The reliability challenge to moral intuitions

Dario Cecchini[a]*

[a]*Department of Philosophy and Religious Studies, North Carolina State University, Raleigh, NC*

*dcecchi@ncsu.edu.com

The reliability challenge to moral intuitions

In recent years, the epistemic reliability of moral intuitions has been undermined by substantial empirical data reporting the influence of cognitive biases. This paper discusses and elaborates upon a promising strategy in response to the reliability challenge to moral intuitions. The argument considered appeals to the fact that moral intuitions come in different levels of strength and agents accept only strong intuitions, not vulnerable to bias under realistic circumstances. This essay aims to reconstruct the defense from the reliability challenge in its most promising form and to evaluate the plausibility of the argument in light of the available empirical evidence. What will emerge from the discussion is that the vindication of moral intuitions fundamentally depends on two distinct premises: first, the hypothesis that agents accept moral intuitions proportionally to their level of confidence, and second, the hypothesis that intuitive confidence is epistemically reliable. Whereas there is consistent evidence for the first hypothesis, there is still no conclusive evidence for the second.

Moral intuition; reliability challenge; intuitive strength; intuitive confidence

1. Introduction

In ordinary reasoning and moral theory, agents tend to accept moral beliefs about the wrongness of torture or the intrinsic goodness of pleasure because they are consistent with some strong intuitions. According to many, this common practice presupposes that moral intuitions are a reliable, trustworthy, and unbiased source of knowledge. Yet, in recent years, widespread pessimism surrounding moral intuitions has undermined the validity of this common assumption. Different authors have argued that moral intuitions are not reliable because they are subject to some biases, such as irrelevant framing effects (Machery 2017, Sinnott-Armstrong 2008, McDonald, et al. 2021). I will call this the *reliability challenge* to moral intuition. The challenge is particularly relevant to the methodology of ethics. If moral intuitions are generally reliable, the practice of accepting and rejecting intuitions can be self-regulated (Bengson, Cuneo and Shafer-Landau 2020).

However, if moral intuitions tend to be biased, moral agents should accept only those intuitions supported by independent evidence (Sinnott-Armstrong 2008).

This paper discusses and elaborates upon a promising strategy in response to the reliability challenge to moral intuitions. In short, the argument appeals to the fact that moral intuitions come in different levels of strength. The empirical evidence against the reliability of intuitions would show the unreliability of only a class of weak intuitions concerning complex and artificial moral scenarios (e.g., sacrificial dilemmas). In contrast, the argument concludes, moral agents tend to justify beliefs on the basis of strong intuitions (e.g., intuitions about the wrongness of rape), not vulnerable to bias under realistic circumstances.

Although this line of reasoning has been pointed out by some authors (Wright 2010, Bengson 2013, Liao 2008, Shafer-Landau 2008), it has not been sufficiently explored. In particular, no one has discussed the empirical plausibility of intuitive strength to vindicate the reliability of moral intuitions.[1] This essay has two goals: first, it aims to reconstruct the defense from the reliability challenge in its most promising form; second, to evaluate the plausibility of the argument in light of the available empirical evidence. What will emerge from the discussion is that the vindication of moral intuitions fundamentally depends on two distinct premises: first, the hypothesis that agents accept moral intuitions proportionally to their level of confidence, and second, the hypothesis that intuitive confidence is epistemically reliable. I will argue that, whereas there is consistent evidence for the first hypothesis, there is still no conclusive evidence for the second. Therefore, the defense of the reliability of moral intuitions remains incomplete at this stage of research.

[1] One exception is Egler (2020), who discusses a similar argument applied to nonmoral intuitions.

The paper proceeds as follows. In section 2, starting from a plausible and commonly accepted definition of moral intuition, I clarify what it means for moral intuition to be reliable and why some evidence undermines its reliability. Then, in section 3, I flesh out the core argument in defense of the reliability of moral intuition. Finally, the last two sections discuss the empirical plausibility of the main two premises of the argument, that is, the hypothesis that agents accept moral intuitions proportionally to their strength (section 4) and the reliability of confidence in moral intuitions (section 5).

2. Setting the stage: moral intuitions and the reliability challenge

Moral intuitions, such as the intuition that *killing infants is wrong*, are typically understood as *automatic* and *strong* responses to a morally relevant situation (Cecchini 2023, Haidt 2001). They are automatic because they derive from processes that are to a large extent autonomous—that is, not controlled, fast, and effortless (Bargh 1992, Evans and Stanovich 2013). Moral intuitions are also *strong* mental states insofar as they are experienced as motivating and compelling such that their content is hard to ignore, sometimes even in the face of contrary reflective considerations (Kauppinen 2013, Railton 2014, Cecchini 2023, Loev 2022). Such *intuitive strength* is what distinguishes moral intuitions from “shallow” automatic mental states, such as guesses or quick hypotheses (Bengson 2015).[2]

Moral intuitions have received considerable interest from philosophers and psychologists in recent decades. In particular, a debated question concerns whether intuitions constitute a reliable source of moral knowledge. There are two main rival

[2] The definition of moral intuition as automatic and strong moral cognition is neutral between theories that understand it as a type of belief (Audi 2015), as an emotion (Railton 2014, Kauppinen 2013), or intellectual seeming (Huemer 2005, Bengson 2015, Chudnoff 2013).

approaches to this issue. According to what I define as *intuitions' optimism,*[3] moral intuitions are *prima facie* reliable; in this view, accepting moral intuitions[4] is epistemically justified in absence of contextual defeaters (Bengson 2015, Chudnoff 2013, Bengson, Cuneo and Shafer-Landau 2020). By contrast, according to *intuitions' pessimism*, moral intuitions are not *prima facie* reliable and some independent confirmation is required before accepting them (Sinnott-Armstrong 2008, Greene 2013, Sauer 2021).

