Cognition 234 (2023) 105379

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

People recognize and condone their own morally motivated reasoning

Corey Cusimano [a,*], Tania Lombrozo [b]

[a] School of Management, Yale University, United States of America
[b] Department of Psychology, Princeton University, United States of America

ARTICLE INFO

Keywords:
Moral judgment
Belief
Motivated reasoning
Metacognition
Bias blind spot
Naïve realism

ABSTRACT

People often engage in biased reasoning, favoring some beliefs over others even when the result is a departure from impartial or evidence-based reasoning. Psychologists have long assumed that people are unaware of these biases and operate under an "illusion of objectivity." We identify an important domain of life in which people harbor little illusion about their biases – when they are biased for moral reasons. For instance, people endorse and feel justified believing morally desirable propositions even when they think they lack evidence for them (Study 1a/1b). Moreover, when people engage in morally desirable motivated reasoning, they recognize the influence of moral biases on their judgment, but nevertheless evaluate their reasoning as ideal (Studies 2–4). These findings overturn longstanding assumptions about motivated reasoning and identify a boundary condition on Naïve Realism and the Bias Blind Spot. People's tendency to be aware and proud of their biases provides both new opportunities, and new challenges, for resolving ideological conflict and improving reasoning.

1. Introduction

Consider the propositions "God exists" or "humans are responsible for climate change." When deciding whether to believe these propositions, and indeed any proposition about matters of fact, one approach is to impartially consider the evidence. Another approach is to think about how one would benefit (or not) from believing. For instance, would belief in God help one feel good or behave ethically? Decades of research show that people's beliefs are a function of both kinds of considerations. People listen to evidence, but they are also "biased" in the sense that their judgments of what would be valuable or risky to believe affect their beliefs. However, popular models of belief formation assume that people do not want to be biased and are unaware that they ever are. In this paper we challenge this latter claim. Specifically, we show that people treat the moral value of holding a belief as a legitimate concern when deciding what to believe, and that as a result, people are sometimes aware that moral considerations have affected their current beliefs.

The finding that people are sometimes aware of and condone their biases has important theoretical and practical implications. First, many popular models of reasoning assume that biases only affect belief unconsciously (e.g., Kunda, 1990; Kruglanski, 1996; Pyszczynski & Greenberg, 1987). The possibility that biases affect belief consciously raises new questions about the processes that enable and constrain biased belief formation. Additionally, people's ignorance of their biases is often cited to explain why biases are common and why disagreements are difficult to resolve (Frantz, 2006; Kennedy & Pronin, 2008; Pronin, Gilovich, & Ross, 2004; Ross & Ward, 1996). However, the possibility that people are knowingly biased offers a complementary explanation for these phenomena. It also offers new opportunities and challenges for debiasing beliefs (Cusimano & Lombrozo, 2021a). We discuss these implications in the General Discussion. Below we review the claim that people are typically unaware of their biases, and then argue that morally motivated biases are a plausible and consequential exception.

1.1. The illusion of objectivity

An influential model of reasoning states that, even though people are biased, they try not to be (Kunda, 1990; Kruglanski, 1996; Pronin, 2007; Pronin et al., 2004; Pyszczynski & Greenberg, 1987). The starting assumption for this view is that people believe that "the only legitimate reasons for adopting a given conclusion as true are informational" (Kruglanski, 1996, p. 503). Thus, accepting or rejecting an idea based on its perceived value would "threaten [an individual's] conception of [themselves] as a reasonable, rational individual" (Pyszczynski & Greenberg, 1987; p. 302). So, at least consciously, people aim to hold beliefs that they could defend to "a dispassionate observer" and form beliefs "only if they can muster up the evidence necessary to support [them]" (Kunda, 1990; p. 482–483). In other words: People believe it is

* Corresponding author.
E-mail address: corey.cusimano@yale.edu (C. Cusimano).

https://doi.org/10.1016/j.cognition.2023.105379
Received 2 March 2022; Received in revised form 27 December 2022; Accepted 16 January 2023
Available online 13 February 2023

bad to be biased, and so aim to be impartial and to base their beliefs on evidence.

If people aim to be impartial, then they should eliminate biases when they detect them (Pronin et al., 2004). After correcting for these biases, they should think that their beliefs "follow from a relatively dispassionate, unbiased, and essentially 'unmediated' apprehension of the information or evidence at hand" (Ross & Ward, 1996, p. 110). As a result, the only biases that should continue to exert an influence on their beliefs are those that they failed to detect. This line of reasoning further entails a constraint (sometimes called a "reality constraint") on reasoning according to which people do not believe something unless they think that they have been impartial and the belief is backed by evidence (Baumeister & Newman, 1994; Festinger, 1957; Kunda, 1990; Kruglanski, 1996; Pyszczynski & Greenberg, 1987).

Many studies vindicate this characterization of reasoning. People often assume that anyone who possesses the same evidence that they do will hold the same beliefs and that anyone who disagrees with them is biased or uninformed (Reeder, Pryor, Wohl, & Griswell, 2005; Robinson, Keltner, Ward, & Ross, 1995; Rogers, Moore, & Norton, 2017; Ross & Ward, 1996). This self-attributed impartiality persists even when people are demonstrably biased. For instance, people deny that they are influenced by prevalent biases such as cognitive dissonance or wishful thinking (Pronin, Lin, & Ross, 2002; West, Meserve, & Stanovich, 2012). And even when people recognize their potential for bias, such as when they acknowledge their past biases, they still deny bias in their current beliefs (Ehrlinger, Gilovich, & Ross, 2005; Hansen, Gerbasi, Todorov, Kruse, & Pronin, 2014). Thus, people appear to operate under an "illusion of objectivity" (Kunda, 1990; Pyszczynski & Greenberg, 1987) and to possess a "bias blind spot" (Pronin et al., 2004). We refer to this view of metacognition, according to which people think they are impartial, evidence-based reasoners, even when they aren't, as the "Objectivity Illusion."

The Objectivity Illusion is often treated as a general (and so, nearly universal) description of conscious reasoning (Ehrlinger et al., 2005; Kruglanski, 1996; Pronin et al., 2004; Ross, 2018). Indeed, as noted above, an important claim of this view is that reasoning is constrained by a requirement to think that one's beliefs are unbiased. Accordingly, any exceptions – i.e., instances in which someone is aware that they are biased – should be rare and unsystematic. Two notable implications follow: First, some common wisdom about belief is incorrect. If people always think they believe on evidence, then no one ever adopts a belief "on faith." Second, an Objectivity Illusion recommends only one strategy for debiasing, namely, educating people about their biases. And it predicts that educating people about their biases should be an effective way to debias them, as people should always be motivated to be unbiased.[1]

Here we challenge the Objectivity Illusion. We argue that people sometimes believe that they ought to be biased. And because people sometimes believe that they ought to be biased, they sometimes recognize and condone biases in their current beliefs. Reasoning is not (always) constrained by a requirement to view oneself as impartial.

1.2. Are people sometimes "Biased and Proud"?

The Objectivity Illusion assumes that people only consider impartial, evidence-based reasoning to be good reasoning. This assumption is wrong. Studies investigating how people evaluate others' beliefs demonstrate that people also consider how helpful and morally good beliefs are – which are *partial* and *non-evidential* qualities of belief – when judging whether others are justified to hold them (Cusimano & Lombrozo, 2021a, 2021b; Metz, Weisberg, & Weisberg, 2018; Tenney, Logg, & Moore, 2015). Among the different non-evidential qualities of belief, people place special emphasis on moral qualities. For instance, people judge others' beliefs as unjustified when they are disloyal or disrespectful to others, even when those beliefs reflect impartial, evidence-based reasoning (Cusimano & Lombrozo, 2021b). Given the importance that people place on morality when evaluating belief, morally desirable biases are a promising place to look for exceptions to the Objectivity Illusion.

Indeed, in many other domains, moral considerations weigh heavily in people's reasoning, often overriding or resisting comparison to people's other goals (Baron & Spranca, 1997; Tetlock, Kristel, Beth, Green, & Lerner, 2000). People generally strive to view themselves as morally good (Blasi, 1980; Hardy & Carlo, 2005), and consider morality to be the most important dimension of a person's character (Goodwin, Piazza, & Rozin, 2014). So even if people want to be rational, they might deprioritize rationality when it conflicts with being respectful, loyal, or protective of others. Likewise, if people evaluate their reasoning based on what they think others demand of them, then because others are likely to evaluate their beliefs against biased criteria, they should hold themselves to standards of belief that incorporate those biases (Tetlock, 2002). If people license moral biases in their reasoning, or indeed actively want to be biased when they reason, then contrary to the Objectivity Illusion, they may not view all their beliefs as bias-free.

But even though moral standards strongly affect people's preferences and behavior, there are good reasons to doubt that people apply moral standards to their own *factual beliefs*.[2] First, as noted above, when it comes to forming beliefs about matters of fact, the Objectivity Illusion remains the dominant view. Second, some prior work has shown that people evaluate their own and others' beliefs differently. For instance, people tend to think that others have voluntary control over their beliefs (Cusimano & Goodwin, 2019) while denying that they themselves do (Cusimano & Goodwin, 2020). Additionally, people are more likely to express outrage at others' politically incorrect beliefs than their own (Cao, Kleiman-Weiner, & Banaji, 2019). And third, feeling convinced of one's objectivity brings certain benefits, such as guarding beliefs from criticism and making people more persuasive (Schwardmann & van der Weele, 2019; von Hippel & Trivers, 2011). So, despite evidence that people value moral biases in others' beliefs, they may not think that they themselves should be (or ever are) morally biased.

It is an open question how people evaluate their own reasoning in domains where they associate beliefs with moral value (and so are likely morally biased). If people remain unaware of their biases, or reject their biases as unjustified, such results would vindicate the Objectivity Illusion as a near-universal description of conscious reasoning. On the other hand, if people recognize and condone their biases, then this would constitute an exception to the Objectivity Illusion with implications for metacognition, critical thinking, and conflict resolution.

1.3. The current studies

Two lines of research yield conflicting predictions about the standards that people hold themselves to when forming beliefs, and accordingly, whether people are ever aware of (or endorse) their own

[1] Of course, people may not always be able to correct their beliefs. For instance, people sometimes have intuitions that on reflection they think are irrational but that they cannot dismiss (Risen, 2016; Walco & Risen, 2017). These irrational intuitions likely constitute one kind of systematic exception to the Objectivity Illusion, as they are beliefs that people hold but acknowledge as irrational. The Objectivity Illusion predicts that people should be motivated to correct these beliefs (and if they cannot correct them, then judge them negatively). Based on the studies we report below, people may want to maintain irrational intuitions when they seem to have moral benefits.

[2] Note that we are not concerned with *moral beliefs* (such as "murder is wrong"). We are concerned with *beliefs about matters of fact* that (may) have moral value. Also, we are not concerned with people's preferences to ignore moralized information when *making decisions* (e.g., Tetlock et al., 2000, on "forbidden base rates"). Instead, our studies investigate people's motivation to consider morality when forming beliefs.

Table 1
Predictions from the Objectivity Illusion and Biased and Proud models of reasoning and metacognition in Studies 1–4.

