not aim to resolve this debate. However, we aim to contribute considerations that may advance it. For, if metacognitive certainty about a first-order belief should be taken as evidence, as some philosophers argue, and if metacognitive certainty is greater for extreme beliefs, as our research shows, then extreme beliefs may enjoy greater evidential support in social discourse.

The plan of our paper is as follows. In the next section, we provide arguments for why we may obtain evidence from testimony in general (at 2.1), and from metacognitive certainty more specifically (at 2.2). We draw the interim conclusion that agents take metacognitive certainty as evidence, even if this is not rational or justified. Then, we present own work supporting the claim that extremity of first-order belief correlates with metacognitive certainty: extremists tend to be more certain than moderates (section 3). We also draw on research from neuroscience that reports and explains this quadratic relationship. Finally, we argue that

if metacognitive certainty is taken as evidence, then it disproportionately favours extreme beliefs (section 4). In other words, because metacognitive certainty tends to be higher for extreme than for moderate beliefs, it may appear that there is stronger evidence for extreme beliefs. We argue that this mechanism may contribute to moralism, virtue signalling, and polarisation, which in turn can be deeply problematic. Therefore, it seems that we should rather not take metacognitive certainty as genuine evidence.

2 Higher-order evidence

2.1 Evidence from testimony

Epistemological research into higher-order evidence discusses whether and how an agent ought to adjust their mental states in light of new higher-order evidence (e.g., Christensen 2007, 2010a; Dorst 2019c; Feldman 2005; Kelly 2010; Salow 2018; Schoenfield 2018). Higher-order evidence is evidence about evidence, such as evidence about the reliability of a process that generates a mental state (Christensen 2010a; Dorst 2019b; Feldman 2005; Kelly n.d.; Lasonen-Aarnio 2014; Whiting 2020). This may be evidence that an epistemic peer disagrees with us (Christensen 2009, 2011; Lasonen-Aarnio 2013; Vavova 2014), evidence that one lacks countervailing testimony or arguments (Eva and Hartmann 2018; Goldberg 2011), evidence that one's education might have crucially determined one's current mental states (Christensen 2010b; Elga n.d.; Lasonen-Aarnio 2014; Schechter 2013; Schoenfield 2014), evidence that one's judgement may suffer from biases or a lack of oxygen (Elga n.d.; Horowitz 2014; Lasonen-Aarnio 2014), or debunking evidence in general (Kahane 2011; Vavova 2018, 2021).

In this paper, we focus on evidence from testimony, particularly from testimony given in social discourse. We receive evidence from speakers when they communicate their beliefs to us (Coady 1990; Graham 2000). Typically, it is not just an assertion itself that provides us with evidence. In addition, we tend to acquire background evidence in testimonial situations (Adler 1994; Faulkner 2000; Siegel 2005). This background evidence may concern, say, the truthfulness and trustworthiness of the interlocutor (Faulkner 2007; Hinchman 2005; Lewis 1973, 1975), or their expertise (Chudnoff 2021; Jones 2002; Kitcher 1993). Even having been given a platform, such as at public events or on social media, may constitute higher-order evidence because it conveys the impression that others second what is communicated on that platform (Levy 2019; cf. Estlund 2018; Simpson and Srinivasan 2018).

Some evidence from testimony may be conveyed by discursive features like intonation, gestures, or phrasing. Whether these features of social discourse constitute genuine evidence or, relatedly, whether we are rational or justified in taking them as evidence, is a matter of ongoing discussion. One side of this debate has suggested that these features do constitute evidence. For one thing, they may convey or signal information that is epistemically valuable to the speaker's audience.

For example, verbal tone, smiling, gestures, polite wording and similar cues may constitute social signals that instil trust between speaker and audience, and consequently promote social cooperation (Sarkissian 2010). If this is true, then these signals seem to constitute genuine evidence for the speaker's trustworthiness. Beyond these epistemic values, taking the communicative signals as evidence may also have pragmatic and moral value, as it may promote cultivation of virtues in all who participate in the discourse (Alfano 2014; Sarkissian 2010; Upton 2017).

In some cases, a piece of testimony may constitute blame, such as when A calls B a liar. Besides communicating a claim to the audience, blaming may signal the speaker's commitment to a set of norms (Shoemaker and Vargas 2019). For example, A does not only voice the claim that B is a liar but also that, other than B, A themselves are committed to honesty and truthfulness. On this view, the blamer provides evidence for their normative competence, demands, and intentions. They do not only convey these signals to the target of the blame but also to bystanders.

Relatedly, Levy (2020) argues that virtue signalling provides genuine evidence. "Virtue signalling" denotes seemingly uncandid expressions of morality that may serve to enhance one's social status. For example, saying that a joke was racist and disgusting might not only convey the speaker's moral belief about the joke but also evidence that the speaker believe themselves to be free from racism and a better person than those who invented the joke. Virtue signalling is not confined to discourse, and it need not be conscious or intentional. For example, buying and displaying a product with an organic or fair-trade label might be virtue signalling even if the buyer did not purchase the product with the intention of showing off environmental or social values. In this paper, we focus on oral or written discourse. In sum, on these accounts social discourse may provide us with evidence from testimony.

However, some authors have developed arguments to the conclusion that at least some features of social discourse do not constitute genuine evidence, and that we are not rational or justified in taking them as evidence. For one thing, evidence from testimony is arguably higher-order evidence, and it is controversial whether and how higher-order evidence can be rationally accommodated at all (Christensen 2010a; Dorst 2019a; Lasonen-Aarnio 2014, 2018). Taking higher-order evidence as genuine evidence may lead to epistemic akrasia, a case where an agent has a first-order belief but is higher-order uncertain that they are justified in this belief. In addition to these epistemic worries, authors have raised practical and moral concerns. For example, Tosi and Warmke (2016) argue that contributions to discourse intended to convey the speaker's high moral status negatively affect social deliberation and are therefore morally bad.

The present paper aims to contribute to the debate about higher-order evidence in social discourse. Although it remains neutral on whether or not discursive features of testimony constitute genuine evidence, it argues that *if* some features, namely evidence from metacognitive certainty, constitute genuine evidence, *then* this evidence disproportionally favours certain first-order beliefs. Before describing these beliefs in detail, we provide some arguments for why metacognitive certainty may be taken as genuine evidence.

2.2 Metacognitive certainty as evidence

Virtue signalling, blaming, and similar pieces of social discourse may provide evidence in that they communicate, inter alia, the speaker's metacognitive certainty (Levy 2020).

Metacognition is cognition about cognition (Carruthers 2009; Proust 2007, 2013). Metacognitive mental states are higher-order: they are mental states about mental states. For example, a higher-order belief is a belief about a (first-order) belief. Mental states can be metacognitive although they are of a different kind than the mental states they are about. For instance, an agent could be thinking about an emotion they are experiencing. In this case, the thought is metacognitive, as it is about the lower-level emotion. This thought is metacognitive although it is not about another thought but about an emotion.

In this paper, we focus on metacognitive certainty. This certainty may concern lower-level certainty and then be of the same kind of mental state. But it may also concern another lower-level mental state: a belief, a perceptual state, a desire, etc.

To distinguish metacognitive certainty from first-order certainty, consider the following example, adapted from Dorst (2019a) (cf. Feldman 2005). An agent is inclined to agree with the claim that Tbilisi is the capital of the republic of Georgia. But they are not fully certain that the claim is true. This is first-order certainty: certainty about whether the proposition "Tbilisi is the capital of Georgia" is true or not. In this case, the agent is not *fully* certain.

Let's assume that the agent assigns a credence of .7 to "Tbilisi is the capital of Georgia". It seems that the agent could be more or less certain about this assignment—imagine they ponder whether they should assign .6 or .8 instead. This would be metacognitive certainty: certainty about whether "my credence is .7" is true or not[1].

The behavioural sciences are particularly interested in a kind of metacognitive certainty called "confidence" (De Martino et al. 2013; Fleming, Weil, et al. 2010; Navajas, Niella, et al. 2018; Pouget, Drugowitsch, and Kepecs 2016). Confidence is defined as the subjective probability that a belief is correct. The more confident an agent is, the higher is the probability they assign, i. e., the more higher-order certain they are. Confidence is intensely researched in the behavioural sciences because it plays a key role in action and choice (Rahnev et al. 2020), such as value judgements (Folke et al. 2016), perception (Navajas, Niella, et al. 2017), confirmation bias (Rollwage et al. 2020), and social cooperation (Bahrami, Olsen, Latham, et al. 2010). It is commonly measured by self-reports of how certain an agent is that a prior choice or judgement was correct, or how confident they are about it (Fleming and Lau 2014). Typically, participants are asked to indicate how confident or certain they are about a prior choice or judgement.