It is important to avoid extremes in framing the dialectic between intuitions' optimism and pessimism. The optimistic claim that moral intuitions have default reliability does not entail that intuitions are infallible, nor indefeasible. In fact, it is largely acknowledged by intuitions' optimists that considerations of different kinds can defeat an intuition; for example, the consideration that I grew up in a devout Catholic home might *undermine* the reliability of my intuition that LGBT adoption is wrong; or it might be the case that the reflective consideration that most children who grew up in LGBT families are happy *outweighs* the intuition that LGBT adoption is wrong. This means that accepting moral intuitions is *generally* permissible and, in the absence of defeaters, a subject can trust her own intuitions. On the other hand, the pessimist position denying that intuitions are *prima facie* reliable does not necessarily entail the eliminativist position according to which intuitions have no role in moral reasoning. Rather, intuitions'

[3] I avoid using the term "intuitionism" because the term has been historically employed to denote a non-naturalist moral realist theory in metaethics. What I call "intuitions' optimism" here is silent about the metaphysics of moral facts.

[4] By "accepting moral intuitions", I mean any voluntary and explicit act of assent to moral intuitions: for instance, *forming* a belief based on intuition, as well as *maintaining, endorsing,* or *sustaining* a belief based on intuition.

pessimism is consistent with the view that intuitions ground moral beliefs if filtered by epistemic tools, such as philosophical expertise or *ad hoc* experimental conditions.

Defining epistemic reliability for moral intuitions is not an easy task. Each theory of knowledge can have a different understanding of the concept. Here, I understand reliability as the feature of a mental process that is inherently capable of preventing those biases that one should avoid in making moral judgments. I mean, for instance, *evolutionary biases*, which aim at fitness rather than rational moral inquiry, *social biases*, i.e., systematic prejudices toward one group of people, or *cognitive biases*, i.e., systematic mistakes in interpreting information from a problem (Kahneman 2011). Therefore, following this definition, intuitions' optimism claims that moral intuitions are generally capable *per se* of preventing such kinds of biases, whereas pessimism argues that moral intuitions need external support for reaching that standard.

Note that the definition of reliability adopted here is particularly apt for framing the discussion between intuitions' optimism and pessimism because it is not committed to any substantive moral truth. The claim that a subject should avoid certain kinds of bias can be understood as an epistemic norm of inquiry valid for both moral and nonmoral judgments. This has the advantage of preventing the objection raised by Rini (2016), according to which any reliability assessment of moral intuitions entails a vicious circularity because it must assume the reliability of other moral intuitions.[5]

In recent years, an increasing number of experimental studies have questioned the reliability of lay-people's moral intuitions. More specifically, many studies report that moral intuitions are subject to cultural, gender, personal, and cognitive biases (see

[5] Here, I am following Yeo's response to Rini (2016) by setting moral truths and epistemic norms on different levels of generality (Yeo 2020).

Machery 2017, 45-89 for a review). Probably, the most substantial and convincing body of evidence concerns the influence of framing effects on moral intuitions (see McDonald et al. 2021 for a meta-analysis). Such evidence includes studies reporting that the *words* describing a moral scenario affect people's judgment. For example, subjects are more prone to judge positively a certain action if told how many people the action *saves* than if they are told how many people it *kills*, even though the outcome is the same. Other studies show that the *order* in which different scenarios are presented influences moral intuitions. For example, it has been reported that manipulating the order of presentation of scenarios involving harm brought by action or omission has some effect on how negatively people regard the omission. Since being systematically influenced by the *mere* framing (e.g., the words used or the order of presentation) is commonly considered a cognitive bias,[6] I take this line of evidence as a paradigmatic case against the reliability of moral intuitions.

In sum, if the evidence on framing effects is compelling, moral intuitions are subject to biases; thus, intuitions' optimism is undermined. Call this the *reliability challenge* to moral intuitions. This challenge has serious implications for how we treat intuitions in moral reasoning and theory.[7] Thus, if accepting moral intuitions is a self-regulated practice as intuitions' optimists suggest, these latter should find a strategy to reject the external validity of the abovementioned evidence; otherwise, moral theorists

[6] Horne and Livengood (2017) have questioned this common assumption, but I will not consider this move here.

[7] Epistemic conservatists like Humer (2005, 2007) may disagree on this point: according to this view, accepting moral intuitions are prima facie justified *regardless* of the reliability of the process generating them. However, discussing this position goes beyond the purposes of this work.

would be constrained to use *ad hoc* debiasing tools before considering their intuitions, as intuitions' pessimists point.

3. A promising defense of intuitions' optimism

Different responses to the reliability challenge have been provided in recent years (Bengson 2013, Rini 2016, Bengson, Cuneo and Shafer-Landau 2020). The line of defense I discuss in this paper appeals to the fact that moral reasoners can autonomously prevent biases according to how a certain intuition is experienced. To assess this capacity, one should consider *what types* of intuitions are subject to biases and whether agents are capable of accepting intuitions of the right type. As Weinberg (2007, 323-327) points out, this is crucial to evaluate whether a certain epistemic source is "hopeful" enough to be trusted.