	Predictions		Observations	
Phenomenon	Objectivity illusion	Biased and proud	Observed	Study
Subjectively, moral value predicts belief and belief evaluation beyond evidence.	–	✓	✓	Studies 1a/1b
Self-attributions of moral bias are sensitive to the presence of moral bias.	–	✓	✓	Studies 2–4
Endorsement of moral biases.	–	✓	✓	Studies 2–4
Greater attributions of moral bias to others compared to self.	✓	–	–	Study 3

motivated reasoning. The "Objectivity Illusion" view predicts that people will judge their beliefs to be justified based on impartial, evidence-based reasoning even when they are morally biased. The alternative "Biased and Proud" view predicts that people consciously incorporate the moral quality of a belief into their self-directed belief evaluation. Accordingly, people should sometimes evaluate their beliefs based on their moral quality and approve of moral biases in their reasoning. We report four studies that test these competing predictions. Studies 1a/1b investigate whether people always think that their beliefs are backed up by evidence, or alternatively, whether people sometimes hold beliefs that they regard as morally good but not supported by evidence. Studies 2–4 build on these results in three ways. In these studies, we induce morally motivated reasoning and demonstrate that people's self-attributions of bias are sensitive to the presence of bias in their reasoning. Second, we demonstrate that, when people engage in morally motivated reasoning, the well-known "bias blind spot" disappears. And third, we demonstrate that people acknowledge and approve of one specific mechanism of biased reasoning – namely, biased hypothesis testing. Table 1 displays competing predictions from the Objectivity Illusion and Biased and Proud models and notes what we observed across each study.

1.3.1. Defining bias

Throughout this paper, we refer to belief formation as "biased" if it incorporates considerations – such as the costs or benefits of belief – that are unrelated to the accuracy of the belief (Cusimano & Lombrozo, 2021a). *Moral* biases concern the moral costs and benefits of adopting a belief. In our studies, moral concerns include whether the belief *promotes morally good behavior* (Study 1), *is respectful* (Studies 2–3), or is *risky* (Study 4). For example, someone would be biased if they held a belief because it was morally desirable and not because it reflected their evidence. Studies 1a and 1b show that people hold beliefs that they think lack evidence but are morally desirable. Additionally, someone might be morally biased if they hold beliefs to different evidentiary standards based on the moral risks of accepting or rejecting those beliefs (Cusimano & Lombrozo, 2021a). Such a person might be motivated to be accurate, and to believe based on evidence, but they would nevertheless be biased because they are more likely to accept beliefs that they consider morally safe relative to beliefs they consider morally risky. Studies 2 and 3 demonstrate "motivated skepticism" consistent with this kind of bias. Study 4 then provides direct evidence that people license this bias in their beliefs. Finally, our focus is purely descriptive: We take no stand on whether people ought to be biased or unbiased (as we define these terms). We will return to this point in the General Discussion.

1.3.2. Transparency and openness

For all studies, all sample sizes, exclusion criteria, and statistical analyses were preregistered. Experimental materials, data, analyses (annotated R scripts), and pre-registrations are available on ResearchBox: https://researchbox.org/150. An online supplement is available; Table A1 in Appendix A summarizes its contents. All studies were approved by the Offices of Research Ethics at Princeton University and Yale University.

2. Studies 1a – 1b

The Objectivity Illusion and Biased and Proud models of reasoning make different predictions about the metacognitive position that people take toward their morally valuable factual beliefs. The Objectivity Illusion predicts that introspection concerning the basis for a given belief is dominated by thinking about one's evidence. Accordingly, when people think about what they believe, how justified they are to believe it, and why they believe it, they focus solely on the evidence that they call to mind. Of course, people may notice that some beliefs have benefits that others do not, but after accounting for how they think about their evidence, these value judgments should not affect whether they accept a given belief or how justified they take themselves to be in that belief. In other words, the Objectivity Illusion predicts that people should never say of one of their beliefs that they lack evidence for it.

Our alternative "Biased and Proud" model is not so strict. On this view, when people assess whether they should believe something, or how justified their beliefs are, they think it is appropriate to consider the potential benefits of believing it separately from the evidence they have. Accordingly, when people evaluate their beliefs, they might draw on both their subjective assessment of the evidence, and separately, the benefits they associate with belief. Over time, these moral judgments may potentiate belief, which would result in people holding some beliefs that they associate with moral value but not evidence. This process would entail that, if we examine a large set of someone's beliefs, their judgments about the moral quality of those beliefs should incrementally predict how they evaluate them over and above how much evidence they think they have. In other words, people should sometimes hold beliefs despite thinking that they lack evidence for those beliefs, especially when those beliefs are associated with moral value.

We test these opposing predictions in Studies 1a and 1b. To this end, we measured participants' introspective reports of confidence, belief, and justification, and then tested whether these judgments were predicted solely by self-assessed evidence or were also predicted by the perceived moral value of the belief. As a secondary goal, these studies tested whether participants' evaluations of their beliefs were incrementally predicted by how pragmatically (so, non-morally) desirable they rated those beliefs to be. Consistent with prior work finding that people routinely deny that they ever engage in mere wishful thinking, we expected that participants would judge merely pragmatic (non-moral) desirability to be an illegitimate influence on their belief. We therefore predicted that, consistent with the Objectivity Illusion, these qualities of belief would not incrementally predict their subjective belief evaluations.

One challenge we faced was reliably measuring people's introspective judgments. For instance, we were concerned that any one way of measuring perceived evidence for a belief might fail to fully account for how people appraise the quality and quantity of information that informs their attitudes. Additionally, we worried that any association we observed between judgments of evidence, moral and pragmatic belief value, and other metacognitive judgments, might be particular to whatever beliefs we happened to measure, rather than more general features of people's introspective belief evaluation. To address these concerns, we conducted two studies, 1a and 1b, that measure the same

metacognitive judgments but with different measures and on different sets of beliefs. Allaying our concerns, the two studies replicated each other. Given large overlap in procedures and findings, we report Studies 1a and 1b together.

2.1. Methods

2.1.1. Participants

In Study 1a, we recruited 122 adults (61% reported male, 39% reported female, mean age 38 years) from Amazon Mechanical Turk (MTurk). An additional 25 participants were excluded for failing an attention check.[3] In Study 1b we recruited 225 adults from Prolific (58% reported female, 38% reported male, 4% reported intersex or preferred not to disclose, mean age 36 years) using the same recruitment criteria. No participants or data points were removed prior to data analysis. For both studies, participation was restricted to users with a US-based IP address and a 95% approval rating based on at least 1000 prior tasks.

2.1.2. Beliefs

Appendix B contains the full text of each proposition we examined in Studies 1a and 1b. In Study 1a, we used nine topics designed to produce wide variation in perceived moral and pragmatic value both within and between participants. Three topics concerned propositions that are often (but not always) associated with high moral and pragmatic value, including (i) whether God exists, (ii) whether people have free will, and (iii) whether people ultimately get what they deserve (i.e., Karma is real). Three topics concerned politicized propositions, including (iv) whether genetically modified foods are safe to eat, (v) whether immigration is good for the United States' economy, and (vi) whether the climate is warming due to human activity. And lastly, three topics concerned propositions without any obvious moral, political, or pragmatic value, including (vii) whether black holes exist, (viii) whether there are more than 35 million different species in tropical rainforests, and (ix) whether social media (like Facebook) is bad for people's mental health. It was not important whether participants categorized these propositions as moralized, politicized, or pragmatic, but only that these propositions generated variation in perceived moral and pragmatic value. Participants in Study 1a saw all nine topics (in a randomized order) and responded to every dependent measure, described below, for each.

In Study 1b, we expanded the set of topics. Several beliefs used in Study 1b were drawn from 1a, including (i) whether God exists, (ii) whether Karma is real, (iii) whether GMFs are safe to eat, and (iv) whether people have free will. The remaining eleven beliefs were new, including (v) whether ghosts/spirits exist, (vi) men tend to score higher than women on standardized math tests, (vii) women score higher than men in most leadership skills, (viii) heaven is real, (ix) the participant will avoid getting a serious illness (like cancer) in their lifetime ("I will avoid getting a serious illness…"), (x) scientists will discover the cure for cancer in the next 10 years, (xi) there is still time to significantly reduce the effects of global climate change, (xii) the participant has an implicit bias against minorities ("I have an implicit bias against minorities"), (xiii) animals, like pigs and cows, feel emotions just like humans do, (xiv) on January 6th, Trump conspired to overturn the election, and (xv) police in the United States tend to be biased against black people. Participants in Study 1b saw five of these fifteen topics (selected randomly and presented in a randomized order) and responded to every dependent measure, described below, for each.

[3] For one of the nine beliefs, selected at random, a free response attention check appeared on the page after participants reported their agreement with the non-evidential value questions described in the methods. To pass the check, participants had to describe the belief that they had just been making judgments about. Responses to this attention check were coded by the first author prior to data analysis. We did not include an attention check in Study 1b.

2.1.3. Procedure

For both Studies 1a and 1b, our primary outcomes were (i) certainty toward the proposition, (ii) endorsement of the proposition, and (iii) evaluations of how justified one is to believe the proposition. We also measured metacognitive judgments of (iv) perceived evidence for the proposition, (v) moral value of believing the proposition, and (vi) pragmatic value of believing the proposition. We measured these judgments in the order described below.

2.1.3.1. Certainty. Participants first reported their subjective confidence in the proposition. In Study 1a, they were asked, "How certain are you of the following claim?", read the text of the proposition (e.g., "God exists", and responded using an 11-point scale with anchors at 0% (Certain it is false), 50% (Completely uncertain), and 100% (Certain it is true). In Study 1b, we asked participants to rate how confident they were in the proposition, and to respond using a 7-point rating scale (1: extremely confident this is false, 7: extremely confident this is true).

2.1.3.2. Evidence. On the next page, participants were asked to consider their evidence for the proposition and to indicate the confidence that their evidence warranted. In Study 1a, we employed a thought experiment based on Kunda (1990)'s supposition that people form beliefs based on what they can defend to an impartial, dispassionate observer:

> Imagine that you were speaking with someone who was 100% perfectly open-minded and willing to carefully listen to and trust you. You have as much time as you want to share all of the evidence and arguments you have in favor and against believing that [God exists]. This person will form a belief based on what you provided them. However, this person is also perfectly objective, logical and rational and will only form a belief that they consider to be perfectly justified. Based on the evidence you can provide to this person, what would they estimate is the probability that [God exists].

Participants responded using the same scale that we used to measure certainty.

Study 1b measured evidence in a different way to account for the possibility that people believe they possess "private" evidence that they cannot convincingly share with others. To this end, we asked participants how much evidence they have and told them that evidence could include "personal experiences that you've had," "testimony that you have heard from sources that you consider to be reliable (like eyewitnesses, experts, or the news)" and/or "high-quality scientific studies that you've learned about." They then rated their agreement with four statements designed to get them to think about the reasons they have for and against the belief, including "I have evidence that [God exists]", "I have evidence that [God does not exist]", "I am aware of rational arguments about why [God exists]", and "I am aware of rational arguments about why [God does not exist]." Then participants responded to our primary measure of evidence, which asked them "Overall, when I think about things rationally and objectively, the evidence and arguments that I have suggest that it is" and then selected an option from a scale that ranged from 1 ("Extremely likely that [God does not exist]") to 4 ("Equally likely that [God does not exist] and [God does exist]") to 7 ("Extremely likely that [God exists]").

2.1.3.3. Belief and knowledge endorsement. On the next page, participants reported whether they endorsed the proposition. To measure belief endorsement, participants were asked, "Overall, what is your stance about whether [God exists]?" and could answer by reporting that they held a confirmatory belief (e.g., "I believe that [God exists]"), a disconfirmatory belief (e.g., "I believe that [God does not exist]"), or that they are withholding belief (e.g., "I am undecided about whether [God exists]"). Studies 1a and 1b measured belief endorsement the same way. Study 1a also asked participants whether they have knowledge on the topic, using the same scale.