"Confidence" may thus be synonymous to "attitude certainty" or "attitude correctness" (Petrocelli and Rucker 2007), on the following understanding of "attitude certainty": it denotes the extent to which someone views their own attitude as valid or correct (Clarkson, Tormala, and Rucker 2008; Dalege et al. 2016; Gross, Holtz, and Miller [1995] 2014;

Tormala, Clarkson, and Petty 2006). Like confidence, attitude certainty is a special kind of metacognition, and may play a key role in action and decision-making. For one thing, the greater an agent's attitude certainty, the more likely they are to act in accordance with the attitude in question (Fazio and Zanna 1978). A related but narrower concept is moral conviction, the metacognitive belief that a first-order belief is based on one's core moral values (Skitka 2010; Skitka, Washburn, and Carsel 2015). E. g., the greater an agent's moral conviction about a belief, the less likely they are to change it based on social influence (Skitka, Bauman, and Sargis 2005).

In this paper, we use the label "confidence" and follow the first usage mentioned above. We thus understand confidence as the metacognitive certainty that a lower-level belief is correct. In this, we align with some philosophers (e. g., Skipper 2020) but we depart from others (e. g., Elga 2007) who use "confidence" as a synonym of "credence" or "first-order certainty". We use "metacognitive certainty" and "confidence" interchangeably.

Current debates in epistemology concern the question of whether agents (rationally) ought to be fully higher-order certain about their first-order mental states (Dorst 2019a; Lasonen-Aarnio 2014, 2018; Salow 2018; Skipper 2020; Titelbaum 2015) or on the question of whether an agent can be rational when their first- and higher-order mental states vary independently from one another (Coates 2012; Greco 2014; Hazlett 2012; Lasonen-Aarnio 2015; Salow 2018; Smithies 2012; Wedgwood 2012; Williamson 2000). We need not enter these debates here.

Let us return to the suggestion that confidence, i. e., metacognitive certainty, may constitute evidence in social discourse. Here is an intuitively plausible way of developing it. Imagine that two agents, A and B, are engaged in a conversation. A claims that p. Now, that A claims that p may be taken by B as positive (or negative) evidence. Furthermore, there may be background evidence concerning, say, A's epistemic status as an expert or peer or their testimonial reliability more generally (Bovens and Hartmann 2003, ch. 5). Similarly, the confidence A conveys or reports about their belief that p may be taken by B as evidence in favour of (or against) p. For example, if A appears highly confident in their claim that p, B may take this as evidence for A's belief that p, or for p itself. Conversely, if A seems hardly confident about their belief, B might take this as evidence against or as a lack of evidence for the belief, or for p.

There are two reasons why B might be justified or rational in taking A's confidence about their claim that p as evidence. One reason is that doing so allows agents with shared interests to exchange relevant information efficiently and leads to optimal solutions in interactions. This has been proved mathematically (Thomas and McFadyen 1995), and confirmed experimentally (Bang, Aitchison, et al. 2017).

Another reason is that empirical research has shown that an agent is more confident in their belief that p when they have stronger evidence for p (Fleming and Daw 2017). For example, when A has strong evidence that suspect Jones is guilty then A tends to report greater confidence in their belief that Jones is guilty (Pulford et al. 2018). Relatedly, it has been found that the degree of confidence in one's perceptual belief correlates with its actual accuracy (Bahrami, Olsen, Latham, et al. 2010).

Furthermore, even if confidence *should not* be taken as evidence in this way, empirical research suggests that we *do* take it as evidence: we follow a confidence heuristic when assessing testimony (Anderson, Brion, et al. 2012; Anderson and Kilduff 2009; Bahrami, Olsen, Latham, et al. 2010; Kappes et al. 2020; Moore et al. 2017; Pulford et al. 2018). For instance, study participants prefer to take advice from more confident financial advisors and believe that their judgements are correct (Price and Stone 2004)[2]. Moreover, when participants report their own performance on an intelligence test, they are evaluated more positively when they appear confident (Schwardmann and van der Weele 2019).

Taken together, empirical and conceptual considerations suggest that confidence in a belief that p is (taken as) evidence about p. More specifically, high confidence in a belief that p is (taken as) evidence in its favour. We have thus established the first claim on which the argument of this paper rests: agents take metacognitive certainty as evidence, even if this is not rational or justified. In the next section, we present own findings that support the second claim.

3 Studies

Empirically, we find that confidence about a first-order belief correlates with the extremity of that belief. In this section, we report five studies supporting this claim, conducted between 2013 and 2019. Unless otherwise stated, they were approved by the Ethics Committee of the University of Konstanz (approval no. 33/2018).

3.1 Pilot study

This study was a pre-test of a bioethical questionnaire about genetic technologies. We measured moral judgments in 79 Amazon Mechanical Turk (MTurk) workers (27 females, mean age 35.2, SD 9.9 years) after excluding participants who failed attention checks. The study was approved by the University of London School of Advanced Study Ethics Committee and conducted in 2017. Participants were compensated with $ 3 for completing the study. Materials and data are available from the Open Science Framework[3].

We administered a questionnaire comprised of 32 items in random order. Half of the items were statements and half were vignettes. Each vignette was matched to a statement. 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.*"

After reading each item, participants indicated on a 6-point Likert scale ranging from "strongly agree" to "strongly disagree" (or vice versa) to what extent they agreed or disagreed with the italicised moral claim.

For each item, they were also asked "How confident are you that your decision is correct?" and answered on a 6-point Likert scale ranging from "very confident" to "not at all confident" (or vice versa). In addition to gender and age, we also collected information about participants' religiosity, education, political affiliation, and prior exposure to genetic testing or cancer (whose treatment often involves genetic testing).

We found that extremity of agreement ratings (positive or negative) and confidence ratings were strongly correlated (Kendall's τ =.63, p< .001)[4]. In other words, the more strongly a participant agreed or disagreed with a moral statement, the more confident they were about it. This finding was neither expected nor predicted.

3.2 Youth workshops

Our second experiment was a field study conducted as partial assessment of a training programme by Germany's national Cancer Research Center (DKFZ). We measured moral judgments and confidence in two groups of teenagers attending outreach events in Heidelberg in 2018. Materials and data for this study are available from the Open Science Framework[5].

Participants signed up for and attended either of two workshops, thereby assigning themselves to the treatment or the control group. We also obtained written consent from their parents. Workshops took place during weekends (Friday through Sunday). After exclusion we analysed 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 significant differences between the two groups with regard to age (t(31)=-1.91, p=.07), gender (χ^2(1)=.02, p=.90), or socioeconomic status (t(64)=-.11, p=.91).

The two-day workshops covered either normative issues surrounding genetic technologies (treatment group) or project management (control group). Twice, we administered the same online questionnaire: first on the Friday before and then during the week after the workshop. Each participant assigned themselves a pseudonym so we could match the pre- and post-workshop responses. The questionnaire consisted of 35 ethical and 12 factual statements about genetic technologies in randomised order, followed by questions about age, gender and parents' education, as an indicator of socioeconomic status.

The ethical statements were largely similar to the ones presented to the MTurkers in the pilot study but translated into German. For all statements, participants were asked to separately 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%. The orientation of agreement and confidence scales was randomised between participants in order to control for any systematic biases.

We found that extremity of first-order belief, positive as well as negative, was positively correlated with level of confidence. 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. This effect was robust within each category of beliefs,

ethical (r_s =.57, treatment: r_s =.60, control: r_s =.55; all p<.001) and factual ones (r_s =.70, treatment: r_s =.73, control: r_s =.65; all p<.001). It was also significant before (r_s =.58, ethical: r_s =.58, factual: r_s =.58; all p<.001) as well as after the outreach event (r_s =.62, ethical: r_s =.58, factual: r_s =.78; all p<.001).

Incidentally, we also observed a significant confidence boost for judgements in the experimental group, an effect we further discuss and explore in Heinzelmann, Höltgen, and Tran (2021).

3.3 Consumer conference

Like the second, our third experiment was a field study. We measured moral judgments and confidence in 20 participants (10 females, mean age: 44.1 years, SD=11.0) before and after they attended a consumer conference on genome editing organised by the German Federal Institute for Risk Assessment (BfR). The conference solicited informed lay opinions for policymaking and public debate during three workshops that took place in Berlin in 2019. Materials and data are available on the Open Science Framework[6].

Participants were selected by BfR from a pool of 147 applicants as a representative and unbiased sample of the German population and compensated with €500. Twice, we administered the same pen-and-paper questionnaire comprised of six claims about the ethical permissibility of genome editing in various contexts. These questions were taken and adjusted from the questionnaire administered during the first field study.