In order to better appreciate this point, a comparison with sensory perceptions is helpful. Perceptions are usually considered reliable not because they are infallible but mainly because the subjects know under what conditions they can trust them; the subjects know, for instance, that perceptual experiences do not deserve much epistemic credit when they are foggy, or like when one is under the effect of drugs or alcohol. Thus, some perceptual mistakes can be easily prevented by how some types of perception are experienced. In a similar vein, it is important to assess whether subjects can prevent cognitive biases by tracking the reliability of their intuitions in a context-sensitive way, beyond evaluating their absolute susceptibility to biases.

A plausible way by which agents can track the reliability of intuitions is through the strength in which intuitions are experienced. Moral intuitions can have different degrees of strength: people have stronger and weaker intuitions. Supposedly, agents tend to accept moral intuitions *proportionally to their level of strength*. To put it more bluntly, the stronger an intuition is experienced, the more a subject is disposed to accept it, i.e., to maintain or endorse it; conversely, the weaker an intuition is, the greater a subject is disposed to revise it. Thus, if intuitive strength is a reliable indicator of the presence of biases, moral reasoners can prevent them by how intuitions are experienced.

Importantly, data on framing effects collected thus far do not consider the strength of moral intuitions, but only to what extent the mere framing affects their content, that is what type of answer subjects are inclined to endorse (typically, consequentialist versus deontological). Furthermore, it is noteworthy that most studies designed to assess framing effects employ complex and unfamiliar moral scenarios, such as different versions of the trolley dilemma. To the extent that problems of such kind pit conflicting moral reasons against each other, they may elicit weak intuitions that the subjects tend not to consider seriously outside of the laboratory. Therefore, to the extent that they do not measure intuitive strength and do not consider the possibility that the subjects prevent biases through it, experimental studies on framing effects might be inadequate to assess the overall reliability of moral intuitions.

One could object that some studies do not collect simple “yes or no” answers but utilize moral acceptability scales (e.g. from 1 to 6) and, as Andow (2016) notices, the mere framing of moral scenarios seems to affect the level of extremity of people’s responses, that is how close the judgments are to the extremes of the scale. However, judgments extremity and strength should not be conflated because they capture different aspects of moral intuition: the strength of an intuition concerns how the intuition is *experienced* (as likely, confident, compelling), regardless of the severity or permissive nature of its content. Possibly, a subject can have an extreme intuition (for example a very harsh or liberal intuition about the legalization of drugs) but still feels uncertain about it and thus disinclined to assign much credibility to it.[8]

Different authors have expressed similar concerns toward empirical studies challenging the reliability of intuitions. For instance, Shafer-Landau, in reply to the evidence from framing effects, points out that there is a class of moral intuitions, such as intuitions about the wrongness of torture, rape, or deliberate humiliation, that are unlikely to be vulnerable to external influences:

> They are genuine moral beliefs, and the evidence about framing effects casts no doubt on their reliability. Neither does this evidence impugn the reliability of more specific, entirely uncontroversial moral beliefs […] These are beliefs that are (for almost everyone) not subject to framing effects: They are invulnerable to change under realistic circumstances. (Shafer-Landau 2008, 92).

In line with such considerations, Liao contends that the experimental evidence against the reliability of moral intuitions does not consider the distinction between "surface" and "robust" intuitions, which are the real justifiers in philosophical theorizing:

> Some might think that one should distinguish between surface intuitions, which are 'first-off' intuitions that may be little better than mere guesses; and robust intuitions, which are intuitions that a competent speaker might have under sufficiently ideal conditions such as when they are not biased. In other words, when philosophers assert that 'Everyone would agree that ...' or 'Intuitively, we would all find it obvious

[8] Note that the conceptual claim that judgment extremity and strength can be dissociated does not exclude that the two notions are empirically correlated for some reason, as reported by some studies (Mata 2019, Vega, et al. 2020, Heinzelmann, Holtgen and Tran 2021).

> that ...' or 'It is clear to us that ...', the 'we' and 'us' should be interpreted as applying only to competent speakers in certain non-distorting conditions. (Liao 2008, 256)

Similarly, Bengson argues that the experimental evidence does not distinguish between "unstable answers", i.e., guesses or quick hypotheses, generated by unfamiliar and not commonsensical scenarios, and "stable answers", i.e., genuine intuitions elicited by commonsensical and familiar scenarios (Bengson 2013, 522-523). Note that Bengson does not discriminate between strong and weak intuitions but separates intuitions from non-intuitions, which he calls "blind answers". However, the core of the argument is the same: the evidence from framing effects misses the target because the class of judgments they undermine is not the same as the one to which intuitions' optimism refers.

Although the aforementioned authors implicitly or explicitly appeal to intuitive strength to address the reliability challenge, none of them seem to consider sufficiently the potential of intuitive strength to vindicate the reliability of moral intuitions. For this purpose, one can construct an argument based on the considerations outlined in this section. The argument is based on two independent premises. The first premise (P1) states that agents accept moral intuitions proportionally to their level of strength. The second premise (P2) affirms that intuitive strength is epistemically reliable. Therefore, in virtue of a bridge premise (P3), the argument concludes that agents accept moral intuitions proportionally to their reliability (C):

> (P1) Agents accept moral intuitions proportionally to their level of strength.
> (P2) Intuitive strength is epistemically reliable.
> (P3) If P1 and P2, then agents accept moral intuitions proportionally to their reliability.
> (C) Therefore, agents accept moral intuitions proportionally to their reliability.

If the argument is sound, moral agents can prevent biases by an intrinsic feature of moral intuition: intuitive strength. That means that intuitions' optimism is true: moral intuitions are prima facie reliable.[9]

P3 is a plausible statement: if intuitive strength is a reliable indicator of the presence (or the absence) of bias and agents weigh their intuitions according to their strength, that means that they track epistemic reliability through it. Taking P3 for granted, in the next sections, I will focus on the most substantive empirical hypotheses assumed by the argument, P1 (section 4) and P2 (section 5).