2.1.3.4. Justification. Participants then reported how justified they felt in their belief. Participants who had just endorsed a belief in the proposition were asked to what extent they agree or disagree with the statement, "I am justified to believe that [God exists]." Participants reported their agreement on a 7-point rating scale anchored at 1 ("completely disagree") and 7 ("completely agree"). Studies 1a and 1b used the same measure of justification.[4]

2.1.3.5. Moral and pragmatic value. In Study 1a, participants were told to think about whether "there would be benefits to believing that [God exists] even if it was not true" and then reported their agreement with four statements. Two statements described possible moral benefits, including (i) "Even if it weren't true, it would make me a more useful and helpful person to society by believing that [God exists]" and (ii) "Even if it weren't true, believing that [God exists] would make me a better friend and family member." These benefits are distinctly moral because they accrue benefits to others, rather than to oneself. By contrast, two other items describe effects of belief that make the belief desirable to the believer, but otherwise lack a clear moral justification, including (iii) "Even if it weren't true, life would be easier for me if I believed that [God exists]" and (iv) "Even if it weren't true, people would like me more if I believed that [God exists]." Participants reported their agreement on a 7-point rating scale anchored at 1 ("completely disagree") and 7 ("completely agree").

Study 1b measured the same four judgments but changed the question format. Participants now judged whether the belief would have moral or pragmatic benefits assuming they had no evidence one way or the other. And, the response scales were converted to be bimodal. For instance, one of the moral value questions now read, "Assume you had no evidence one way or the other. What effect would believing that [God exists] have on how useful and helpful you are to society?", and participants provided a response on a scale from 1 ("much less useful/helpful") to 7 ("much more useful/helpful").

2.1.4. Reasons for belief

Study 1b included one measure that Study 1a did not include in any form. As the last question for each proposition, participants who had reported believing the proposition were given the following prompt, "For each of the reasons below, indicate how much that reason - from 0% to 100% - explains why you believe that [God exists]." Below this prompt were six items, each with a scale that ranged from 0% (not at all explains) to 100% (completely explains). Participants' responses to the scales had to sum to 100. The first item always cited evidence ("I have evidence for this belief"). The last item was always a catch-all for other or unknown reasons ("Other reasons"). The middle four items stated each of the four value judgments measured above as potential reasons, including the two moral value items ("Believing this makes me a better person" and "Believing this makes me a more useful and helpful person to others"), and the two pragmatic value items ("Believing this makes others like me" and "Believing this makes my life easier"). These four items were displayed in a randomized order.

At the end of the survey, participants reported their sex and age.

2.2. Results

We examined whether perceived moral value or pragmatic value incrementally predicted each of our primary outcomes – which included (i) one's certainty or confidence in the proposition, (ii) one's endorsement of the proposition, and (iii) how justified one feels believing the proposition, all after accounting for self-attributions of evidence. Because of the large number of measures, and because results were the same across measures and study, we will provide only a summary of the results here. Detailed reporting can be found in Supplemental Materials #1 and #2.

Perceived evidence for belief strongly predicted every metacognitive judgment in both Study 1a and 1b, no matter whether moral value or pragmatic value was included in the model (*ps* $< .001$). However, even after accounting for self-assessments of evidence, the perceived moral value of belief incrementally predicted certainty (Study 1a: $b = 2.80$, $SE = 0.35$, $df = 756.27$, $t = 7.96$, $p < .001$), confidence (Study 1b: $b = 0.21$, $SE = 0.03$, $t = 7.61$, $p < .001$), belief endorsement (Study 1a: $b = 0.30$, $SE = 0.11$, $z = 2.76$, $p = .006$; Study 1b: $b = 0.61$, $SE = 0.15$, $z = 4.13$, $p < .001$), and in Study 1a, knowledge endorsement ($b = 0.49$, $SE = 0.18$, $z = 2.76$, $p = .006$; Study 1b did not measure knowledge endorsement). Among participants who reported believing the proposition, and controlling for perceived evidence, those who associated the proposition with moral value felt more justified in their belief relative to participants who did not (Study 1a: $b = 0.08$, $SE = 0.02$, $df = 138.28$, $t = 3.24$, $p = .002$; Study 1b: $b = 0.18$, $SE = 0.03$, $t = 5.14$, $p < .001$). In all cases, a model that included the perceived moral value of belief as a covariate did a substantially better job predicting belief and metacognitive judgment compared to evidence-only models (*ps* $< .001$).

By contrast, variation in the perceived pragmatic benefits of belief was a poor predictor of belief and metacognition. In Study 1, models that included pragmatic value did not substantially improve model fit over evidence-only models (*ps* $> .19$). In Study 1b, pragmatic value incrementally predicted certainty ($b = 0.12$, $SE = 0.03$, $t = 3.67$, $p < .001$), justification ($b = 0.09$, $SE = 0.04$, $t = 2.30$, $p = .021$), and belief endorsement ($b = 0.34$, $SE = 0.14$, $z = 2.4$, $p = .016$). However, this may have been because, despite our attempts to generate items that independently varied in moral and pragmatic desirability, ratings of moral and pragmatic desirability were highly correlated in this study ($r = .53$). In exploratory analyses, we found that, when included together in the same model (alongside perceived evidence), moral value continued to incrementally predict participants' judgments, but pragmatic value did not. Across all outcomes, once perceived moral value was included in the model, adding pragmatic value did not improve model fit. But adding moral value to a model that included pragmatic value both improved model fit (*ps* $< .001$) and eliminated the predictive power of pragmatic value.

Finally, we analyzed what participants in Study 1b cited as reasons for why they held the belief. To calculate the total proportion of their belief that they explained by appeal to moral reasons, we added together their responses to the two moral reason items. We then did the same for the two pragmatic reason items to calculate the total proportion of pragmatic reasons. On average, participants cited evidence as the strongest reason for their belief ($M = 59$, $SD = 35$). Moral reasons were the second most cited ($M = 16$, $SD = 21$), followed by the catch-all item "Other reasons" ($M = 13$, $SD = 25$), followed by pragmatic reasons ($M = 12$, $SD = 19$). Participants were more likely to cite the moral benefits of belief as reasons for their belief when they associated those beliefs with moral benefits, $r(520) = 0.28$, $p < .001$, 95% CI [0.19, 0.35]. Participants were overall less likely to cite the pragmatic benefits of belief than the moral benefits of belief as reasons, $t(521) = 3.13$, $p = .002$, 95% CI [1.35, 5.91]. However, they were also more likely to cite pragmatic reasons as reasons for why they held a belief when they associated the belief with pragmatic benefits, $r(520) = 0.22$, $p < .001$, 95% CI [0.13, 0.30].

2.2.1. Demonstrating a clear counterexample to the Objectivity Illusion

Recall that the Objectivity Illusion states that people should never say of one of their beliefs that they lack evidence for it. We conducted

[4] Participants who reported either that they believe the opposite, or that they are withholding belief, were told, "We now want you to think about how justified you would be if you did decide to believe that [God exists]" and to "Assume that you have not learned anything new or forgotten anything that you know." They then reported their agreement with the statement, "I would be justified to believe that [God exists]." This measure of justification showed the same results as our primary measure, but because it did not concern an actual belief participants held, only indirectly speaks to our hypothesis.

additional, exploratory analyses to directly test this claim. To this end, we split all trials across Studies 1a and 1b into those in which participants reported that their evidence favored the proposition in question (i.e., their evidence rating was above the midpoint of the scale), and those in which they reported that their evidence did not favor the proposition (i.e., all other trials). By and large, participants believed propositions when their (perceived) evidence favored them (90%) and disbelieved them when their evidence did not (17%). Moreover, consistent with the Objectivity Illusion, participants' beliefs aligned with evidence whether they found the belief pragmatically beneficial or not (Fig. 1A).

However, participants sometimes did report believing a proposition despite thinking that the belief was not favored by their evidence. Indeed, as we expected, participants held beliefs like this when the belief was associated with high moral value (Fig. 1B). Among the top quintile of morally desirable beliefs, when participants indicated that their evidence did not favor the proposition, they nevertheless believed it in 43% of trials. And in these trials, participants generally felt justified in their belief, with the average justification rating equaling 4.92 ($SD = 1.54$, *median* = 5), on a 7-point scale. Common beliefs that participants endorsed (despite thinking they lacked evidence) included, "animals experience suffering the same way that humans do," "people have free will," "God exists," and "Police are biased against minorities."[5]

2.3. Discussion

One of the central claims of the Objectivity Illusion view is that people treat the evidence that they think they have as the only relevant feature to their believing it. Accordingly, people should never say of a proposition that they believe it but that they lack evidence for it. Studies 1a and 1b supported this claim for beliefs that people associate with pragmatic benefits. In Studies 1a and 1b, variation in participants' judgments that a belief would be pragmatically desirable did not predict the likelihood that they would endorse the proposition or how justified they thought their belief was (after accounting for their subjective evaluations of their evidence). We found the same results in Study 1b once accounting for the perceived moral value of belief. And finally, participants rarely cited the pragmatic benefits of belief as reasons for holding the belief.

However, participants thought about the moral value of their beliefs in a way that challenges the Objectivity Illusion. Across Studies 1a and 1b, participants who thought that believing a proposition would provide them moral benefits were more likely to report a belief in the proposition, and more likely to feel justified in their belief, holding constant their perceived evidence. Indeed, when believing something was thought to be especially morally valuable, participants were most likely to believe it, and feel justified to do so, despite thinking that their evidence did not favor the belief. And in Study 1b, we found that participants were more likely to cite moral considerations as reasons for their belief when those beliefs conferred moral benefits. The relative strength of moral value, but not non-moral value, as an influence on metacognitive judgments is consistent with the observation that people believe that moral value, but not non-moral value, is a legitimate influence on belief (Cusimano & Lombrozo, 2021b). These results were robust to two different methods for measuring perceived evidence, to two different ways of measuring the perceived moral and non-moral value of belief, and across studies that asked participants to introspect about different sets of beliefs.

In summary, people do not appear to evaluate their beliefs based solely on the evidence that they think they have for them. And people do not think that all their beliefs are supported by an impartial evaluation of their evidence. Instead, people sometimes endorse beliefs, and feel justified doing so, when they derive moral benefits from the belief. In other words, across a range of commonplace beliefs, people appear to believe that they are morally biased and appear to condone their biased beliefs. However, Studies 1a and 1b do not demonstrate that people accurately appraise their reasoning. Indeed, participants may have been suffering from an ironic lack of insight by underestimating the extent to which their beliefs are unbiased. This is an important limitation: If psychologists want to leverage people's awareness of their biases, it is important to know whether people accurately diagnose their biases. Studies 2–4 address this limitation by inducing morally motivated reasoning and then showing that people self-attribute bias commensurate with the influence of bias in their judgment.

3. Study 2

In Study 2 we induced morally biased reasoning and then examined whether participants were aware of their bias and whether they approved of it. Here the key challenge we faced was finding a candidate proposition that was both morally and evidentially ambiguous from the perspective of our target population, and one that we could experimentally manipulate using real-world data. The proposition we identified was: "On average, Black people tip less than White people." The moral status of this belief is the subject of ongoing debate in moral philosophy (e.g., Basu, 2018, 2019; Gardiner, 2018), and this controversy inspired us to investigate it. But while this topic is morally controversial, it is not widely discussed. So, coming into the study, participants were unlikely to have much evidence on the matter. Thus, this belief is morally and evidentially ambiguous, and therefore an ideal target for our investigation.