For each item, participants indicated on a 6-point Likert scale to what extent they agreed or disagreed with a moral claim (from "completely agree" to "not agree at all") and indicated how certain they were about their response (between 0% and 100% certain). Participants created and assigned an individual code for themselves that allowed us to match questionnaires completed before and after the conference.

We analysed data from 16 participants who provided at least one valid confidence estimate. Again, we found that extremity of agreement or disagreement correlated with confidence (r_s =.75, p<.001). In other words, the more strongly a participant agreed or disagreed with an ethical statement, the more confident they were about it. As our correlation analysis indicates, this correlation was strong and highly significant.

3.4 Neuroimaging study

For the present purpose, we re-analysed data from a published functional magnetic resonance imaging study (Heinzelmann, Weber, and Tobler 2020). Materials and data for this study are available on the Open Science Framework[7].

30 participants (19 females, mean age: 23 years) participated in the study, a figure chosen a priori on the basis of previous research with sample sizes ranging from 10 to 28 (Avram et al. 2013; Kawabata and Zeki 2004; Tsukiura and Cabeza 2011; Wang et al. 2015). Data from three participants (2 males, 1 female) were excluded from the analysis because of excessive movements or failure to complete the task. The study was

approved by the ethics committee of the Canton of Zurich and conducted in 2013.

In the scanner, participants consecutively viewed 24 artistic images depicting morally salient actions. In this experiment, we measured only first-order belief. Using a continuous slider, participants separately rated

- the moral goodness or badness of the depicted action (from "very bad" to "very good"),
- the beauty or ugliness of the image (from "very ugly" to "very beautiful"),
- the speed of the depicted action (from "very slow" to "very fast"), and
- the age of the image (from "very new" to "very old").

Each participant rated each image thrice for each criterion, i. e., 12 times overall. Orientation of the rating scale was randomised between trials to correct for any systematic bias.

We also measured reaction time from the onset of the image to participants' response. Reaction time is known to correlate negatively with confidence ratings (Bang, Fusaroli, et al. 2014; Lebreton et al. 2015; Patel, Fleming, and Kilner 2012; Pleskac and Busemeyer 2010; Pulford et al. 2018). That is, the more confident we are about a judgment, the faster we give it, and vice versa.

For all four kinds of judgments, we found that reaction time was smaller when judgments were more extreme. We found that extremity of ratings (i. e., squared ratings) correlated significantly with reaction time for each domain (goodness: r_s =-.28, beauty: r_s =-.18, speed: r_s =-.23, age: r_s =-.17; all $p<.0001$). In other words, if a participant found an image very beautiful, very ugly, very new or very old, or an action morally very bad, morally very good, very fast, or very slow, then they were more likely to report this assessment fast.

This analysis further corroborates the claim that extremity of judgment correlates positively with confidence.

3.5 Computer lab study

After having found in four studies that extremity of first-order belief correlated with confidence, we set up a fifth study to directly test this effect in a controlled experiment.

To that end, we examined chat interactions in a computer lab. We measured bioethical judgments and confidence in a student population (N=48, 21 males, 23 females, 2 diverse, mean age: 23.2 years, SD=2.8). This sample size was based an a priori power analysis (two-tailed Wilcoxon signed rank test, power=.9, alpha=.05, medium effect size=.5) that we conducted using G*Power (Faul et al. 2009), and which resulted in a total sample of 47. We recruited 48 students because we needed an even number of participants for dyadic interactions. Before data collection we pre-registered the study on the Open Science Framework[8].

The experiment consisted of three stages. First, participants assessed 18 claims about genetic technologies in random order on a desktop computer. For each claim, we measured participants' first-order belief and

their metacognitive certainty. More specifically, participants provided a moral judgment measured on a 6-point Likert scale from "morally good" to "morally bad". Then they indicated their confidence about that first-order judgment on a continuous scale from 0% to 100% (stage 1).

Afterwards, each participant consecutively discussed nine of the 18 claims through an anonymous chat interface with one other randomly assigned participant (stage 2). The nine claims were selected randomly as well. Participants were shown the claim together with their own and their interlocutor's assessment. Participants were instructed to explain their assessment to their partner. After an item had been discussed, couples were newly matched and the next item was put up for discussion.

The last stage (stage 3) was identical to the first one. At the end of the experiment, we asked participants three personal questions about their age, gender, and socioeconomic status, respectively.

Analysing data from all 48 participants, we found a positive correlation between first-order extremity and confidence (stage 1: $r_s = .59$, $p < .001$, stage 3: $r_s = .55$, $p < .001$). In other words, the more extreme a participant's moral belief was (positive or negative), the more confident they were about this belief. This relationship held before and after the chat interaction. That is, regardless of whether participants had or had not chatted with others about their views, they were more confident about extreme than about moderate ones.

Interestingly, we also found a significant confidence boost for the discussed items (independently of extremity), thus replicating a finding from the workshop field study, above. The Wilcoxon signed rank test for repeated measures ($V = 17743$, $p < .001$) revealed an increase of confidence after the discussion ($M = 81.9$, $SD = 20.3$) versus before ($M = 75.5$, $SD = 24.7$). There was no effect for items not discussed ($V = 25312$, p-value$= .13$) in stage 1 ($M = 76.0$ $SD = 23.0$) versus stage 2 ($M = 77.7$ $SD = 22.2$). The difference in confidence boost was significantly higher (Mann-Whitney U test, $W = 101104$, $p < .01$) for the discussed items ($M = 64.9$, $SD = 25.3$) versus not discussed items ($M = 15.8$, $SD = 22.8$). The first-order beliefs converged for the discussed items (Wilcoxon signed rank test on within group variance, $V = 5412$, $p = .03$).

We also calculated a linear mixed model for confidence with item and subject as random effects. For this model, we used data from only 44 participants because two participants did not specify their gender, and two more participants self-identified as neither male nor female, a subcategory too small for meaningful analysis. Overall, we analysed 1575 observations (occasionally, participants missed an item).

The results are presented in table 1. We find that extremity, age, and an interaction of stage and discussed items have effects on confidence.

Extremity has a positive effect on confidence. That is, the more extreme a first-order judgement was (in either direction), the more confident the agent was about it. Extremity ranged from 1 (that is, a rating of 3 or 4 on the 6-point Likert scale) to 3 (i. e., a rating of 1 or 6 on the Likert scale). More specifically, if a participant's rating was more extreme by one point on the Likert scale, then they were about 14% more confident about this rating.

	confidence			
Predictors	*Estimates*	*CI*		*p*
(Intercept)	12.84	–12.51	– 38.18	0.321
extremity	14.43	13.28	– 15.58	**<.001**
stage [3]	0.74	–1.66	– 3.14	0.547
discussed [yes]	–0.79	–3.19	– 1.60	.516
age	1.43	0.33	– 2.54	**.011**
gender [female]	3.29	–2.03	– 8.61	.226
SES [at least BA degree]	–3.64	–10.45	– 3.18	.296
stage [3] * discussed [yes]	4.09	0.70	– 7.48	**.018**

Table 1: *Effects of age, extremity, and discussion on confidence.* Table displays results from a linear mixed model with variables for extremity, stage (after discussion), discussed items, age, gender (female rather than male), socioeconomic status (SES), and interaction of stage and discussed items. Significant findings are printed in bold. p-values were calculated by t-tests using Satterthwaite's method conducted with the lmerTest R package. Marginal R^2 =.302, conditional R^2 =.449. Random effects: σ^2 =294.39, $\tau_{00subjects}$ =68.41, τ_{00item} =9.85, ICC=.21.

Age also had a positive effect on confidence. This means that the older a participant was, the more confident they were in their judgement. On average, if a participant was one year older, they were also more confident by about 1.4% percentage points. However, as the age of our participants ranged between 19 and 29 years and was thus comparatively narrowly spread, we suggest that not too much weight should be given to this finding. Therefore, we do not discuss it further but note it as a possible avenue for future research. Across a wider age range, future work could, e. g., test the hypothesis that older people are more confident than younger ones.

Discussion also seems to increase confidence. We calculated the effect of an interaction of stage and discussed items. That is, we examined whether the fact that an item had been discussed in an anonymous chat interaction led to an increase in confidence. At stage 1, participants assessed all items before discussion. At stage 2, they sequentially discussed half of those items with others. At stage 3, they assessed all items again, the ones they had discussed and the ones they had not discussed.