4. Intuitive strength and its cognitive function

Strength is a characteristic feature of moral intuitions that has been neglected for many years. Only recently, some authors (Wright 2013, Andow 2016, Cecchini 2023) have discussed the important *cognitive function* it performs, that is how moral reasoning is guided by the strength with which certain intuitions are experienced. Intuitions' optimism predicts that agents tend to accept strong intuitions and reject weak ones. This is in line with the first premise of the argument stated above (P1). To argue for this hypothesis, one has to provide a psychologically plausible account of intuitive strength (4.1) and, based on it, find evidence that links intuitive strength with beliefs acceptance (4.2). I will argue that interpreting strength as *confidence* is conceptually plausible and consistent with its supposed cognitive function.

4.1. Intuitive strength as confidence

Understanding intuitive strength is crucial for assessing the reliability of intuitions. As

[9] This argument can be classified as a *vindicating argument* in favor of the legitimacy of moral intuitions. On the opposite of a debunking argument, a vindicating argument aims to defend the legitimacy of a class of beliefs (in the present case, of intuition-based beliefs) by pointing out that the psychological process on which they are based is reliable (Sauer 2018, 209).

stated previously, unlike mere guesses and quick hypotheses, intuitions are experienced as "compelling" and their content is perceived as likely and credible. However, these general phenomenological features apart, specific accounts diverge in how to explain the strength of moral intuitions. In what follows, I consider three main views: the *perceptual* account, the *emotional* account, and the *metacognitive* account. Then, I provide some reasons why the metacognitive account is preferable to the other views.

According to the perceptual account, moral intuitions are *presentational states* (Chudnoff 2013, Bengson 2015). Like sensory perceptions, intuitions present the world as being in a certain way. In other words, intuitions provide the impression that things stand in the way they are represented; for example, if one has the intuition that killing babies is wrong, one has the vivid impression that killing babies is wrong. In this view, intuitive strength denotes the degree of *presentational phenomenology* of an intuition. In more simple words, the strongest intuitions are those that present some content that strikes most as true.

A second candidate is the emotional account, according to which intuitive strength is reducible to the intensity of the moral emotion elicited by a certain fact (Haidt 2001, Railton 2014, Kauppinen 2013). For example, the intuition that killing babies is wrong is as strong as one feels outraged while having that intuition. This account is attractive since it explains the documented correlation between moral emotions and intuition (Ugazio, Lamm and Singer 2012, Decety and Cacioppo 2012), the alleged *motivational force* of moral intuitions (Kauppinen 2013), and their *recalcitrant* nature, that is the fact that strong intuitions can occur in tension with reflective beliefs.

Finally, according to the metacognitive account, the strength of moral intuitions denotes the level of subjective confidence about a certain moral content (Cecchini

2023).[10] In this view, intuitive strength results from the *fluency* with which the information is processed. That is, the more fluently and easily a certain stimulus is processed, the greater the confidence in the automatic response and the stronger the resultant intuition. For instance, the intuition that killing babies is wrong is typically experienced as strong to the extent that for people who have western moral education, it is quite easy and familiar to think about this fact as wrong; by contrast, people may have a weaker intuition that turning the switch in the trolley dilemma is permissible given the complexity and unfamiliarity of the problem.

Arguably, the metacognitive account seems to be the most convincing one for different reasons. First, unlike the perceptual account, it demystifies moral intuition by understanding it as a confident automatic cognition, rather than a *sui generis* mental state such as an intellectual perception. Second, intuitive confidence is easy to operationalize by self-reported measures and indirect measures, such as response time. Third, the metacognitive feeling of confidence explains the phenomenology of strong intuition, that is the fact that the content of strong intuitions is experienced as "compelling" and likely, as opposed to guesses or quick hypotheses, which are accompanied by a sense of uncertainty. Fourth, as I will show in the next subsection in more detail, the metacognitive account explains the cognitive function of intuitive strength, i.e., how subjects regulate the activation of cognitive resources in relation to the strength of their intuitions. Fifth, and finally, the metacognitive account can accommodate the correlation between moral emotions and intuitive strength, while at the same time explaining why the two concepts

[10] Even though he labels his account as "affectivism", Loev (2022) converges on this view to the extent that he explains the strong character of intuitions though epistemic metacognitive feelings.

can diverge. Specifically, the metacognitive account connects moral emotions with confidence through processing *fluency*: since emotions are great cognitive facilitators, they tend to speed up information processing and, consequently, favor the generation of confident intuitions; nevertheless, other determinants of fluency, such as the familiarity or "paradigmaticity" of a moral concept (Wright 2010), may replace emotion in this role and that explains why people can have strong intuitions with low emotional force (e.g., the intuition that benevolence is a virtue or that freedom is a fundamental value) (Cecchini 2023, 15).

This is not the place to defend the metacognitive account in detail (see Cecchini 2023 and Loev 2022 for an extensive discussion). However, the considerations outlined above suggest that intuitive strength can be legitimately interpreted as confidence. An additional reason for this assumption is pragmatic: the most relevant studies published to date focusing on the cognitive role of intuitive strength operationalize it as confidence. To my knowledge, no empirical study has explicitly tested if moral emotions or perceptual veridicality predict moral belief acceptance, and without empirical evidence it is hard to defend the key premises of the argument in defense of moral intuitions. Therefore, in the remainder of the paper, I will refer to strong intuitions as automatic moral responses experienced with a substantial degree of confidence.