We found two studies published in the 1990s – Lynn and Graves (1996) and Mok and Hansen (1999) – that investigated racial differences in tipping behavior and that form a naturally-occurring near-minimal pair. These two studies used highly similar methods: They recruited similar sample sizes and both recruited participants from restaurants in Houston, Texas. They used similar interview methods to gather data, used similar response scales, and investigated and reported on the same questions about what factors predict tipping. They also reported the same findings about what factors do and do not predict tipping behavior – with one important exception. Lynn and Graves (1996) found that on average ethnic minorities tip less than Whites, even after controlling for other variables; Mok and Hansen (1999) found no difference in tipping behavior between ethnic minorities and Whites. Thus, these two studies are nearly identical with respect to their scientific merit and conclusions but differ in that one reports a morally controversial finding and the other does not. In Study 2 we randomly assigned participants to read about one of these two studies and then measured their reactions.

In the face of undesirable new information, people exhibit "motivated skepticism" such that they more heavily scrutinize that information and raise the standard of evidence required to believe it (Ditto & Lopez, 1992; Edwards & Smith, 1996; Gilovich, 1991; Plunkett, Buchak, & Lombrozo, 2020). Given that the potential harm of falsely accepting Lynn and Graves' (1996) finding is greater than the potential harm of falsely accepting Mok and Hansen (1999) finding, we predicted that participants would judge Lynn and Graves (1996) more harshly, but *only* when they learned the results of the study (rather than just the methods of the study). This reaction would constitute morally motivated skepticism. However, we have found in prior work (on evaluations of others' beliefs) that people believe that motivated skepticism, when done for moral reasons, is justified (Cusimano & Lombrozo, 2021b). Thus, we

[5] A few readers have wondered if this finding represents a kind of *dumbfounding* (reminiscent of 'moral dumbfounding' in the moral judgment literature). It does not. Dumbfounding refers to a phenomenon in which people remain stubbornly committed to a claim despite self-consciously lacking any defensible reason for it (Royzman, Kim, & Leeman, 2015). Participants in our studies do not think that their beliefs lack any defensible reason. Even though participants think they lack evidential reasons to justify their belief, they also think that they possess moral reasons that justify their belief. At most people may be epistemically dumbfounded, but they are not overall dumbfounded.

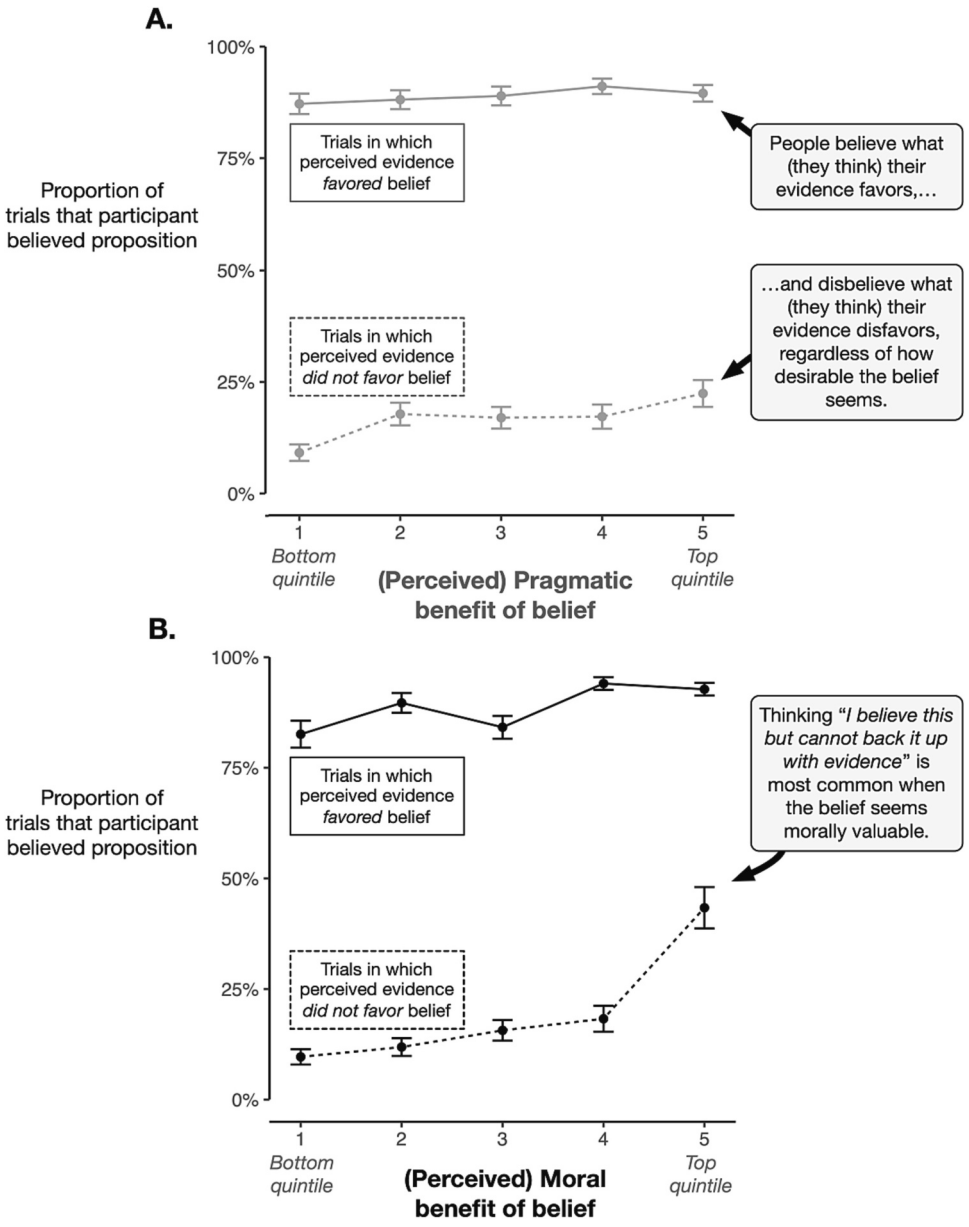

Fig. 1. Across Studies 1a/1b, the proportion of propositions believed (and standard error) grouped by perceived pragmatic benefit rating (top) and perceived moral benefit rating (bottom) quintiles. Belief rates are further divided into trials in which the participant provided subjective evidence ratings that favored the belief over its opposite (solid line) and trials in which the participant provided subjective evidence ratings that did not favor belief (dashed line).

further predicted that people would be aware of their morally motivated skepticism and that they would condone it.

When documenting motivated reasoning it is important to rule out information-based explanations for differences in people's beliefs. In the context of racial differences, people might be unbiased (in the sense of being consistent with Bayesian reasoning) to reject a study documenting racial differences if they come into the study with a lot of evidence that there are no racial differences (Baron & Jost, 2019; Jern, Chang, & Kemp, 2014; Koehler, 1993; Lord, Ross, & Lepper, 1979). We can rule out such information-based explanations in the current studies. A pilot study ($N = 200$) that we conducted on Prolific found that when participants give their "best guess" about the tipping practices of Black and White customers, they on average assume that Black people tip less than White people, $t(199) = -7.18$, $p < .001$, 95% CI [−1.08, −0.62]. We replicate this finding in Study 3, reported below, using a different method for probing prior belief. More importantly, participants whose prior guesses favor racial differences in tipping, and participants whose prior guesses favor *no* racial differences, are on average both highly *un*confident in their guess, and equally (un)confident in their guesses. Thus, based on purely information-based reasoning, participants should update their beliefs the same amount (or their beliefs should slightly favor evidence of racial differences) when presented with equally strong new evidence for or against racial differences (see Supplemental Materials #6). However, we predicted that participants would favor the morally safe belief (that there are no racial difference) over the morally risky one (that there are racial differences).

3.1. Methods

3.1.1. Participants

We recruited 1070 participants (48% Male, 50% Female, 1% unreported or other, mean age = 36 years) from Prolific. Participation was

restricted to users currently living in the United States and with a 95% approval rating based on at least 150 prior tasks. As preregistered, we excluded data from another 130 participants who incorrectly answered an attention check.

3.1.2. Design

Participants were randomly assigned to read a description of either Lynn and Graves (1996) or Mok and Hansen (1999). We will refer to the Lynn and Graves (1996) condition as "Race-Different," because this study reports racial differences in tipping, and the Mok and Hansen (1999) condition as "Race-Same." The descriptions that participants read comprised a thorough summary of the methods of the studies (see Appendix C for full text). We also manipulated whether participants then read about the questions the experimenters were trying to answer ("methods-and-goals" condition) or what results the experimenters reported ("methods-and-results" condition). For instance, in the methods-and-goals condition, participants read that the authors were interested in, among other things, studying whether "demographic factors (sex, age, and ethnicity)" predict tipping behavior. This description of the study goals did not differ across the race-same and race-different conditions. In the methods-and-results condition, participants read a short summary of what each study found. In the race-same condition, part of the results description included, "there was no variation across different demographic variables," while the race-different condition included, "White respondents left larger tips (even after controlling for the other variables) than did non-white respondents" (see Appendix C for full text). Participants were randomly assigned to one of these four conditions at the beginning of the study.

We also asked participants to self-attribute (and evaluate) two different kinds of moral bias. One form of moral bias reflected the kind that we predicted would affect participants' reasoning, namely, judgments that it is disrespectful to form beliefs about others based on their race (hereafter, "race-respect bias"). The second was a similar norm against forming beliefs about others based on their sex (hereafter, "sex-respect bias"), but one that was irrelevant with respect to what either study reported finding. Including this form of bias helps us rule out an alternative explanation for our anticipated results according to which people self-attribute apparently desirable moral biases merely because they identify them as desirable (rather than because those biases played any detectable role in their reasoning).

3.1.3. Procedure

Participants first read a brief introduction to the topic of tipping. This introduction noted that social scientists are interested in studying how people spend money, and that many have specifically investigated what factors predict how people tip. Participants then read a 230-word description of either the Race-Different or Race-Same study. This description provided details about the methods and procedures used in each study, including where the study was located, the number of participants recruited, and how the authors collected demographic data. Then, on the next page, participants read either about the goals or the results of the corresponding study. After reading about the study, participants (i) evaluated the quality of the study, (ii) reported whether they believed the results of the study, (iii) rated the moral acceptability of forming beliefs about tipping behavior based on different factors, and, (iv) self-attributed and self-evaluated bias (Fig. 2).

3.1.3.1. Study quality. After reading about the study, participants evaluated five dimensions of the study's quality. These included, (i) "this was a high-quality study," (ii) "the methods/procedures are balanced and leave little-to-no room for bias," (iii) "the authors have collected a sufficient sample size (i.e., number of customers) to make claims about people's tipping behavior," (iv) "the customers recruited for this study are representative of the rest of the population," and (v) "the location and type of restaurant provided is representative of restaurants in general." Participants rated each of these qualities on 7-point agreement scales (1: Strongly disagree; 7: Strongly agree), with higher values indicating judgments of higher quality. These items were presented in a random order. After three items, participants received an attention check that instructed them to select the lowest point on the scale.