As shown in table 1 (predictor "stage"), whether an item was assessed at stage 3 or at stage 1 did not have an effect on confidence by itself. Likewise, whether an item was discussed or not had no significant effect on confidence (predictor "discussed"). However, the interaction of the two factors did have a combined effect on confidence (interaction of stage and discussion). That is, when a claim was discussed and then assessed again, participants were more confident about their first-order rating. As the second rating took place after the discussion, this indicates that the chat interaction causally contributed to the boost in confidence.

Overall, our chat experiment replicated the finding of our previous studies that extremity and confidence are positively correlated.

3.6 General discussion

This section discusses the main finding of our studies, possible limitations, and mechanisms that may explain it. Across five studies and with different measures, stimuli, and participant samples, we find a quadratic relationship between first-order mental states and metacognitive certainty (confidence). That is, metacognitive certainty is greater for more extreme mental states (positive or negative) than for moderate ones. Figure 1 illustrates this finding: confidence ratings were higher for more extreme first-order ratings. Each panel depicts this for one of four studies separately.

To our best knowledge, this phenomenon has not been discussed in philosophical research. However, it has been reported and discussed in neuroscience (Barron, Garvert, and Behrens 2015; Lebreton et al. 2015). In experiments by Lebreton et al. (2015), participants' reaction times and confidence ratings varied quadratically for a range of different domains (age of paintings, pleasantness of visuals, desirability of objects and future events, probability of future events).

For example, in one study, participants were shown a verbal description of an event (e. g., "France wins the 2014 World Cup"). On the next screen, they were then asked "How much would you like this?", and indicated the desirability of the event on a scale from −10 to 10. Subsequently, the next screen said, e. g., "You gave a rating of 5. How confident are

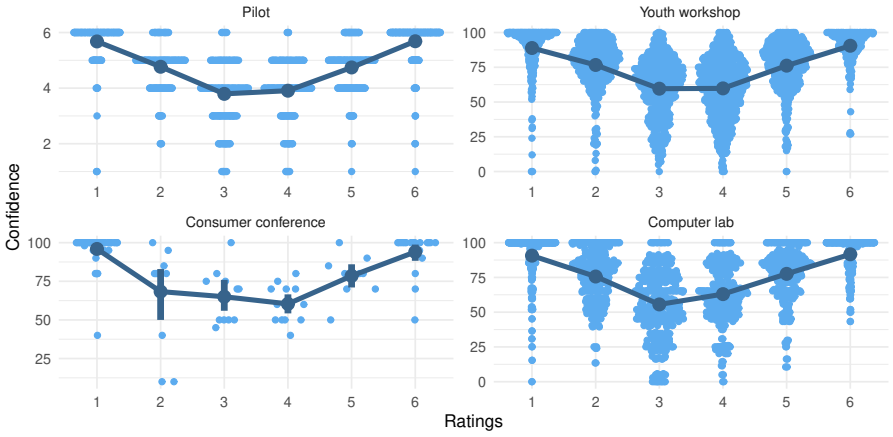

Figure 1: *First-order ratings and confidence.* Each panel shows data from one of four studies: the Mturk pilot, the two field studies (at youth workshops and a consumer conference), and a computer lab study. In all studies, we measured first-order ratings on a 6-point Likert scale (x axes). In all but the pilot study for which we used another Likert scale, we measured confidence as ratings between 0% and 100% (y axes). Raw data is plotted in the background. The foreground shows a line connecting the means with bootstrapped confidence intervals. In each case, confidence is higher for more extreme first-order ratings, positive and negative.

you?". Participants indicated their confidence on a continuous rating scale ranging from "not at all" to "totally".

Behaviourally, the experimenters identified a quadratic relationship between first-order ratings and confidence. Moreover, they found the same relationship neurally: the neural signal in the ventromedial prefrontal cortex was quadratically related to first-order ratings. The ventromedial prefrontal cortex has long been known as a neurocorrelate of value.

The effect we report in this paper is thus not only well established across our own studies but also supported by the empirical literature. However, our work also has at least two limitations.

First, all our studies suffered from a ceiling effect for confidence. That is, throughout all studies, participants tended to report high confidence measures, with even the lowest means above the midpoint of the rating scale. When adjusting the design from one study to the next, we tried to eliminate the ceiling effect by toning down the wording in our questionnaire on the one hand and by switching to other measures of confidence on the other hand (e. g., we switched from a Likert scale to continuous rating scales after the pilot study). However, these measures failed to eliminate the ceiling effect, as is plainly visible in figure 1. Nevertheless, the quadratic relationship between belief and confidence is sufficiently pronounced to be statistically significant in all cases. Although participants are thus in general highly confident about all their beliefs, they are even more confident about extreme beliefs.

Second, although our studies find the same effect in a variety of settings, they focus on moral beliefs. That is, in all studies we asked participants to report their first-order moral beliefs and to indicate how certain they were about them. However, some authors have argued that moral testimony is unlike testimony in other domains (henceforth: "non-moral testimony"). For one thing, relying on moral testimony seems more problematic than relying on non-moral testimony, or even downright impermissible (Andow 2020; Hills 2013; McGrath 2011; Williams 1995). If this is true, then we probably ought not rely on moral testimony at all. We should not take other people's moral testimony as genuine evidence.

However, even if we should not take moral testimony as evidence, it is still possible that we do. We may in this respect be irrational, epistemically irresponsible, or even immoral. In this case, the empirical work presented in this paper draws attention to a problem that arises when we rely on moral testimony: it shows that we take ourselves to have greater evidence for more extreme moral beliefs.

Moreover, two of our studies also examined beliefs in the non-moral domain. On the one hand, in our youth workshops study, we asked participants about a range of factual beliefs, and we observed that the extremity of first-order beliefs correlated significantly with metacognitive certainty about these beliefs. We thus found the same main effect in both the ethical and the factual domain.

On the other hand, our neuroimaging study examined non-moral domains as well. We measured participants' judgments about morality, beauty, age, and speed. Separately for all four judgements, we found that confidence was quadratically related to first-order belief.

Overall, our findings thus promise to generalise beyond the moral domain. However, in the case they do not, they still draw attention to an issue that raises concerns about moral testimony.

Despite these limitations, evidence from our studies and the literature indicates that the extremity of judgements and confidence about those judgements is related. The question thus arises of how the relationship is best explained. In the remainder of this section, we discuss two proposals.

The first has been suggested in the neuroscientific literature (Barron, Garvert, and Behrens 2015; Lebreton et al. 2015). When participants communicate their first-order mental state, they might be (higher-order) confident about this mental state, whatever it is, and report it accordingly. However, when they have no opinion, they might be inclined to choose a moderate first-order rating rather than an extreme one. In other words, when someone takes a moderate position, they might either truly have that moderate mental state and be confident about, or they might in fact suspend first-order judgement.

On a neural level, we may in effect represent low metacognitive certainty as a moderate first-order mental state (Barron, Garvert, and Behrens 2015). For, the brain (and, more precisely, the ventromedial prefrontal cortex) automatically encodes confidence when an agent makes first-order judgements. It might do so in two ways: it might either encode confidence and first-order mental states separately. Alternatively, confidence itself might be processed as a value signal. That is, the brain might treat confidence as intrinsically valuable, perhaps because it predicts the accuracy of first-order judgements.

Metacognitive certainty thus appears to be inherent in first-order mental states (Folke et al. 2016). This indicates that the brain does not, or even cannot, distinguish low higher-order confidence and absence or suspension of first-order judgement. The psychological or evolutionary function of this connection, and of metacognitive certainty more generally, is still an object of scientific debate.

Within this debate, one suggestion is that metacognitive certainty guides learning and exploration (cf. Boorman et al. 2009; Daw et al. 2006; Schulz et al. 2020): low certainty may prompt us to explore and to be more open to revising our first-order mental states. In other words, it may inform one's own future judgement and action. Within a predictive processing framework (Clark 2013; Friston 2010; Hohwy 2016), confidence might encode signal strength[9]. According to predictive processing theory, the brain tries to minimise the prediction error, a discrepancy between the predicted and the actual state of affairs. It also seeks to minimise the noise of the prediction error signal, i. e., aims to increase its precision. Confidence might encode the precision. A more precise prediction error, that is, higher confidence, influences updating of the brain's model to a greater degree.

Another, possibly complementary function of metacognitive certainty might be that it allows agents to communicate their first-order mental states more efficiently to others, and thereby improve group decisions (Bahrami, Olsen, Bang, et al. 2012; Bang, Fusaroli, et al. 2014). That is,

an individual's metacognitive certainty may inform collective first-order judgement and joint action.