4.2. Confidence and moral beliefs acceptance: a rapid review

With the metacognitive account in mind, we can now consider the first premise of the argument in defense of intuitions' optimism, which states that agents accept moral intuitions proportionally to their level of strength (i.e., confidence). That means that the more a subject feels confident about a certain intuition, the more she tends to accept it, i.e., endorse it, and maintain it. Said otherwise, the most confident intuitions are the most *stable* ones, that is, those less prone to revisions. Therefore, P1 entails a positive correlation between confidence and stability.

The link between confidence and stability is an empirical hypothesis. Indeed, confidence and stability are two logically distinct concepts, captured by distinct experimental measures. Confidence is the subjective feeling of ease and fluency with which a certain intuition comes to mind. In the literature, the confidence of intuitions is measured by different tools. Among these, experimenters often use self-reports by directly asking them *how confident they feel* about the response provided (Thompson, Turner and Pennycook 2011) or *how difficult* the judgment was (Bago and De Neys 2019); others use indirect questions like "how *conflicted* did you feel when responding to the problem?" or "How much are you *aware of possible disagreement* on your response?" (Mata 2019). The most common non-self-reported measure employed to assess confidence is response time: since confident intuitions are supposed to come fluently and easily, the time employed by a subject to respond to a moral problem is considered an indirect indicator of confidence. Furthermore, mouse-tracking technology has been utilized by some (Koop 2013, Gürcay and Baron 2017) to infer confidence. In this kind of experimental set, answers to a moral problem are located on opposite sides of the screen and the experimenters can observe the subjects' trajectory of the mouse while they respond: supposedly, the more straightforward the trajectory and the less it swings, the more confident the subject.

In contrast with confidence, stability is not a subjective feeling but a behavioral tendency, which can be defined as the extent to which a subject tends to endorse an intuition. This tendency can be captured by assessing how certain moral responses remain unchanged despite circumstantial factors, such as exposure to relevant (or irrelevant) information or changes in the experimental setting (Wright 2013). Another employed method is to measure people’s reflection time as an indicator of their tendency to reconsider their automatic intuition (Bago and De Neys 2019). Accordingly, the less time a subject spends to reflect, the more she is disposed to trust her intuition. Therefore, since confidence and stability are distinguishable concepts, a positive correlation between the two cannot be taken for granted, though it appears to be plausible at first glance.

Some studies seem to support the hypothesis that intuitive confidence is predictive of stability. For example, Zamzow and Nichols (2009, 373-374) find that the most confident moral judgments are also the most resistant to changes in the order of presentation of dilemmas; similar results have been replicated by Wright (2010, 2013). Other studies report a significant negative correlation between confidence and the time spent before arriving at a judgment (Bialek and De Neys 2016, 2017). Possibly, that means the more a subject feels confident in a certain intuition, the less she is disposed to reflect and reconsider it through reasoning. Thus, these data may count as indirect evidence for the link between intuitive confidence and stability.

In recent years, two studies have confirmed the data described above (Bago and De Neys 2019, Vega, et al. 2020). Following Thompson and colleagues (2011), these studies adopt a “two response” paradigm to assess how intuitive confidence (defined as “Feeling of rightness”) is predictive of the subjects’ tendency to reflect and revise their responses. This method divides the experiment into two stages. In the first stage, to knock out their cognitive resources, the participants are instructed to respond to a moral dilemma as quickly as possible with the first answer that comes to mind. Then, in the second stage, the moral problem is presented again, and the subjects can spend as much time as they need to provide a final answer. Importantly for our purposes, in both the mentioned studies, the confidence of the participants’ initial responses turns out to be predictive of the time spent before providing the final judgment and their tendency to revise their initial answer in the second stage. Furthermore, consistent with the metacognitive account, the level of intuitive confidence turns out to be significantly correlated with the fluency (measured by response time) of the initial judgment in Vega and colleagues' study.

In sum, although the quantity of collected data is not overwhelming, all these findings are consistent with the hypothesis that intuitive confidence is a reliable indicator of the stability of moral judgment. This supports the first premise of the argument according to which agents accept their intuitions in proportion to the level of experienced strength. Additionally, the evidence discussed above confirms that understanding intuitive strength as confidence is a promising way to explain its cognitive function.

5. Is confidence epistemically reliable?

Thus far, I have shown that subjective confidence is the most plausible way to understand the strength of moral intuitions. Additionally, I have reviewed some evidence for the claim according to which agents accept moral intuitions proportionally to their level of confidence. However, as mentioned, the mere fact that people tend to accept strong intuitions and revise weak ones is not sufficient to conclude that moral intuitions are reliable. Rather, for this conclusion, intuitions' optimists should show that intuitive strength is a reliable indicator of the presence (and the absence) of bias; and this empirical hypothesis is independent of the one considered in the previous section. Therefore, assuming the plausibility of the metacognitive account, intuitions' optimists should consider whether intuitive confidence is epistemically reliable (P2). This section discusses that claim. More specifically, starting from some general rational principles, I review the evidence on the reliability of confidence in the moral domain (5.1). Then, I consider whether confidence is generally reliable in nonmoral domains (5.2).

5.1. Confidence and moral judgment: rational principles and rapid review

The epistemic reliability of confidence is an empirical conjecture and not a self-evident truth. One can be confident about certain claims according to good or bad reasons. Whether agents' confidence *generally* responds to rational principles is an open empirical question. This issue has been widely investigated in the literature on metacognition (Shekhar and Rahnev 2021). Nevertheless, evidence on confidence in moral judgment is scarce, and, hence, at this stage of research, considerations about this question can be at best conjectural.