3.1.3.2. Study acceptance. On the following page, participants rated their agreement with three statements about whether they do, or would, accept the results of the study. Specifically, participants reported whether they (i) "accept the results of the study" (ii) "believe the major claims the authors make based on these methods" and (iii) "believe these findings would replicate in future studies." Participants assigned to the goal information condition responded to similar statements lightly modified to ask whether they "would" accept whatever results the study found. As above, participants reported their agreement with 7-point rating scales (1: Strongly disagree; 7: Strongly agree). These statements were shown in a random order.

3.1.3.3. Moral judgments. Participants then reported what information they thought would be disrespectful to use when forming a belief about someone's tipping behavior. Participants rated their agreement that it would be disrespectful to incorporate information for seven attributes, including (i) race/ethnicity, (ii) sex/gender, (iii) wealth/social status, (iv) quality of food the person ate, (v) quality of service received, (vi) quality of restaurant they went to, and (vii) the location of the restaurant they went to. These attributes were shown in a random order. As noted above, we were most interested in participants' responses to 'race/ethnicity' and 'sex/gender'; the remaining attributes were distractor items. Participants used a 7-point rating scale with every point labelled between "Strongly disagree" (1) and "Strongly agree" (7).

3.1.3.4. Bias attribution and evaluation. Participants then self-attributed moral bias and judged the use of moral bias in their evaluation of the study. We were primarily interested in self-attribution of bias related to judging that it is disrespectful to use race/ethnicity. To measure self-attributions of race-respect bias, participants read the following prompt:

> You indicated that you "[race-respect response]" that it is disrespectful to use information about a person's race/ethnicity to form beliefs about their tipping habits. How much did this influence how you personally evaluated the study?

The text [race-respect response] was filled in with the response they provided in the *moral judgments* phase of the procedure above. So, if a participant reported that they "slightly agree" (corresponding to point-4 on the 7-point rating scale) that it is disrespectful to use information about race/ethnicity, then that participant would have read, "You indicated that you 'slightly agree' that it is disrespectful…" Participants attributed to themselves their degree of bias using a 7-point rating scale anchored at 1 ("Did not influence me at all") and 7 ("Influenced me a lot"). On the next page participants evaluated their bias after reading the prompt:

> You just (on the previous page) reported that thinking it is disrespectful (or not) to use information about race/ethnicity influenced you "[bias attribution]" on a scale from 1 (not at all) to 7 (a lot). Do you think that this influenced you the right amount, too much, or not enough?

The phrase [bias attribution] was replaced with the response they had just provided on the previous screen. In response to this question, participants used a 7-point rating scale anchored at −3 "Did not influence me enough," 0 "Influenced me the right amount," and + 3

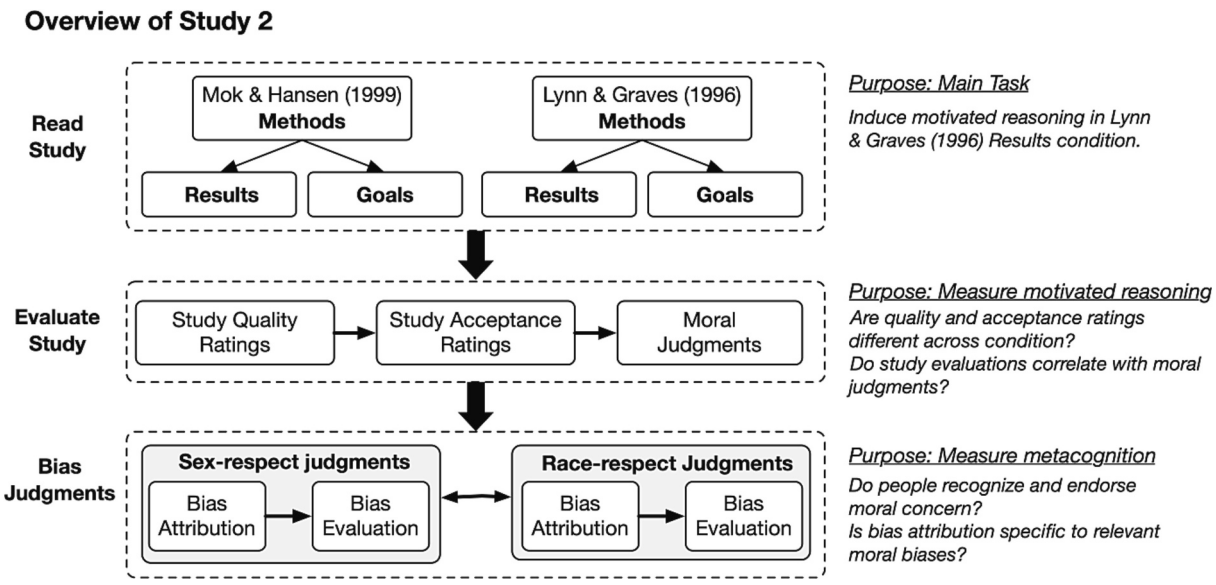

Fig. 2. Overview of Study 2 procedure. Arrows indicate the order in which steps occurred, with branches corresponding to randomly assigned between-participants conditions, and the bidirectional arrow indicating that task order was counterbalanced.

"Influenced me too much." Participants repeated the same procedure for sex/gender respect bias. Bias attribution and evaluation were always paired in the way described above. The order of race-respect bias and sex-respect bias was counterbalanced across participants.

Participants then reported their age and sex, filled out a life satisfaction survey (which we included as a distraction task) and a 10-item social desirability scale (Strahan & Gerbasi, 1972).

3.2. Results

We created composite measures of "study quality" (α = .89) and "study acceptance" (α = .93) from our five- and three-item scales, respectively. We then submitted these composite study quality scores and composite study acceptance scores to 2 (Study: Race-Different vs Race-Same) × 2 (Study Information: Methods-and-goals vs Methods-and-results) ANOVAs. Looking at composite study quality ratings, we did not observe a reliable main effect of information condition, $F(1, 1066) = 3.56$, $p = .059$, or study condition, $F(1, 1066) = 0.12$, $p = .732$. However, we observed the predicted study × information interaction, $F(1, 1066) = 4.61$, $p = .032$. Examining this interaction in more detail, participants evaluated the race-same study nearly identically in the methods-and-goals ($M = 3.63$, $SD = 1.37$) and methods-and-results ($M = 3.65$, $SD = 1.33$) conditions, $t(540.6) = -0.17$, $p = .867$, 95% CI [−0.25, 0.21]. However, participants who read about the race-different study evaluated that study more harshly in the methods-and-results condition ($M = 3.49$, $SD = 1.45$) compared to participants in the methods-and-goals condition ($M = 3.84$, $SD = 1.42$), $t(523.88) = 2.77$, $p = .006$, 95% CI [0.10, 0.59]. Analyzing study acceptance ratings revealed similar evidence for motivated reasoning (see Supplemental Materials #3).

Further evidence for morally motivated reasoning comes from examining the relationship between people's moral judgments and their study evaluations. The one condition that presents disrespectful information is the results condition for the race-different study. And indeed, in this condition only, moral judgments about the use of racial information negatively correlated with both study quality, $r(258) = -0.24$, $p < .001$, 95% CI [−0.36, −0.13] and study acceptance, $r(258) = -0.34$, $p < .001$, 95% CI [−0.44, −0.22]. Taken together, the results demonstrate clear evidence of morally motivated reasoning. Additional analyses relating patterns of judgment to responses on a social desirability scale suggest that response bias cannot account for the observed effects.[6]

Participants in the race-different results condition exhibited some self-awareness of their morally motivated reasoning (Fig. 3a). We submitted self-attributions of race-respect bias to a 2 (Study) × 2 (Information condition) ANOVA. This analysis revealed main effects of information condition, $F(1, 1066) = 5.39$, $p = .02$, $\eta^2_G = 0.02$, and study, $F(1, 1066) = 15.02$, $p < .001$, $\eta^2_G = 0.01$. However, and as predicted, these effects were qualified by a significant information × study interaction, $F(1, 1066) = 8.95$, $p = .003$, $\eta^2_G = 0.01$. Participants who read about the race-different study self-attributed more race-respect bias when they learned the results of the study ($M = 4.24$, $SD = 1.94$) compared to participants in the methods-and-goals condition ($M = 3.59$, $SD = 2.02$), $t(524.92) = -3.78$, $p < .001$, 95% CI [−0.99, −0.31]. Likewise, participants who read the results of the race-different study self-attributed more race-respect bias to themselves compared to participants who read the results of the race-same study ($M = 3.40$, $SD = 2.00$), $t(528.94) = 4.89$, $p < .001$, 95% CI [0.50, 1.17]. Indeed, examining the means in Fig. 3b shows that participants uniquely self-attributed race-respect bias at higher levels in the race-different results condition. This was the only condition in which participants demonstrated morally motivated reasoning. Importantly, we only observed this pattern of bias attribution for self-attributions of race-respect bias. When we analyzed self-attributions of sex-respect bias, we observed no variation in bias across conditions (see Supplemental Materials #3).

Finally, we investigated whether the increase in self attributions of bias was associated with changes in self evaluations of bias. Across conditions, the majority (74%) of participants rated their bias as ideal (i.e., by selecting "0" on the −3 to +3 scale). Importantly, evaluations of

[6] In the race-different methods-and-result condition, social desirability weakly correlated positively with overall study quality, $r(258) = 0.12$, $p = .05$, 95% CI [0, 0.24], and not at all with study acceptance, $r(258) = 0.05$, $p = .41$, 95% CI [−0.07, 0.17]. And, social desirability did not correlate with judgments that it is disrespectful to use racial information, $r(1,068) = 0.01$, $p = .77$, 95% CI [−0.05, 0.07]. We discuss response bias in more detail in Supplemental Materials #8.

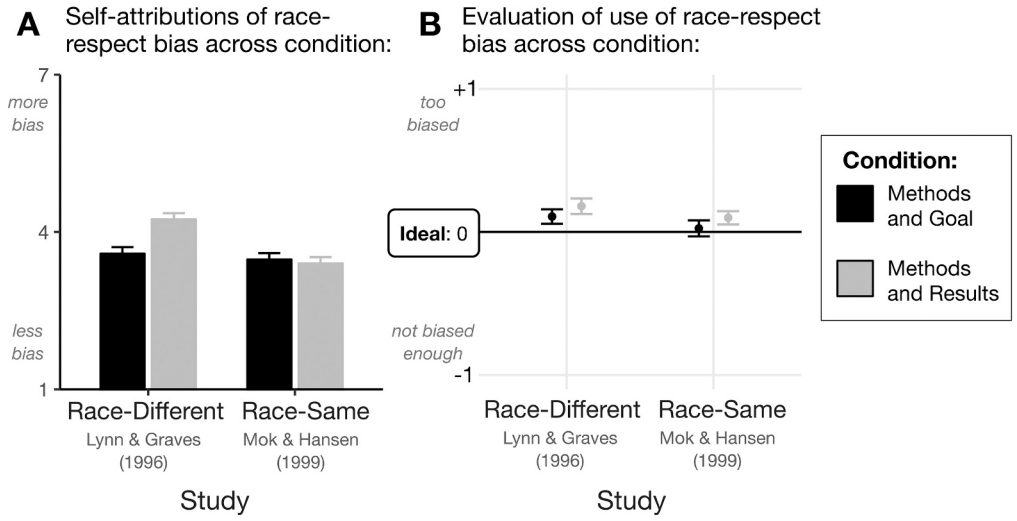

Fig. 3. Self-evaluation of bias in Study 2. (A) Participants in the race-different methods-and-results condition were uniquely likely to self-attribute race-respect bias (figure displays means and standard errors). (B) However, participants in all conditions rated their reasoning as ideal (figure displays mean and standard errors, scale ranged from −3 to +3).

race-respect bias did not vary across conditions when we submitted responses to a 2 (Study) × 2 (Information condition) ANOVA ($ps > .16$; Fig. 3b). It therefore appears that participants in the race-different methods-and-results condition attributed to themselves more bias while still judging this increased bias as ideal. To formally test this claim, we grouped participants' bias attribution and bias evaluation judgments into a new dependent variable, "bias judgment," and created an independent variable, "bias judgment kind," that indicated whether the judgment was an "attribution" or "evaluation." We then submitted "bias judgments" to a 2 (Study) × 2 (Information condition) × 2 (Bias judgment kind) mixed within-between ANOVA. This analysis revealed the predicted 3-way interaction, $F(1, 1066) = 10.76$, $p = .001$: When participants were self-attributing bias, we observe the study × information condition interaction whereby participants self-attributed bias in the race-different results condition but no other condition. By contrast, when participants were evaluating their bias, they evaluated it as ideal in all conditions.