Especially if it turns out that metacognitive certainty plays a crucial adaptive role in communication of an agent's mental states, the findings we and others have reported may raise an important caveat for empirical methodology and social discourse. For, it seems that when we elicit a report of an agent's mental state either in a behavioural experiment or in social discussion, we implicitly also elicit a report of their metacognitive certainty concerning this mental state. For instance, imagine that a talk show participant advocates a moderate view. Thereby, they may implicitly communicate that they are first-order uncertain about their first-order mental state or that they are confident that a moderate view is correct. The participant and their interlocutors may not distinguish between these two cases. As a consequence, contributions to social discourse may transport information that we may not have sufficiently appreciated.

A second possible explanation for why extremity of beliefs correlates positively with confidence is the following: more extreme beliefs have been formed on the basis of stronger or more evidence than moderate beliefs. As confidence tracks evidence, we are more confident of more extreme beliefs; they enjoy greater evidential support.

That more extreme beliefs are supported by more or greater evidence is a claim for which we have no empirical support. However, on the basis of the range of possible beliefs we can argue for it as follows. First, imagine an agent with a moderate belief (say, they believe that 0.5 in a case where belief content may range between 0 and 1) and contrast them with another agent with an extreme belief (say, they believe that 1). In social discourse, the first agent may encounter dissenting beliefs that differ by at most 0.5 from their own. However, the second agent may encounter beliefs that differ up to 1 from their own. This difference in belief range may be one dimension of disagreement (of course, there are others). Therefore, overall, disagreement may be greater or stronger for an agent with an extreme belief than for an agent with a moderate belief.

So, at least in one respect, extreme beliefs may face greater disagreement. Disagreement is, in turn, often (taken as) higher-order evidence in social discourse: encountering disagreement with one's belief is prima facie evidence against that belief. More specifically, *greater* disagreement may be (taken as) *stronger* evidence against that belief. That is, if the agent with a belief that 1 encounters an interlocutor who believes that 0.5, then there is weaker evidence against their belief that 1 than when they encounter an interlocutor who believes that 0. If this is correct, then one may expect weaker evidence from social discourse against a moderate belief than against an extreme belief.

Again, this need not be true in all cases, as disagreement and evidence have different dimensions. For one thing, it is of course conceivable that the agent with the moderate belief may encounter greater disagreement just because all their interlocutors happen to believe that 1. However, the converse is also conceivable. The argument we are developing here proceeds on the assumption that all other conditions are held constant.

We merely consider the difference between an extreme and a moderate belief, everything else being equal.

Now, it seems that in principle the agent with an extreme belief may be challenged by greater evidence against their belief than an agent with a moderate belief. Next, imagine that both agents succeed in rebutting the challenges and retain their beliefs. The agent with the extreme belief has defended their view against stronger disagreement than the agent with the moderate belief. It seems plausible, even justified, for the former to become more confident in their belief than the latter (we develop this argument more fully in Heinzelmann and Hartmann (2022); cf. Mill, *On Liberty*, Book II: 39–40). Whilst both agents may increase their metacognitive certainty about their beliefs, the agent with the extreme belief may increase their confidence even more than the agent with the moderate belief. Indeed, there is empirical evidence that agents are more confident in their belief when they have stronger evidence for it (Fleming and Daw 2017; cf. Clarkson, Tormala, and Rucker 2008; Tormala, Clarkson, and Petty 2006).

In sum, then, we may be more confident of extreme beliefs because we have previously defended them against greater disagreement and therefore we have better evidence for them.

Note that the two explanations just presented may not be mutually exclusive, and they may not be the only possible ones. Examining in greater detail the mechanism and rationality of the quadratic relationship between belief and metacognitive certainty may be avenues for future research. The relationship may also have implications of potential interest for normative and philosophical research. We shall turn to these in the next section.

4 Implications

From the claims established so far, the current section concludes that confidence may be disproportionately taken as metacognitive evidence for extreme views. Therefore, extreme beliefs may (appear to) enjoy greater evidence in social discourse. Then, we discuss possible implications pertaining to moralism, virtue signalling, and polarisation, respectively.

4.1 Evidence for extreme beliefs

We have established in section 2.2 that high confidence in a belief that p is taken as evidence in favour of the belief that p, or p itself. That is, as a matter of fact we take confidence as evidence. This may or may not be rational.

Moreover, our own research has shown that the relationship between confidence and first-order mental states is quadratic (section 3). That is, confidence in a belief is high when the belief is extreme. Again, this may or may not be rational.

These two claims, taken together, suggest that confidence may be disproportionately taken as positive evidence for extreme views. In other

words, if we take high confidence about a belief that p as positive evidence, then, because confidence tends to be higher for more extreme views, we may take ourselves to have greater evidence for more extreme views compared to moderate views.

Note that high confidence about a belief that p may also be taken as indirect evidence for a related belief or proposition, such as (the belief that) p'. For instance, consider an agent who is highly confident in their belief that genome editing is morally very bad. This may be taken by another agent as evidence for the claim that genome editing is somewhat bad. However, this evidence is indirect. It relies on the further assumption that evidence for p is also evidence for p', and thus presumably on some substantial claim about how p and p' are related.

First and foremost, then, confidence is (taken as) disproportionate evidence for extreme views. For example, imagine that two agents, A and B, have diverging views on the morality of genome editing. A has an extreme moral belief, they think that genome editing is morally very wrong. B has a moderate moral belief, they think that genome editing is morally neither good nor bad. Based on our research, we would expect that A expresses higher confidence in their moral belief than B. However, if we take high confidence in a belief as evidence in favour of that belief, then we seem to have good evidence in favour of A's extreme belief or its content. In contrast, B's lower confidence is presumably taken as weaker or no evidence in favour of their moderate belief or its content. In social discourse, A's belief or its content may therefore spread more easily than B's, and become more popular. Overall, this mechanism facilitates belief in more extreme views, such as A's belief that genome editing is morally very wrong.

4.2 Extreme views may be problematic

If the argument made in this paper is sound, then we take ourselves to have stronger evidence for more extreme views. As a result, individuals may be more likely to have extreme views, and they may prevail in a society. Is this, in itself, problematic? This question is our target in the current section.

Prima facie, it seems that it is not per se problematic that we take ourselves to have stronger evidence in favour of more extreme views. After all, having extreme views is at least in some cases entirely unproblematic. It seems epistemically and ethically justifiable that we should all believe, say, that murder is morally very bad or that it is nearly-certain that climate change is largely caused by humans. Accordingly, it seems unproblematic if these views prevail in individuals or in a society.

However, if an individual has extreme views, or if extreme views are popular in a society, then this may have further implications that are problematic, such as moralism, virtue signalling, and polarisation. Let us consider them in turn.

First, having extreme moral views may overlap with *moralism*. Moralism has variably been characterised as the illicit introduction of moral considerations (Driver 2005, p. 37), a failure to recognise the requirements

of moral thought or reflection (Taylor 2011, p. 153), and an inflated sense of the extent to which moral criticism is appropriate (Archer 2018).

Although moralism thus understood is not identical to moral extremity, the two may be closely related. For one thing, an agent with extreme moral beliefs might be inclined to moralise: "making extreme or excessive moral judgments [...] is a feature of moralism on many occasions" (Taylor 2011, p. 2). Relatedly, someone expressing an extreme moral belief may appear to others as moralising. Overall, then, if extreme views and moralism overlap, then, if extreme views spread more easily, at least some instances of moralism may spread more easily.

However, if this is true, then it may be problematic because moralism is typically regarded as deeply problematic. Arguably, moralism undermines moral criticism (Archer 2018) and moral self-improvement (Dean 2012), and causes uncharitable and thus harmful actions (Fullinwider 2005). Moralism may also be damaging to social discourse in that it threatens open-mindedness, an epistemic virtue supposedly conducive to truth (Kwong 2017; Song 2018; but see Fantl 2018; Levy 2006). If these concerns are warranted, then, they suggest at least some caution about extreme views in moral discourse.

Second, it seems that confidence is at least sometimes conveyed in virtue signalling, i. e., a contribution to social discourse aimed at communicating to others one's own high social status (Levy 2020; Tosi and Warmke 2020). As some philosophers have argued, virtue signalling may be problematic due to its effect on others. Therefore, at least in some instances it may be problematic to take a speaker's confidence as evidence.

Arguably, virtue signalling is problematic when it it is first and foremost intended to satisfy a desire to convince others that the speaker is morally respectable, i. e., when the speaker is grandstanding (Tosi and Warmke 2016, 2020). Because not all virtue signalling amounts to grandstanding—think of the feathers of a peacock signalling fitness—let us focus on those instances that do. It is likely that agents having more extreme views are more likely to grandstand (and vice versa). Indeed, grandstanding has been found to reliably relate to ideological extremism (Grubbs, Warmke, Tosi, and James 2020). That is, people who grandstand identify with more extreme political views on either the left or the right end of the spectrum. These findings converge with the ones we reported above: our data indicates that extreme beliefs are correlated with confidence.