To review the available studies, I proceed as follows (see Table 1). First, I divide possible determinants of confidence into two groups: *situational* and *personal* factors.[11] Then, for each group of factors, I establish a rational principle that discriminates between what factors rationally justify some change in confidence and those that do not; in other words, what is epistemically relevant and what is a bias. Such criteria must be plausible epistemic principles, independent of any substantive disagreement among ethical theories. Finally, in light of the principles, I consider the main determinants studied in the literature and whether the collected data follow rational expectations.

Let us start with *situational* factors, that is, characteristics of either the moral problem at stake or its setting. Arguably, if a situational factor undermines an agent's cognitive resources (e.g., available time, attention, or information), one rationally expects that the resultant intuition will be less confident than it would be without that factor. Conversely, if a situational factor increases the agent's resources, one expects a more confident intuition. Therefore, generalizing from these considerations, one can assume the following principle:

[11] I take this distinction from Klenk (2021).

(S) A situational factor *s* justifies some change in confidence if, and only if, *s* affects agents' cognitive resources required for making a moral judgment

Attention is one of the crucial resources that a subject has to rely on to consider a moral problem. Therefore, a rational prediction is that inhibiting subjects' attention through a distracting task makes subjects' confidence in their intuitions decrease. This condition is called *cognitive load* and the rational expectation is confirmed in Bago and De Neys (2019). The time available to consider a moral problem is also an important factor that should affect confidence. In line with this prediction, both Bago and De Neys' and Vega et al.'s studies report that instructing subjects to respond as quickly as possible (*time pressure* condition) decreases subjects' confidence in their intuitions. Furthermore, since the quantity and the quality of information available contribute to the difficulty of a moral problem, confidence should be modulated according to them. Indeed, the two abovementioned studies and Mata (2019) report that increasing the conflict between reasons in a moral problem (utilitarian vs. deontological) makes people's confidence decrease. Finally, even exposure to experts' disagreement is relevant information that should undermine confidence and this rational expectation is confirmed in Wright's study (2013).

The results outlined above are good news for intuitions' optimism because they show that people's confidence about moral intuitions seems to respond to relevant factors. However, to get more robust confirmation for the reliability of confidence in the moral domain, one should test whether it does *not* respond to biases, i.e., factors that do not affect agents' cognitive resources. For example, confidence should not track factors like the order of presentation of moral problems, the words used, or incidental induced feelings (e.g., questioning subjects from dirty desks). Unfortunately, no data about the effect of such biases on confidence have been published to date.[12] Moreover, in a recent study, Heinzelmann et al. (2021) find that exposing some subjects to peer disagreement increases their confidence in their moral opinions. This finding goes against our rational expectation since one expects that awareness of disagreement will either diminish one's confidence or leave it unchanged. Therefore, all considered, there is not enough evidence to conclude that confidence in moral intuitions responds to principle S.

Let us consider now *personal* factors, namely dispositions of the agents that may influence moral judgment. Arguably, if a subject is extremely competent with moral judgment, she will be entitled to be more confident, while a less competent judge should be more prudent in confidence. On this basis, one can state that, in general, a subjective trait is relevant in determining confidence only if it contributes to an agent's overall competence with the moral problem at stake:

(P) A personal factor *p* justifies some change in confidence if, and only if, *p* is relevant to the competence of the agent in making a moral judgment

There is metaethical disagreement on what constitutes moral expertise. Probably, one can assume that *experience* with a certain class of moral problems, as well as ethical knowledge, justify some increase in confidence. In contrast, any irrelevant demographic factors or social prejudices do not justify any change in confidence since they do not contribute to the agent's competence. These predictions would be very important to test the reliability of moral confidence, but no study has investigated them yet.

[12] Some studies investigating the variation of confidence in nonmoral problems report the influence of framing effects (see Egler 2020, 55-57 for a review).

To summarize, confidence is reliable only if it responds to rational principles governing situational and personal factors. It is too premature to conclude that agents' confidence in their moral intuitions responds to such principles since the evidence is scarce and mixed. The hope of this review is to enhance future empirical studies on how people are influenced by certain factors in calibrating confidence in the moral domain.

Principle	Factor	S/P	Rational expectation	Reference
A situational factor *s* justifies some change in confidence if, and only if, *s* affects agents' cognitive resources required for making a moral judgment	Cognitive load	S	Decrease in confidence	Bago and De Neys 2019
	Conflicting reasons (and reasons weight)	S	Decrease in confidence	Bago and De Neys 2019, Vega, et al. 2020, Mata 2019
	Peer disagreement	S	Either no change or decrease in confidence	Heinzelmann, Holtgen and Tran 2021
	Time pressure	S	Decrease in confidence	Bago and De Neys 2019, Vega, et al. 2020
	Expert disagreement	S	Decrease in confidence	Wright 2013
	Order of presentation	S	No change	
	Words used	S	No change	
	Incidental feelings	S	No change	
A personal factor *p* justifies some change in confidence if, and only if, *p* is relevant to the competence of the agent in making a moral judgment	Experience with the problem	P	Increase in confidence	
	Ethical knowledge	P	Increase in confidence	
	Demographic factors	P	No change	
	Social prejudices	P	No change	

Table 1. Review of confidence determinants according to rational principles.

5.2. Confidence beyond the moral domain

To the extent that evidence on confidence reliability in the moral domain is still little, intuitions' optimists could look beyond the moral domain and consider the reliability of confidence in other kinds of problems. If people's confidence is generally reliable across different domains, there is no reason to be skeptical about it in the moral domain. This strategy seems promising: after all, cognitive capacities have evolved to help humans navigate the world and, consistent with that, subjective feelings of confidence should be reliable indicators for receiving vital information from the environment. Therefore, stating that confidence is unreliable would be at odds with human evolution.