3.3. Discussion

Participants in Study 2 engaged in morally motivated reasoning. On average, they judged the quality of a scientific study more negatively, and were less accepting of the study, when that study provided evidence for a morally disrespectful result compared to a morally respectful one. According to the Objectivity Illusion, participants should deny that they engaged in morally motivated reasoning and instead insist that their judgments regarding how disrespectful certain beliefs are played no role in their reasoning. But this is not what participants did: They recognized that they engaged in morally motivated reasoning, and moreover, judged their moral bias as justified. These self-attributions of morally motivated reasoning were sensitive to the actual presence of morally motivated reasoning, as participants only self-attributed bias in the one condition for which moral judgments affected their reasoning. And, participants only self-attributed biases that were consistent with their motivated reasoning (namely moral biases related to race). This latter observation, when considered with the fact that socially desirable responding did not correlate with any of our measures, speaks against the possibility that participants self-attributed biases merely because they judged those biases to be desirable.

4. Study 3

Researchers have operationalized the "bias blind spot" as a tendency to attribute more bias to others than to oneself (Ehrlinger et al., 2005; Pronin et al., 2002; West et al., 2012). In past studies, higher attributions of bias to others are interpreted as accurate recognition of others' biases, while lower attributions of bias to the self are interpreted as inaccurate unawareness of one's own biases. Based on the findings from Study 2, we predicted that this self-other difference would attenuate or reverse for moral biases that affect people's reasoning. Thus, insofar as self-other differences in bias attribution constitute evidence for unawareness of bias, Study 3 provides evidence against unawareness of bias. As a secondary goal, we sought to replicate Study 2 using measures similar to those used in Studies 1a and 1b. Instead of asking participants to evaluate the quality of a scientific study, we directly asked participants to report their beliefs about differences in tipping behavior between races.

4.1. Methods

4.1.1. Participants

We recruited 700 people (48% reported male, 51% reported female, 1% other or unreported, mean age in years = 37 years) from Prolific. This sample size yielded 95% power to detect an effect $d \geq 0.30$, which we observed in a pilot study.

4.1.2. Design

Similar to Study 2, participants were randomly assigned either to read a short summary of Lynn and Graves (1996), which reports racial differences in tipping, or of Mok and Hansen (1999), which reports no demographic differences. These summaries included descriptions of the methods of each paper as well as their results (see Appendix D). Participants assigned to read about Lynn and Graves reported whether they accepted the study's conclusion that "Black people tip less than White people on average" (race-different condition) and participants assigned to Mok and Hansen reported whether they accepted the study's conclusion that "Black people and White people on average tip the same amount" (race-same condition).

Within each study condition, participants attributed bias both to themselves ("self" condition) and to other participants in the study ("other" condition) (Bias target manipulation). Additionally,

participants attributed two forms of bias: moral bias, in this case exhibiting respect toward minorities ("respect" condition), and non-moral bias (i.e., "wishful thinking" condition; non-moral condition) (bias type condition). Fully crossed, this yielded a 2 (study) × 2 (bias target) × 2 (bias type) mixed within-between design.

4.1.3. Procedure

Study 3 comprised five tasks: (1) make a prior guess about racial tipping differences, (2) read one of two studies documenting evidence for either a morally acceptable or unacceptable proposition, (3) evaluate one's belief about tipping behavior in light of exposure to the study, (4) attribute bias to self and others, and (5) evaluate one's own reasoning.

4.1.3.1. Prior guess. At the beginning of the study, participants were told that marketers and scientists are interested in how people spend money and how this differs (or not) across demographics. Participants then provided their best guess about how Black and White tipping compared in the 1990s. Specifically, participants were told, "Please provide your best guess for how much more or less – on average – Black people in the United States tipped compared to White people in the United States in the 1990s. For instance, imagine that White people tip on average 15% of their bill. How much, on average, do you think Black people tipped, relative to this 15% anchor?" Participants responded using an 11-point scale that ranged from −10% to +10% in steps of 2%. Below this, participants reported how confident they felt in their guess using a 7-point rating scale anchored at 1 ("Not at all confident") and 7 ("Extremely confident"). After providing their best guess, participants learned that social scientists have studied this question and that they were about to read about one study that has been published on this topic.

4.1.3.2. Read and evaluate scientific study. After reading the summary of either Lynn and Graves (1996) (race-different condition) or Mok and Hansen (1999) (race-same condition), participants reported whether they accepted the conclusions of the study. We used the same procedure for measuring belief from Study 1. Participants could select an affirmatory belief that matched the conclusions of the study (e.g., "Yes I believe that, when this paper was published, [Black people in the United States on average tipped less than White people in the Houston, TX],"), endorse believing the opposite claim, or indicate that they are withholding belief. Participants who endorsed the claim matching the conclusions of the study were designated "believers"; participants who selected one of the other options were designated "non-believers." We provided a text box below this question where participants could optionally provide context for their response if they wished. Using the same method from Studies 1a and 2, participants reported whether they had knowledge of tipping behavior and rated how justified it is to accept the conclusion of the study. Believers rated how justified they were to hold their belief (actual justification). Non-believers rated how justified someone would be, in general, to believe (counterfactual justification). Both believers and non-believers responded to these prompts using the same 7-point rating scale anchored at 1 ("Not at all justified") and 7 ("Completely justified").

4.1.3.3. Bias attribution. On the next page, participants reported whether they thought that they or other participants in the study were biased. To this end, participants reported how much they agreed or disagreed with a series of four statements that varied according to a 2 (bias type) × 2 (bias attribution target) crossed design. One factor manipulated the type of bias: respect-bias or wishful thinking. For instance, the respect-bias statement about oneself read, "When I formed my belief about the tipping study, I was affected by how respectful or disrespectful I thought it would be to form beliefs about others based on their race." The wishful thinking item read, "When I formed my belief about the tipping study, I was affected by how desirable or undesirable I thought it would be for Black people and White people to tip the same amount." These two items were duplicated to be about "other people in this experiment." These four statements were presented in a random order for each participant.

4.1.3.4. Reasoning evaluation. After attributing bias, participants evaluated whether they were biased the right amount. As in Study 2, for both wishful thinking and respect bias, participants were reminded of the response they provided and then evaluated whether they "should have" been that biased. Participants responded to this prompt using a 7-point rating scale anchored at −3 ("I was affected less than I ought to have been"), 0 ("I was affected to the extent that I ought to have been"), and 3 ("I was affected more than I ought to have been"). We counterbalanced the order of wishful thinking and respect bias evaluation.

Lastly, participants reported their sex and age.

4.2. Results

4.2.1. Prior guess and motivated reasoning

Participants' prior guesses regarding tipping leaned toward the proposition that Black people tip less than White people on average ($M = -0.20$, $SD = 2.34$), $t(699) = -2.41$, $p = .016$, 95% CI [−0.39, −0.04]. Additionally, participants whose prior guesses favored racial differences, and participants whose prior guesses were that there were no racial differences, were equally confident in their prior guess. Given participants's prior guesses, we observed biased acceptance of the race-different and race-same studies: Participants were less likely to believe the results of Lynn and Graves's (race-different) study (53%) than they were Mok and Hanson's (race-same) study (70%), $b = -0.65$, $SE = 0.16$, $z = -4.06$, $p < .001$. Participants also felt less justified to believe the race-different result ($M = 4.62$, $SD = 1.44$ vs. $M = 5.13$, $SD = 1.38$), $t(695.5) = 4.86$, $p < .001$, 95% CI [0.31, 0.73]. However, knowledge self-attribution did not differ across the race-different (53%) and race-same (58%) conditions, $b = -0.22$, $SE = 0.15$, $z = -1.42$, $p = .156$. See Supplemental Materials #6 for a detailed analysis of motivated reasoning in this study.

4.2.2. Bias attribution

Attributions of bias were subjected to a 2 (Target: self vs other) × 2 (Bias kind: wishful thinking vs respect) repeated-measures ANOVA. Replicating prior work, participants on average attributed more bias to others ($M = 4.48$, $SD = 1.43$) compared to themselves ($M = 3.77$, $SD = 1.88$), $F(1, 699) = 171.87$, $p < .001$, $\eta^2_G = 0.04$. Averaged over themselves and others, participants attributed more respect bias ($M = 4.33$, $SD = 1.70$) than wishful thinking ($M = 3.92$, $SD = 1.69$), $F(1, 699) = 77.47$, $p < .001$, $\eta^2_G = 0.02$. However, as predicted, these main effects were qualified by a significant target × bias type interaction, $F(1, 699) = 11.96$, $p = .001$, $\eta^2_G < 0.01$, such that the difference in bias attribution between self and other was significantly smaller for respect bias compared to wishful thinking.

If bias attribution reflects actual bias, then self-attributions of bias should vary according to whether (1) the belief that participants formed elicited moral concerns and (2) whether participants formed a belief that was morally desirable. To test this, we submitted respect bias attributions to a 2 (Condition: race-same vs race-different) × 2 (Target: self vs other) × 2 (Believer status: believer vs non-believer) mixed-design ANOVA. As predicted, we observed a significant 3-way interaction of condition, target, and believer status, $F(1, 696) = 13.78$, $p < .001$. The best way to understand this interaction is to consider how attributions of race-respect bias vary as a function of target (self vs other) and believer status (believer vs non-believer) within the race-same condition and race-different condition. We report these follow-up analyses below.