Grandstanding may be problematic for a range of epistemic, practical, and moral reasons. For one thing, grandstanding seems insincere and hypocritical: the speaker makes a contribution to social discourse that would seem to stem from noble intentions, yet their true intention is entirely self-serving. Grandstanding may also be condescending, and thus disrespectful towards other participants in a social debate.

Furthermore, grandstanding may create an air of hypocrisy not merely about the speaker who grandstands but about moral talk in general (Tosi and Warmke 2020, p. 79). When contributions to social discourse are all-too-often guided by egoistic intentions, then we may become sceptical and cynical even about honest instances. Over time, this may erode trust and negatively affect social discourse and practice (Tosi and Warmke 2020,

pp. 67–92). Indeed, grandstanding has been associated with status-seeking personality traits and may contribute to conflict with others (Grubbs, Warmke, Tosi, James, and Campbell 2019).

Therefore, it seems plausible that at least in instances of grandstanding, taking confidence as evidence in social discourse may be epistemically and ethically problematic.

Third, if extreme views spread more easily in social discourse, this may contribute to *polarisation* in a society. Polarisation is the phenomenon that the average opinions of groups diverge between groups through discussion (Dorst 2019c; Myers 1975; Myers and Lamm 1976; Sunstein 2002). For example, imagine that one group is slightly left-leaning and another one is slightly right-leaning. After within-group discussion, the first group can be expected to be somewhat more left-leaning and the second one somewhat more right-leaning. In other words, between-group differences in opinions has increased; the two groups have polarised.

If people take high confidence as positive evidence and if, on average, we are more confident about extreme views, it can be expected that discussion will lead to a dominance of more extreme views over time. If the average opinion within each group thus becomes more extreme and if there is initially diversity of opinions between groups, then we may reasonably expect greater between-group differences over time. In other words, we may expect polarisation.

Indeed, there is evidence that polarisation has increased at least since the 1970s. In particular, Republicans and Democrats in the United States of America have become more divided in their political opinions (Abramowitz and Saunders 2008; Fiorina, Abrams, and Pope 2008; Levendusky 2009; McCarty, Poole, and Rosenthal [2006] 2016; Pew Research Center 2014). Although this may not have been true of the overall electorate until the mid-2000s (DiMaggio, Evans, and Bryson 1996; Evans 2003; Fiorina, Abrams, and Pope 2004), voters who identify strongly with the two major parties are recently becoming more and more polarised, and the share of voters who do strongly identify with a major party has been rising (Green, Palmquist, and Schickler 2002; McCarty, Poole, and Rosenthal [2006] 2016; Pew Research Center 2014). Moreover, there is evidence that affective polarisation is rising, i. e., the tendency of individuals identifying with one major party to evaluate counterpartisans negatively and copartisans positively (Iyengar, Lelkes, et al. 2019; Iyengar and Westwood 2015). Empirical evidence suggests that affective polarisation may be caused by extremity of opinions as well as the perceived extremity of others (Levendusky and Malhotra 2016; Rogowski and Sutherland 2016; Stroud 2010; Webster and Abramowitz 2017).

Polarisation is usually seen as problematic for epistemic, practical, and moral reasons. For one thing, it may increase the risk that participants in a discourse advocate false views, it may discredit the practice of social discourse and deter people from participating in it (Tosi and Warmke 2016, p. 212). It may also lead to conflict between groups and to moral wrongdoing towards out-group members (Greene [2013] 2014; Haidt 2013). Finally, polarisation may make it harder to reach society-wide consensus or compromise on issues of polarisation (Navajas, Heduan, et al. 2019). As

a result, social and political decisions may be more difficult to make and to enact. For example, it might make it harder for democratic parliaments to pass laws, or for elected representatives to form a government.

5 Conclusion

This paper has provided empirical evidence for Russell's presumed factual claim that extreme views are associated with greater confidence, and it has followed him in drawing attention to potentially problematic implications. More specifically, across five studies we find a quadratic relationship of first-order beliefs and metacognitive certainty about those beliefs: the more extreme the first-order beliefs are (on either side of the spectrum), the more confident agents are about them. It has been suggested that metacognitive certainty constitutes genuine evidence in social discourse. That is, when a speaker is more confident about a belief, then this may be taken as evidence in favour of that belief or the content of their belief. However, if this is true, then there will be disproportionate evidence in favour of extreme beliefs or their contents, as we tend to be more confident about them. This may have a range of implications, including problematic ones. In particular, individuals with extreme views and high confidence may be more prone to moralising or grandstanding, and a greater influence of extreme views in social discourse may contribute to polarisation. Overall, we therefore caution against taking confidence as genuine evidence in social discourse[10].

Notes

[1] As an anonymous reviewer points out, on imprecise probability models of first-order credence (Bradley 2019), this metacognitive certainty may be understood as an interval around 0.7. This raises the interesting question of whether imprecise credence models may be better suited to account for confidence than precise credence models. Here, we have to set this question aside.

[2] As an anonymous reviewer suggests, participants may prefer an advisor who expresses high confidence in a judgment if they know that the advisor expressed low credence in other claims (Skipper 2021). To date, there seems to be no empirical evidence to support Skipper's hypothesis but future work may test it directly.

[3] `https://osf.io/xntm9/`

[4] As we used a 6-point Likert scale to measure confidence, our data contained multiple ties and therefore we calculated Kendall's τ rather than Spearman's r. In our other studies, we used a more fine-grained measure of confidence.

[5] `https://osf.io/23ajd/`

[6] `https://osf.io/z93jk/`

[7] `https://osf.io/9q286/`

[8] Pre-registration: `https://osf.io/wxpba/`; data, analyses, etc: `https://osf.io/nrgyc/`

[9] We thank Neil Levy for suggesting this role of confidence within a predictive processing framework.

[10] We thank Alexander Dinges and Neil Levy for detailed comments on the manuscript. We are indebted to Alex Soutschek, Benedikt Höltgen, Giacomo Günther May, Johannes Doerflinger, Luis Hillebrand, Svenja Küchenhoff, Tomas Folke, and audiences in Berne, Cracow, and Munich. We also thank Katrin

Platzer, Frank Rösl and the German Cancer Research Center (DKFZ), Leonie Dendler, Emilia Böhm, and the German Federal Institute for Risk Assessment (BfR), the University of Munich (LMU), the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA), and the University of Zurich, notably Philippe Tobler. We thank two anonymous reviewers for their help to improve the paper. Our research was largely funded by the German Federal Ministry for Education and Research (BMBF).

References

Abramowitz, A. and K. Saunders (2008). "Is polarization a myth?" *The Journal of Politics* 70 (2):542–55.

Adler, J. (1994). "Testimony, trust, knowing". *Journal of Philosophy* 91 (5):264–75.

Alfano, M. (2014). *Character as moral fiction*. Cambridge: Cambridge University Press.

Anderson, C., S. Brion, et al. (2012). "A status-enhancement account of overconfidence." *Journal of Personality and Social Psychology* 103 (4):718.

Anderson, C. and G. Kilduff (2009). "Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance." *Journal of Personality and Social Psychology* 96 (2):491.

Andow, J. (2020). "Why don't we trust moral testimony?" *Mind & Language* 35 (4):456–44.

Archer, A. (2018). "The problem with moralism". *Ratio* 31 (3):342–50.

Aristotle ([n. d.] 1894). *Nicomachean ethics*. Ed. by I. Bywater. Oxford: Oxford University Press.

Avram, M. et al. (2013). "Neurofunctional correlates of esthetic and moral judgments". *Neuroscience Letters* 534:128–32.

Bahrami, B., K. Olsen, D. Bang, et al. (2012). "What failure in collective decision-making tells us about metacognition". *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1594):1350–65.

Bahrami, B., K. Olsen, P. Latham, et al. (2010). "Optimally interacting minds". *Science* 329 (5995):1081–5.

Bang, D., L. Aitchison, et al. (2017). "Confidence matching in group decision-making". *Nature Human Behaviour* 1 (6):s41562–017.

Bang, D., R. Fusaroli, et al. (2014). "Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making". *Consciousness and Cognition* 26:13–23.

Barron, H., M. Garvert, and T. Behrens (2015). "Reassessing VMPFC: full of confidence?" *Nature Neuroscience* 18 (8):1064–6.

Boorman, E. et al. (2009). "How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action". *Neuron* 62 (5):733–43.

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

Bradley, S. (2019). "Imprecise probabilities". In: *The Stanford encyclopedia of philosophy*. Ed. by E. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University.