Some evidence on the ecological validity of cognitive fluency supports the line of reasoning above. Specifically, a large amount of studies reviewed in Herzog and Hertwig (2012) shows that how fluent certain thoughts are is a valid cue to infer objective properties of the environment, such as numerical quantities, true statements, the dangerousness of objects, or social information. That said, there are still some important findings that cast doubts about the reliability of the confidence hypothesis. In particular, two challenges are worth mentioning.

The first challenge is the finding that subjective confidence tracks "consensuality" rather than truth (Koriat 2012). That means that people can be very confident about statements commonly believed as true even if they are false (e.g., that Sydney is the capital of Australia). This happens because repeated exposure to certain propositions increases familiarity and fluency of thought regardless of the content of the propositions. Because of their hedonic feeling, fluency and confidence discourage continued effort in a task (Fiedler 2012). Consequently, this may prevent subjects from learning uncommon truths. If confirmed in the moral domain,[13] the close link between confidence and perceived consensus has the effect of blocking moral progress by not tracking non-mainstream moral truths. Therefore, the "consensuality principle" constitutes a serious case against the reliability of confidence.

The second challenge is the well-known *Dunning-Kruger effect* (Dunning 2011). According to this hypothesis, incompetent subjects in a given domain tend to be

[13] Mata (2019) explores this hypothesis.

overconfident because they overestimate their competence; in contrast, competent subjects tend to be slightly underconfident. The rationale behind this is that ignorance is invisible under a certain threshold of competence: the incompetent subject, exactly because of her lack of competence, does not have any clue to detect the flaws in her knowledge. The trend has been reported consistently across a large variety of tasks (see Dunning 2011 for a review) and, if confirmed in the moral domain, it is highly problematic. In particular, the Dunning-Kruger effect is a clear violation of principle P, which prescribes that confidence should be proportional to the level of confidence possessed by a subject.

In sum, although appealing to the general reliability of confidence seems plausible at first glance, this move is not devoid of problems. Specifically, one has to face two important empirical challenges. At this point, the intuitions' optimist has two options: either showing that the two challenges described above do not apply to the moral domain for some reason or arguing that their impact is not strong enough to undermine the hypothesis that confidence is generally reliable. Another possibility would be to embrace other accounts of intuitive strength. Yet, it is hard to see this move as advantageous since perceptual states are based on processes very close to the ones producing confidence and the general reliability of emotions is at least as disputed as the reliability of confidence.

6. Conclusion

The reliability of moral intuitions has been questioned by empirical data showing that defective processes influence moral intuitions. These results challenge the optimistic claim that moral reasoning is a self-regulated practice inherently capable of preventing biases.

This paper had two goals: reconstructing a promising argument in defense of the reliability of intuitions and evaluating the plausibility of its main premises. For the first goal, I have shown that the argument is based on two main logically independent empirical hypotheses: first, the claim that agents accept moral intuitions proportionally to their level of confidence, and second, the claim that intuitive confidence is epistemically reliable. Regarding the second goal, I have shown that while there is substantial evidence for the claim that agents assign credibility to intuitions in proportion to how confident they feel, the evidence for the reliability of intuitive confidence is still inconclusive in the moral domain and beyond. This means there is still work to do for intuitions' optimism to tackle the reliability challenge. However, this discussion has highlighted the importance of confidence in the evaluation of the reliability of moral intuitions. The hope is that this enhances new empirical research on confidence calibration in the moral domain.

References

Andow, J. 2016. "Reliable but not home free? What framing effects mean for intuitions." *Philosophical Psychology* 29 (6): 904-911.

Audi, R. 2015. "Intuition and Its Place in Ethics." *Journal of the American Philosophical Association* 1 (1): 57–77.

Bago, B., and W. De Neys. 2019. "The Intuitive Greater Good: Testing the Corrective Dual Process Model of Moral Cognition." *Journal of Experimental Psychology: General* 148 (10): 1782-1801.

Bargh, J.A. 1992. "The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects." *The American Journal of Psychology* 105 (2): 181-199.

Bengson, J. 2013. "Experimental Attacks on Intuitions and Answers." *Philosophy and Phenomenological Research* 86 (3): 495-532.

Bengson, J. 2015. "The Intellectual Given." *Mind* 124 (495): 707-760.

Bengson, J., T. Cuneo, and R. Shafer-Landau. 2020. "Trusting Moral Intuitions." *Nous* 54 (4): 956-984.

Bialek, M., and W. De Neys. 2016. "Conflict Detection During Moral Decision-Making: Evidence." *Journal of Cognitive Psychology* 28 (5): 631–639.

Bialek, M., and W. De Neys. 2017. “Dual processes and moral conflict: Evidence for deontological reasoners’ intuitive utilitarian.” *Judgment and Decision Making* 12: 148-167.

Cecchini, D. 2023. “Moral intuition, strength, and metacognition.” *Philosophical Psychology* 36 (1): 4-28.

Chudnoff, E. 2013. *Intuition.* Oxford: Oxford University Press.

Decety, J., and S. Cacioppo. 2012. “The speed of morality: a high-density electrical neuroimaging study.” *Journal of Neurophysiology* 108: 3068–3072.

Dunning, D. 2011. “The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance.” In *Advances in Experimental Social Psychology. Vol 44,* edited by M.P. Zanna and J.M. Olson, 247-296. San Diego: Elsevier.