In the race-same condition, we observed a main effect of target, such that on average participants attributed more respect bias to others ($M = 4.61$, $SD = 1.35$) than they did to themselves ($M = 4.05$, $SD = 1.90$), $F(1,$

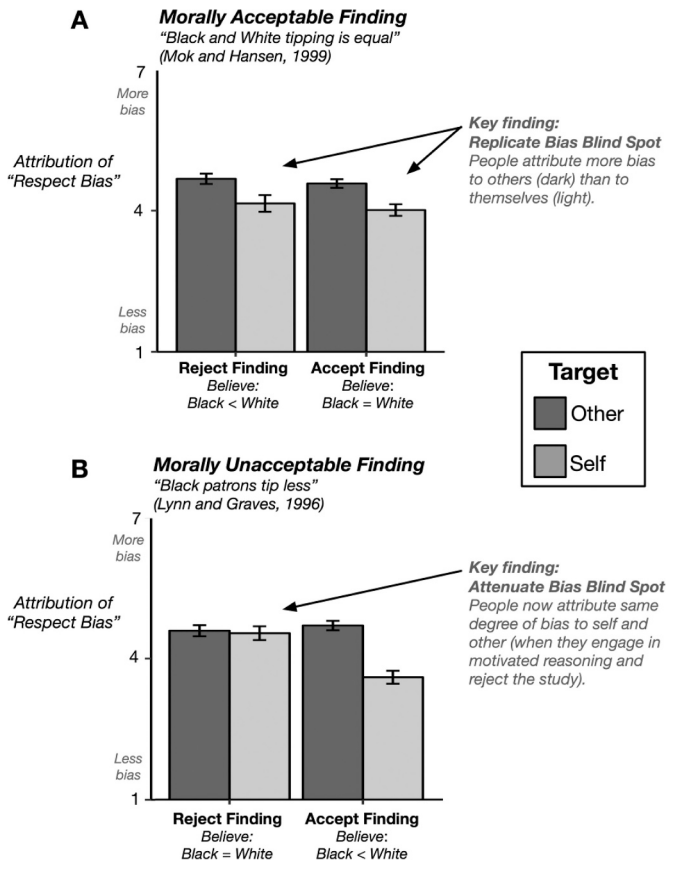

Fig. 4. Study 3 respect bias attribution to self and other (means and ± standard error of the mean) across moral acceptability condition and belief endorsement.

351) = 40.52, $p < .001$, $\eta^2_G = 0.03$. These attributions were not qualified by a believer × target interaction, $F(1, 351) = 0.05$, $p = .83$, nor did we observe differences between believers and non-believers, $F(1, 351) = 0.51$, $p = .475$ (Fig. 4A). This pattern of results perfectly replicates the "bias blind spot." We observed a different, but predicted, pattern of results in the race-different condition (Fig. 4B). Here we also observed a main effect of target, such that on average participants attributed more respect bias to others ($M = 4.65$, $SD = 1.43$) than they did to themselves ($M = 4.02$, $SD = 1.93$), $F(1, 345) = 43.48$, $p < .001$, $\eta^2_G = 0.03$. However, bias attributions were now qualified by a significant target × believer interaction, $F(1, 345) = 29.59$, $p < .001$, $\eta^2_G = 0.02$. Believers – those who accepted the race-different finding – attributed more respect bias to others ($M = 4.70$, $SD = 1.39$) than they did to themselves ($M = 3.60$, $SD = 1.92$), $t(191) = 8.15$, $p < .001$, 95% CI [0.83, 1.36]. However, non-believers attributed equal (and equally-high) respect bias to themselves ($M = 4.59$, $SD = 1.47$) and others ($M = 4.53$, $SD = 1.82$), $t(154) = 0.38$, $p = .701$, 95% CI [−0.21, 0.32]. In other words, non-believers in the race-different condition did not attribute bias to themselves and others in a way predicted by the bias blind spot.

Further examination of these results reveals that bias attribution was sensitive to the presence of bias. Among participants assigned to read about the study reporting racial differences (Fig. 4), non-believers attributed to themselves greater respect bias ($M = 4.54$, $SD = 1.82$) than believers did ($M = 3.60$, $SD = 1.92$), $t(336.01) = 4.65$, $p < .001$, 95% CI [0.54, 1.33] (Fig. 4B, compare the light gray bars). This difference in bias attribution is sensible because believers, who accepted the new evidence, by nature of their acceptance, are unlikely to have engaged in motivated skepticism. We can also compare participants who hold the same belief but, by virtue of being assigned to read about different studies, were exposed to different information. For instance, non-believers in the race-different condition and believers in the race-same condition both, at the end of the study, deny racial differences. However, the "non-believers" in the race-different condition read evidence against this rejection, and so needed to engage in motivated reasoning to hold that belief. By contrast, the "believers" in the race-same condition did not need to engage in motivated reasoning. And indeed, non-believers (who received evidence against their belief) self-attributed more bias compared to believers (who received evidence for their belief), $t(342.98) = 2.74$, $p = .007$, 95% CI [0.15, 0.90]. (For reference, the cells we are comparing here are represented by the light gray bar in the "reject belief" condition in Fig. 4B, and the light gray bar in the "accept belief" condition in Fig. 4A, respectively.)

4.2.3. Evaluations of reasoning

Finally, we turned to participants' self-evaluations of how biased they were (or were not). Overall, 60% of participants reported weighing respect bias "the right amount." If people who self-attributed more respect bias thought they were wrong to do so, then we would observe differences in bias evaluation across condition and believer status. To test this prediction, we submitted self-attributions of respect bias evaluation scores to a 2 (Study Condition) × 2 (Believer) ANOVA. We found that, despite differences in self-attributed respect bias across these conditions (Fig. 4, light gray bars), there were no significant effects nor interactions in this model ($ps > .266$). Replicating our analysis from Study 2, we grouped respect-bias self-attributions and respect-bias self-evaluations into a new dependent variable, "bias judgment," and created an independent variable "bias judgment kind" that indicated whether the judgment was an "attribution" or "evaluation." We then submitted "bias judgments" to a 2 (Condition) × 2 (Believer) × 2 (Bias judgment kind) mixed within-between ANOVA. We observed the same three-way interaction, $F(1, 696) = 8.53$, $p = .004$, such that participants varied in their self-attributions of respect bias, but they did not vary in their evaluations of respect-bias. Additional analyses revealed that the findings above, including the full attenuation of the bias blind spot, replicate within the subset of participants who rated their reasoning as ideal (i.e., participants who responded "0" to the bias evaluation questions) (see Supplemental Materials #4).

4.3. Discussion

Study 3 replicated the same morally motivated reasoning that we observed in Experiment 2 while using different methods to measure belief and expanding to belief-related metacognitive judgments. In this study, participants more-readily endorsed, and felt more justified believing, a study that reported no racial differences in tipping behavior compared to a study that reported racial differences. Our primary question, again, was how participants evaluated their morally motivated reasoning. Consistent with our expectations, participants who engaged in morally motivated reasoning – i.e., those who tempered their acceptance of evidence of racial differences in tipping – seemed largely aware that they did so. These participants self-attributed greater moral bias in their reasoning compared to two groups of participants who did not engage in morally motivated reasoning – i.e., those who received the same morally risky information but did not temper their belief, and those who received morally safe information (and so were not induced to engage in motivated reasoning). Finally, just as in Study 2, most participants endorsed their morally motivated reasoning as ideal. Thus, Study 3 shows additional evidence that people are biased, aware, and proud.

Study 3 also identified a novel boundary condition on the *bias blind spot*: Participants who engaged in morally motivated reasoning attributed similar amounts of moral bias to themselves and others. This finding did not reflect a general tendency among our participants to self-attribute bias. Replicating prior work on the bias blind spot, Study 3 found across all conditions that participants attributed less wishful thinking to themselves than they did to others. Additionally, participants attributed more moral bias to others (replicating the bias blind

spot for moral bias) in all conditions except for the one condition in which participants themselves engaged in morally motivated reasoning. Thus, insofar as the bias blind spot constitutes evidence for the Objectivity Illusion, Study 3 demonstrates evidence against the Objectivity Illusion.[7]

5. Study 4

One aim of Study 4 was to replicate findings from Studies 2 and 3 with other propositions, thereby helping to establish the generality of our results. The propositions that we investigated concerned the impact of providing gender affirming care[8] to adolescents with gender dysphoria. A second aim of Study 4 was to test whether people condone their tendency to engage in a specific form of morally biased reasoning wherein they consider the harms of making a wrong judgment when deciding whether they have sufficient evidence for belief (i.e., *biased hypothesis testing*; Trope & Liberman, 1996, or *value-based evidential reasoning*; Cusimano & Lombrozo, 2021a).[9] Consistent with past work, we expected people's acceptance of new information to depend on how risky they regarded believing that information to be. And consistent with the current investigation so far, we expected people to be aware of, and to condone, this bias.

When forming a belief about gender affirming care for adolescents, there are two opposing errors to worry about. One error would be to incorrectly reject the claim that gender affirming care helps (a Type 2 error). This error risks (among other things) harming youth who would benefit from early care but not receive it. The opposing error would be to incorrectly accept this claim (a Type 1 error). This error risks (among other things) harming youth who may get affirming care but not need it (or might otherwise regret it). We hypothesized that people's moral concerns regarding which harms are more important to prevent would predict their acceptance of new information about gender affirming care. For instance, some people are especially worried about the risk of incorrectly concluding that affirming care is helpful because they think that it is a bigger tragedy for someone to get gender affirming care and regret it compared to the tragedy of failing to get care that one needs. These individuals should be especially skeptical toward – i.e., biased against – new evidence that gender affirming care is helpful. By contrast, people who rank the tragedy of these outcomes in the opposite way should be biased against accepting new evidence that gender affirming care is not helpful. We tested these predictions by exposing participants to new information about gender affirming care and then measuring the impact of their concerns on their acceptance of that information.

According to the Objectivity Illusion, people do not realize the role that value-laden error management plays in their belief formation. Instead, people should think that they view the world "directly" in a way "unmediated" by their cares and concerns (e.g., Ross & Ward, 1996). And if people did realize that their concerns were pivotal to their belief, then they should judge their beliefs as unjustified. We suspected otherwise. We expected that people would recognize the influence of their values on their reasoning, and in particular, recognize (and condone) the value-laden way that error management tempers their acceptance of new information.

Pilot testing confirmed a few important facts about how our target sample of adult Prolific users thinks about gender affirming care in adolescents. First, their primary concern about providing adolescents with gender affirming care is its potential harms and benefits. In one pilot (N = 101), both liberals (M = 4.66, SD = 0.54) and conservatives (M = 4.11, SD = 1.03) rated this concern highly (on a 5-point scale), and both liberals (79%) and conservatives (71%) selected "the potential harms and benefits" as the most important consideration from a list of common considerations. Second, most of our participants (around 65–70%) reported having no direct experience or knowledge about this topic. This means that if we present people from this population with new data about gender affirming care, their impression should be responsive to the new information (because they start off with little information on the topic) but selectively tempered by their concerns about making a harmful error in judgment.

5.1. Methods

5.1.1. Participants

We recruited 800 people (52% reported Male, 48% reported Female, <1% other/did not report; average age = 43 years) from Prolific. We limited recruitment to Prolific users who had completed at least 30 studies and no more than 10,000 and who had an approval rate of at least 90%. We also limited recruitment to users over the age of 30 to reduce the number of participants that we would later exclude for having prior knowledge on the topic.[10]

5.1.2. Design and stimuli

In the main section of the study, participants read about a published study investigating the impact of puberty suppressants on gender dysphoric adolescents in the Netherlands (De Vries, Steensma, Doreleijers, & Cohen-Kettenis, 2011). De Vries and colleagues' study used two outcome measures: (1) "psychological functioning" (measured via a depression scale and parent-report and self-report measures of attitude and behavior problems), and (2) "gender dysphoria" (measured via gender dysphoria and body image scales). The authors measured these outcomes in a sample of 70 adolescents before the teens started taking puberty suppressants and then measured them again two years later. Results for these two outcomes differed. Psychological functioning improved over time, providing superficial support for the claim that gender affirming care is overall good for gender dysphoric teens. But feelings of gender dysphoria did not change over time, providing superficial support against the claim that gender affirming care is helpful for teens.[11] At the start of the study, participants were randomly assigned to read one of these two outcomes: psychological functioning ("Improvement / Psychological Functioning" Condition) or gender dysphoria ("No Improvement / Gender Dysphoria" Condition). Full text of the stimuli for these two conditions is available in Appendix D.