Carruthers, P. (2009). "How we know our own minds: The relationship between mindreading and metacognition". *Behavioural and Brain Sciences* 32:121–82.

Christensen, D. (2007). "Epistemology of disagreement: the good news". *The Philosophical Review* 116 (2):187–217.

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

— (2010a). "Higher-order evidence". *Philosophy and Phenomenological Research* 81 (1):185–215.

— (2010b). "Rational reflection". *Philosophical Perspectives* 24 (1):121–40.

— (2011). "Disagreement, question-begging and epistemic self-criticism". *Philosophers' Imprint* 11.

Chudnoff, E. (2021). "Two kinds of cognitive expertise". *Noûs* 55 (2):270–92.

Clark, A. (2013). "Expecting the world: perception, prediction, and the origins of human knowledge". *Journal of Philosophy* 110 (9):469–46.

Clarkson, J., Z. Tormala, and D. Rucker (2008). "A new look at the consequences of attitude certainty: the amplification hypothesis." *Journal of Personality and Social Psychology* 95 (4):810.

Coady, A. (1990). *Testimony: a philosophical study*. New York: Oxford University Press.

Coates, A. (2012). "Rational epistemic akrasia". *American Philosophical Quarterly* 49 (2):113–24.

Dalege, J. et al. (2016). "Toward a formalized account of attitudes: the Causal Attitude Network (CAN) model." *Psychological Review* 123 (1):2.

Daw, N. et al. (2006). "Cortical substrates for exploratory decisions in humans". *Nature* 441 (7095):876–9.

De Martino, B. et al. (2013). "Confidence in value-based choice". *Nature Neuroscience* 16:105–10.

Dean, R. (2012). "A plausible Kantian argument against moralism". *Social Theory and Practice* 38 (4):577–97.

DiMaggio, P., J. Evans, and B. Bryson (1996). "Have American's social attitudes become more polarized?" *American Journal of Sociology* 102 (3):690–755.

Dorst, K. (2019a). "Evidence: a guide for the uncertain". *Philosophy and Phenomenological Research*.
— (2019b). "Higher-order uncertainty". In: *Higher-Order Evidence: New Essays*. Ed. by S. Rasmussen and A. Steglich-Peterson. Oxford: Oxford University Press.
— (2019c). "Why rational people polarize". In: *The Phenomenal World*.
Driver, J. (2005). "Moralism". *Journal of Applied Philosophy* 22 (2):137–51.
Elga, A. (2007). "Reflection and disagreement". *Noûs* 41 (3):478–502.
— (n.d.). "Lucky to be rational".
Estlund, D. (2018). "When protest and free speech collide". In: *Academic freedom*. Ed. by J. Lackey. New York: Oxford University Press, pp. 151–69.
Eva, B. and S. Hartmann (2018). "When no reason for is a reason against". *Analysis* 78 (3):426–31.
Evans, J. (2003). "Have Americans' attitudes become more polarized? An update". *Social Science Quarterly* 84 (1):71–90.
Fantl, J. (2018). *The limitations of the open mind*. Oxford: Oxford University Press.
Faul, F. et al. (2009). "Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses". *Behavior Research Methods* 41 (4):1149–60.
Faulkner, P. (2000). "The social character of testimonial knowledge". *Journal of Philosophy* 97 (11):581–601.
— (2007). "On telling and trusting". *Mind* 116 (464):875–902.
Fazio, R. and M. Zanna (1978). "Attitudinal qualities relating to the strength of the attitude-behavior relationship". *Journal of Experimental Social Psychology* 14 (4):398–408.
Feldman, R. (2005). "Respecting the evidence". *Philosophical Perspectives* 19 (1):95–119.
Fiorina, M., S. Abrams, and J. Pope (2004). *Culture war? The myth of a polarized America*. New York: Longman.
— (2008). "Polarization in the American public: misconceptions and misreadings". *Journal of Politics* 70 (2):556–60.
Fleming, S. and N. Daw (2017). "Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation". *Psychological Review* 124 (1):91.
Fleming, S. and H. Lau (2014). "How to measure metacognition". *Frontiers in Human Neuroscience* 8 (443):1–9.
Fleming, S., R. Weil, et al. (2010). "Relating introspective accuracy to individual differences in brain structure". *Science* 329:1541–3.
Folke, T. et al. (2016). "Explicit representation of confidence informs future value-based decisions". *Nature Human Behaviour* 1.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11 (2):127–18.
Fullinwider, R. (2005). "On moralism". *Journal of Applied Philosophy* 22 (2):105–20.
Garg, A. (2002). *A word a day*. Hoboken: Wiley.
Goldberg, S. (2011). "If that were true I would have heard about it by now". In: *Social epistemology: essential readings*. Ed. by A. Goldman and D. Whitcomb. New York: Oxford University Press, pp. 92–108.
Graham, P. (2000). "The reliability of testimony". *Philosophy and Phenomenological Research* 61 (3):695–709.
Greco, D. (2014). "A puzzle about epistemic akrasia". *Philosophical Studies* 167 (2):201–19.
Green, D., B. Palmquist, and E. Schickler (2002). *Partisan hearts and minds: political parties and the social identities of voters*. New Haven: Yale University Press.
Greene, J. ([2013] 2014). *Moral tribes*. London: Atlantic Books.
Gross, S., R. Holtz, and N. Miller ([1995] 2014). "Attitude certainty". In: *Attitude strength: Antecedents and consequences*. Ed. by R. Petty and J. Krosnick. Vol. 4. New York: Psychology Press, pp. 215–45.
Grubbs, J., B. Warmke, J. Tosi, and S. James (2020). "Moral grandstanding and political polarization: a multi-study consideration". *Journal of Research in Personality* 88.
Grubbs, J., B. Warmke, J. Tosi, S. James, and K. Campbell (2019). "Moral grandstanding in public discourse: status-seeking motives as a potential explanatory mechanism in predicting conflict". *PLoS ONE* 14 (10).
Haidt, J. (2013). *The righteous mind*. London: Penguin.
Hazlett, A. (2012). "Higher-order epistemic attitudes and intellectual humility". *Episteme* 9 (3):205–23.
Heinzelmann, N., B. Höltgen, and V. Tran (2021). "Moral discourse boosts confidence in moral judgments". *Philosophical Psychology* 34 (8):1192–216.
Heinzelmann, N. and S. Hartmann (2022). "Deliberation and confidence change". *Synthese* 200:1–13.
Heinzelmann, N., S. Weber, and P. Tobler (2020). "Aesthetics and morality judgments share cortical neuroarchitecture". *Cortex*.
Hills, A. (2013). "Moral testimony". *Philosophy Compass* 8 (6):552–9.
Hinchman, E. (2005). "Telling as inviting to trust". *Philosophy and Phenomenological Research* 70 (3):562–87.
Hohwy, J. (2016). "The self-evidencing brain". *Noûs* 50 (2):259–85.
Horowitz, S. (2014). "Epistemic akrasia". *Noûs* 48 (4):718–44.

Iyengar, S., Y. Lelkes, et al. (2019). "The origins and consequences of affective polarization in the United States". *Annual Review of Political Science* 22:129–46.

Iyengar, S. and S. Westwood (2015). "Fear and loathing across party lines: new evidence on group polarization". *American Journal of Political Science* 59 (3):690–707.

Jones, W. (2002). "Dissident versus loyalist: which scientists should we trust?" *Journal of Value Inquiry* 36 (4):511–20.

Kahane, G. (2011). "Evolutionary debunking arguments". *Noûs* 45 (1):103–25.

Kappes, A. et al. (2020). "Confirmation bias in the utilization of others' opinion strength". *Nature Neuroscience* 23 (1):130–7.

Kawabata, H. and S. Zeki (2004). "Neural correlates of beauty". *Journal of Neurophysiology* 91 (4):1699–1705.

Kelly, T. (2010). "Peer disagreement and higher order evidence". In: *Social Epistemology: Essential Readings*. Ed. by A. Goldman and D. Whitcomb. Oxford: Oxford University Press, pp. 183–217.

— (n.d.). "Evidence". In:

Kitcher, P. (1993). *The advancement of science: science without legend, objectivity without illusions*. New York: Oxford University Press.

Kwong, J. (2017). "Is open-mindedness conducive to truth?" *Synthese* 194 (5).

Lasonen-Aarnio, M. (2013). "Disagreement and evidential attenuation". *Noûs* 47 (4):767–94.

— (2014). "Higher-order evidence and the limits of defeat". *Philosophy and Phenomenological Research* 88 (2):314–45.

— (2015). "New rational reflection and internalism about rationality". In: *Oxford Studies in Epistemology*. Ed. by T. Gendler and J. Hawthorne. Vol. 5. New York: Oxford University Press.