Egler, M. 2020. “Testing for the phenomenal: Intuition, metacognition, and philosophical methodology.” *Mind and Language* 35: 48-66.

Evans, J., and K. Stanovich. 2013. “Dual-Process Theories of Higher Cognition: Advancing the Debate.” *Perspectives on Psychological Science* 8 (3): 223-241.

Fiedler, K. 2012. “Fluency and behavior regulation: Adaptive and maladaptive consequences of a good feeling.” In *The Experience of Thinking: How the Fluency of Mental Processes Influences Cognition and Behaviour*, edited by C. Unkelbach and R. Greifeneder, 234-254. Abingdon: Taylor and Francis.

Greene, J. 2013. *Moral Tribes: Emotions, Reason, and The Gap Between Us and Them.* New York: The Penguin Press.

Gürcay, B., and J. Baron. 2017. “Challenges for the Sequential Two-system Model of Reasoning.” *Thinking and Reasoning* 23(1): 49-80.

Haidt, J. 2001. “The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814-834.

Heinzelmann, N., B.T.A. Holtgen, and V. Tran. 2021. “Moral discourse boosts confidence in moral judgments.” *Philosophical Psychology* 1-25.

Herzog, S.M., and R. Hertwig. 2012. “The ecological validity of fluency.” In *The Experience of Thinking: How the Fluency of Mental Processes Influences Cognition and Behaviour,* edited by C. Unkelbach and R. Greifeneder, 190-219. Abingdon: Taylor and Francis.

Horne, Z., and J. Livengood. 2017. “Ordering effects, updating effects, and the specter of global skepticism.” *Synthese* 194: 1189-1218.

Huemer, M. 2005. *Ethical Intuitionism.* HoundmiUs, Basingstoke: Palgrave Macmillan.

Huemer, M. 2007. "Revisionary Intuitionism." *Social Philosophy and Policy* 25 (1): 368-392.

Kahneman, D. 2011. *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kauppinen, A. 2013. "A Humean theory of moral intuition." *Canadian Journal of Philosophy* 43 (3): 360–381.

Klenk, M. 2021. "The Influence of Situational Factors in Sacrificial Dilemmas on Utilitarian Moral Judgments." *Review of Philosophy and Psychology* 1-33.

Koop, G. 2013. "An Assessment of the Temporal Dynamics of Moral Decisions." *Judgment and Decision Making* 8(5): 527–539.

Koriat, A. 2012. "The Self-Consistency Model of Subjetive Confidence." *Psychological Review* 119 (1): 80-113.

Liao, S.M. 2008. "A defense of intuitions." *Philosophical Studies* 140: 247–262.

Loev, S. 2022. "Affectivism about intuitions." *Synthese* 200 (274): 1-24.

Machery, E. 2017. *Philosophy Within Its Proper Bounds.* Oxford: Oxford University Press.

Mata, A. 2019. "Social Metacognition in Moral Judgment: Decisional Conflict Promotes Perspective Taking." *Journal of Personality and Social Psychology* 117(6): 1061–1082.

McDonald, K., R. Graves, S. Yin, T. Weese, and W. Sinnott-Armstrong. 2021. "Valence framing effects on moral judgments: A meta-analysis." *Cognition* 212: 104703.

Railton, P. 2014. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics* 124: 813-859.

Rini, R. 2016. "Debunking debunking: a regress challenge for psychological threats to moral judgment." *Philosophical Studies* 173: 675–697.

Sauer, H. 2021. "Against moral judgment. The empirical case for moral abolitionism." *Philosophical Explorations* 24 (2): 137-154.

—. 2018. *Debunking Arguments in Ethics.* Cambridge: Cambridge University Press.

Shafer-Landau, R. 2008. "Defending Ethical Intuitionism." In *Moral Psychology. Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, edited by W. Sinnott-Armstrong, 83-96. Cambridge, MA: The MIT Press.

Shekhar, M., and D. Rahnev. 2021. "Sources of Metacognitive Inefficiency." *Trends in Cognitive Sciences* 25 (1): 12-23 .

Sinnott-Armstrong, W. 2008. "Framing Moral Intuitions." In *Moral Psychology, vol. 2*, edited by W. Sinnott-Armstrong, 47-76. Cambridge, MA: The MIT Press.

Thompson, V., J.P. Turner, and G. Pennycook. 2011. "Intuition, Reason and Metacognition." *Cognitive Psychology* 63: 107-140.

Ugazio, G., C. Lamm, and T. Singer. 2012. "The Role of Emotions for Moral Judgments Depends on the Type of Emotion and Moral Scenario." *Emotion* 12 (3): 579-590.

Vega, S., A. Mata, M.B. Ferreira, and A.R. Vaz. 2020. "Metacognition in Moral Decisions: Judgment Extremity and Feeling of Rightness in Moral Intuitions." *Thinking and Reasoning* 20 (2): 215-244.

Weinberg, J.M. 2007. "How to Challenge Intuitions Empirically Without Risking Skepticism." *Midwest Studies in Philosophy* 21: 318-343.

Wright, J.C. 2010. "On intuitional stability: The clear, the strong, and the paradigmatic." *Cognition* 115: 491–503.

Wright, J.C. 2013. "Tracking instability in our philosophical judgments: Is it intuitive?" *Philosophical Psychology* 26 (4): 485–501.

Yeo, S.L. 2020. "Defusing the Regress Challenge to Debunking Arguments." *Canadian Journal of Philosophy* 50 (6): 785-800.

Zamzow, J.L., and S. Nichols. 2009. "Variations in Ethical Intuitions." *Philosophical Issues* 19: 368-388.