— (2018). "Enkrasia or evidentialism? Learning to love mismatch". *Philosophical Studies*:1–36.

Lebreton, M. et al. (2015). "Automatic integration of confidence in the brain valuation signal". *Nature Neuroscience* 18 (8):1159–67.

Levendusky, M. (2009). *The partisan sort: how liberals became Democrats and conservatives became Republicans*. Chicago: University of Chicago Press.

Levendusky, M. and N. Malhotra (2016). "(Mis)perceptions of partisan polarization in the American public". *Public Opinion Quarterly* 80 (S1):378–91.

Levy, N. (2006). "Open-mindedness and the duty to gather evidence". *Public Affairs Quarterly* 20 (1):55–66.

— (2019). "No-platforming and higher-order evidence, or anti-anti-no-platforming". *Journal of the American Philosophical Association* 5 (4):487–502.

— (2020). "Virtue signalling is virtuous". *Synthese*:1–18.

Lewis, D. (1973). "Convention: a philosophical study". *Synthese* 26 (1):153–7.
— (1975). "Languages and Language". In: *Minnesota Studies in the Philosophy of Science*. Ed. by K. Gunderson. University of Minnesota Press, pp. 3–35.
McCarty, N., K. Poole, and H. Rosenthal ([2006] 2016). *Polarized America: The dance of ideology and unequal riches*. 2nd ed. Cambridge (MA): MIT Press.
McGrath, S. (2011). "Skepticism about moral expertise as a puzzle for moral realism". *Journal of Philosophy* 108 (3):111–37.
Mill, J. ([1859] 2014). *Collected works of John Stuart Mill*. Ed. by J. Robson. London: Routledge.
Moore, D. et al. (2017). "Confidence calibration in a multiyear geopolitical forecasting competition". *Management Science* 63 (11):3552–65.
Myers, D. (1975). "Discussion-induced attitude polarization". *Human Relations* 28 (8):699–714.
Myers, D. and H. Lamm (1976). "The group polarization phenomenon". *Psychological Bulletin* 83 (4):602.
Navajas, J., F. Á. Heduan, et al. (2019). "Reaching consensus in polarized moral debates". *Current Biology* 29 (23):4124–9.
Navajas, J., T. Niella, et al. (2017). "Deliberation increases the wisdom of crowds". *arXiv preprint arXiv:1703.00045*.
— (2018). "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds". *Nature Human Behaviour* 2 (2):126–32.
Patel, D., S. Fleming, and J. Kilner (2012). "Inferring subjective states through the observation of actions". *Proceedings of the Royal Society B: Biological Sciences* 279 (1748):4853–60.
Petrocelli John Vand Tormala, Z. and D. Rucker (2007). "Unpacking attitude certainty: attitude clarity and attitude correctness." *Journal of Personality and Social Psychology* 92 (1):30.
Pew Research Center (2014). *Political polarization in the American public*. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/. accessed 10 Dec 2020.
Pleskac, T. and J. Busemeyer (2010). "Two-stage dynamic signal detection: A theory of confidence, choice, and response time". *Psychological Review* 117 (3):864–901.
Pouget, A., J. Drugowitsch, and A. Kepecs (2016). "Confidence and certainty: distinct probabilistic quantities for different goals". *Nature Neuroscience* 19 (3):366.
Price, P. and E. Stone (2004). "Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic". *Journal of Behavioral Decision Making* 17 (1):39–57.

Proust, J. (2007). "Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition?" *Synthese* 159:271–95.
— (2013). *The philosophy of metacognition*. Oxford: Oxford University Press.
Pulford, B. et al. (2018). "The persuasive power of knowledge: Testing the confidence heuristic." *Journal of Experimental Psychology: General* 147 (10):1431.
Rahnev, D. et al. (2020). "The confidence database". *Nature Human Behaviour* 4 (3):317–25.
Rogowski, J. and J. Sutherland (2016). "How ideology fuels affective polarization". *Political Behavior* 38 (2):485–508.
Rollwage, M. et al. (2020). "Confidence drives a neural confirmation bias". *Nature Communications* 11 (1):1–11.
Salow, B. (2018). "The externalist's guide to fishing for compliments". *Mind* 127 (507):691–728.
Sarkissian, H. (2010). "Minor tweaks, major payoffs: the problems and promise of situationism in moral philosophy". *Philosophers' Imprint* 10.
Schechter, J. (2013). "Rational self-doubt and the failure of closure". *Philosophical Studies* 163 (2):428–52.
Schoenfield, M. (2014). "Permission to believe: why permissivism is true and what it tells us about irrelevant influences on belief". *Noûs* 48 (2):193–218.
— (2018). "An accuracy based approach to higher order evidence". *Philosophy and Phenomenological Research* 96 (3):690–715.
Schulz, L. et al. (2020). "Dogmatism manifests in lowered information search under uncertainty". *Proceedings of the National Academy of Sciences* 117 (49):31527–34.
Schwardmann, P. and J. van der Weele (2019). "Deception and self-deception". *Nature Human Behaviour*:1–7.
Shoemaker, D. and M. Vargas (2019). "Moral torch fishing: a signaling theory of blame". *Noûs*.
Siegel, H. (2005). "Truth, thinking, testimony and trust: Alvin Goldman on epistemology and education". *Philosophy and Phenomenological Research* 71 (2):345–66.
Simpson, R. and A. Srinivasan (2018). "No platforming". In: *Academic freedom*. Ed. by J. Lackey. New York: Oxford University Press, pp. 186–209.
Skipper, M. (2020). "Does rationality demand higher-order certainty?" *Synthese*.
— (2021). "The humility heuristic. People worth trusting admit to what they don't know". *Social Epistemology* 35 (3):323–36.
Skitka, L. (2010). "The psychology of moral conviction". *Social and Personality Psychology Compass* 4 (4):267–81.

Skitka, L., C. Bauman, and E. Sargis (2005). "Moral conviction: Another contributor to attitude strength or something more?" *Journal of Personality and Social Psychology* 88 (6):895.

Skitka, L., A. Washburn, and T. Carsel (2015). "The psychological foundations and consequences of moral conviction". *Current Opinion in Psychology* 6:41–4.

Smithies, D. (2012). "Moore's paradox and the accessibility of justification". *Philosophy and Phenomenological Research* 85 (2):273–300.

Song, Y. (2018). "The moral virtue of open-mindedness". *Canadian Journal of Philosophy* 48 (1):65–84.

Stroud, N. (2010). "Polarization and partisan selective exposure". *Journal of Communication* 60 (3):556–76.

Sunstein, C. (2002). "The law of group polarization". *Journal of Political Philosophy* 10 (2):175–95.

Taylor, C. (2011). *Moralism: a study of a vice*. Routledge.

Thomas, J. and R. McFadyen (1995). "The confidence heuristic: a game-theoretic analysis". *Journal of Economic Psychology* 16 (1):97–113.

Titelbaum, M. (2015). "Rationality's fixed point". In: *Oxford Studies in Epistemology*. Ed. by T. Gendler and J. Hawthorne. Vol. 5. New York: Oxford University Press.

Tormala, Z., J. Clarkson, and R. Petty (2006). "Resisting persuasion by the skin of one's teeth: the hidden success of resisted persuasive messages". *Journal of Personality and Social Psychology* 91 (3):423.

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

— (2020). *Grandstanding: the use and abuse of moral talk*. New York: Oxford University Press.

Tsukiura, T. and R. Cabeza (2011). "Remembering beauty: roles of orbitofrontal and hippocampal regions in successful memory encoding of attractive faces". *Neuroimage* 54 (1):653–60.

Upton, C. (2017). "Meditation and the cultivation of virtue". *Philosophical Psychology* 30 (4):373–394.

Vavova, K. (2014). "Confidence, evidence, and disagreement". *Erkenntnis* 79 (1):173–83.

— (2018). "Irrelevant influences". *Philosophy and Phenomenological Research*:134–52.

— (2021). "The limits of rational belief revision". *Noûs* 55 (3):717–34.

Wang, T. et al. (2015). "Is moral beauty different from facial beauty? Evidence from an fMRI study". *Social Cognitive and Affective Neuroscience* 10 (6):814–23.

Webster, S. and A. Abramowitz (2017). "The ideological foundations of affective polarization in the US electorate". *American Politics Research* 45 (4):621–47.

Wedgwood, R. (2012). "Justified inference". *Synthese* 189 (2):273–95.

Whiting, D. (2020). "Recent work on higher-order evidence". *Analysis* 80 (4):789–807.

Williams, B. (1995). *Making sense of humanity*. Cambridge: Cambridge University Press.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.