[7] In Supplemental Materials #5 we report a conceptual replication of Study 3 (N = 500). We replicate differences in belief acceptance and justification, observe the same pattern of bias attribution between the self and others, and importantly, observe the same pattern of bias attribution based on condition and believer status.

[8] Gender affirming care refers to interventions designed to support and affirm an individual's gender identity when that identity does not correspond to the gender assigned at birth. In Study 4 we focus on people's beliefs about puberty blockers.

[9] This kind of bias may have been operative in Studies 2–3. That is, participants might have held the claim that there are racial differences to a stricter evidential standard because they found it relatively morally risky. One of the ways that Study 4 goes beyond Studies 2–3 is by directly measuring the concern participants place on certain judgment errors.

[10] A recent Pew Research Poll reported that 53% of people under 30 knew at least one person who had received or wanted to receive gender affirming care (Pew Research Center, 2021).

[11] We say "superficial" because of the lack of a control group in this study. For instance, the finding that gender dysphoria did not change seems like a frustrating result for advocates of gender affirming care, but it is possible that a control group would have shown worsening feelings over time as their puberty advanced. Likewise, the finding that psychological functioning improved over the two-year period seems like a helpful result for advocates of gender affirming care, but it is possible that 16-year-olds tend to function better than 14-year-olds. Nevertheless, describing these results as "superficially" supportive vs unsupportive is warranted given how our participants responded to them.

5.1.3. Procedure and dependent measures

At the start of the study, participants read an introduction to the topic of gender affirming care that familiarized them with common terms and summarized the reasons that advocates and opponents of such care typically have for their views. Participants then read that they were going to learn about a real study on the topic. They were further told that the study is one of only a few that has looked at the topic and that (in part for this reason) it is widely referenced in the scientific literature. Then participants completed the following tasks in order: (1) reported their prior experience with the topic, (2) reported relevant moral judgments, (3) read an overview of the methods of De Vries et al. (2011), (4) read about one of the two key dependent measures, either psychological functioning or gender dysphoria, based on condition, (5) reported whether they believed the outcome of the study, (6) reported what factors influenced their judgment of the study, and finally (7) reported whether they reasoned as they ought to. We will describe each these steps below.

5.1.3.1. Background and current beliefs. Participants reported whether each of four statements accurately described their background with gender affirming care (e.g., "I know several people who, as teenagers, received, or wanted to receive, gender affirming care"). Participants then reported their "best guess" about whether gender affirming care is overall helpful or harmful for teens with gender dysphoria. To do this, participants reported the expected outcome, after two years, for fourteen-year-olds with gender dysphoria who receive puberty suppressants compared to teens who do not receive puberty suppressants. They provided their "best guess" using a 7-point rating scale with the following values: "extremely better," "a lot better," "a little better," "no different," "a little worse," "a lot worse," and "extremely worse." They then reported how confident they felt in their guess (1: not at all confident, 5: extremely confident).

5.1.3.2. Moral concerns about errors in judgment. Participants next reported the importance they place on the Type 1 (false acceptance) and Type 2 (false rejection) errors discussed above. To measure sensitivity to false acceptance, participants reported their agreement with the statement, "Given how bad it would be for someone to get gender affirming care but then regret it, it would be risky to conclude that gender affirming care is overall helpful to teenagers with gender dysphoria." To measure sensitivity to false rejection, participants reported their agreement with the statement, "Given how bad it would be for someone to need gender affirming care but never get it, it would be risky to completely rule out the possibility that gender affirming care is overall helpful to teenagers with gender dysphoria." Participants used 7-point rating scales anchored at 1 ("completely disagree") and 7 ("completely agree"). As an indirect measure of the relative importance of these two kinds of errors, participants also filled in the blank from the following statement, "Needing gender affirming care but never getting it would be [blank] compared to getting gender affirming care but regretting it," using a 5-point scale ("much worse", "a little worse", "equally bad", "a little better", "much better").

5.1.3.3. Exposure to scientific study. Participants read about the study in two stages. First, participants read a general overview of the methods of the study. This overview included a short description of the sample population, description of the general methodology, and the main analytic strategy (namely, comparing responses within this single group across the two time points).

On the next screen, participants read about one of the two main outcomes – either gender dysphoria or psychological functioning – based on the condition they had been assigned. Participants first read about the methods that De Vries and colleagues used to measure either gender dysphoria or psychological functioning. This description included a list of the key dependent measures (e.g., "gender dysphoria scale" or "depression inventory") alongside short descriptions of the measure. Participants then read the results. For each measure, participants read what the authors found at Time 1 (before taking puberty blockers) and Time 2 (two years later), and whether there was a statistically significant difference between the two time points. At the bottom of the screen, participants read a quote from the authors of the study summarizing the finding. In the "Improvement / Psychological Functioning" condition, participants read, "Based on these results, the authors wrote, 'psychological functioning of adolescents diagnosed with gender identity dysphoria had improved in many respects after an average of nearly 2 years of [puberty suppressants]' (p. 2281)." In the "No Improvement / Gender Dysphoria" condition, participants read "Based on these results, the authors wrote, 'puberty suppression did not result in an amelioration of gender dysphoria' (p. 2281)."

5.1.3.4. Belief. Participants then indicated whether they believe the study's results. As in Study 3, participants could report whether they believe the study (e.g., "Yes, I believe that puberty suppression improves psychological functioning in teens with gender dysphoria."), whether they believe the opposite of what the study found (e.g., "No, I believe that puberty suppression does not improve psychological functioning in teens with gender dysphoria."), or whether they are withholding belief (e.g., "No, I am withholding belief on this matter for now."). We modified the text of the items across condition to match the finding.

5.1.3.5. Report influences on belief. On the next screen, participants were reminded of what the study found and reminded of the belief they just reported. We then explained that we were interested in what factors influenced their belief about the study, and that they should indicate what considerations affected their judgment by reporting their agreement with a series of statements. Two of the statements corresponded to moral concerns about making a Type 1 or Type 2 error. The first statement read, "In deciding whether to believe the conclusions of the study, I was influenced by a concern about how believing the wrong thing could hurt future teens who might get gender affirming care and regret it." The second statement read, "In deciding whether to believe the conclusions of the study, I was influenced by a concern about how believing the wrong thing could hurt teens who might need gender affirming care but never get it." Endorsement of these statements constituted self-attribution of moral bias in belief formation. Participants reported their agreement using a 7-point rating scale anchored at 1 ("completely disagree) and 7 ("completely agree").

We were worried that participants might erroneously agree with these statements in the absence of providing them the opportunity to cite or discuss what else might have influenced their judgment. To address this concern, this part of the study contained two additional features. First, the page included three filler items which comprised other factors that likely did and did not influence people's judgment. The complete set of items – the two target items above and the three filler items – were presented in a randomized order. Additionally, at the bottom of the page we provided a large textbox with instructions to optionally write any additional comments they have about the study. By providing these alternative ways to explain and justify their judgment, participants should only self-attribute a concern about the moral risk of making a particular error if they detected that concern in their reasoning.

5.1.3.6. Self-evaluate reasoning. As the final task in the study, participants evaluated whether the degree to which they exhibited concern about the moral risk of error was justified. Participants were reminded what response they provided to the bias attribution question, and then asked whether the moral concern weighed too little, or too much, in their judgment. Participants responded to this question using a 5-point rating scale anchored at −2 ("I weighed this concern too little"), 0 ("I weighed this concern the right amount"), and +2 ("I weighed this

concern too much"). On the same screen, participants responded to the same kind of question for the opposing moral concern. These two questions were presented in a randomized order.

5.1.3.7. Demographics and debriefing. At the end of the study, participants filled out a brief demographics form, and finally, were debriefed. As a part of our debriefing, we included a summary of the American Medical Association's stance on gender affirming care, a link to a full PDF of De Vries et al. (2011) that the participant could download, and references for two recent studies investigating the impact of gender affirming care on adolescents.

5.2. Results

5.2.1. Descriptive statistics

Participants mostly came into the study expecting that gender affirming care would be helpful to teens with gender dysphoria. The median and modal prior guess was that affirming care would make teens "a little better off" ("3" on the 7-point rating scale), while the mean guess was between making teens "a little better off" and "no different" ($M = 3.59$, $SD = 1.57$). This prior guess was mostly weakly held: 62% of participants reported no prior experience with or knowledge of the topic, and their average confidence in their prior guess was in the middle of the scale (*Mean* = 3.01, *Median* = 3, $SD = 1.09$).

Participants on average felt equally concerned about the risks of false acceptance ($M = 4.39$, $SD = 1.85$) and false rejection ($M = 4.31$, $SD = 1.81$) of the hypothesis that gender affirming care is helpful, $t(799) = 0.79$, $p = .428$. However, those who cared more about one of these risks cared less about the other, $r(798) = -0.42$, $p < .001$.[12] For the analyses below, we created a composite "moral judgment" predictor by subtracting participants' concern about false acceptance (i.e., how bad it would be to wrongly believe that care is helpful) from their concern about false rejection (i.e., how bad it would be to wrongly believe that care is not helpful). This yielded a *relative moral risk judgment* for each participant such that higher values indicated greater concern about false acceptance and lower values indicated greater concern about false rejection. Even though moral judgment correlated with prior belief about affirming care ($r = 0.49$), there was also clear separation between the two. Consider participants who provided the most common prior guess (that affirming care would make adolescents a little better off). These participants split 41%/59% in terms of caring more about false acceptance than false rejection, respectively. Given some independence of participants' prior beliefs, and moral concerns, we turned to whether their moral concerns predicted their acceptance of new information.

5.2.2. Evidence of morally biased reasoning

Moral judgments about the risks of accepting or rejecting a particular belief predicted acceptance of the belief, thereby demonstrating motivated reasoning. We regressed belief (1: yes, 0: no) on prior guess, condition, and moral judgment, the interaction of prior guess and condition, and the interaction of moral judgment and condition, using a logistic regression. The interaction of prior guess and condition accommodates the observation that participants may rationally discount a study's findings when those findings are incongruent with their prior beliefs. However, we also observed a significant moral judgment × condition interaction, $b = -0.44$, $SE = 0.07$, $z = -6.15$, $p < .001$ (Fig. 5), indicating that participants' relative weighting of the moral risks of error affected their beliefs differently depending on whether accepting the finding risked one kind of error or the other.

We next examined the role that participants' relative moral risk judgments played within each condition. Within each condition, we regressed belief (1: yes, 0: no) on participants' prior guesses and their moral judgments. In the Psychological Adjustment condition, participants who were especially concerned about wrongly believing that affirming care is helpful were *less* likely to accept the study's conclusion (after accounting for their prior belief on the matter), $b = -0.21$, $SE = 0.05$, $z = -3.85$, $p < .001$ (Fig. 5B). By contrast, participants with this moral concern profile in the Gender Dysphoria condition were *more* likely to believe the study's finding, $b = 0.24$, $SE = 0.05$, $z = 4.90$, $p < .001$ (Fig. 5A). These analyses provide evidence for morally baised reasoning. Indeed, these results go beyond Studies 2 and 3 by demonstrating how *opposing* moral biases lead people to accept or reject *opposing* claims.[13] Given evidence of morally biased hypothesis testing, we turn to whether people were aware of and approved of this bias.

5.2.3. Awareness of morally biased reasoning

If participants were aware that their moral values affected their reasoning, then they should cite them as reasons for their belief (or disbelief). And, they should do so in a way that matches the impact that the moral concern had on their belief. To test for this possibility, we regressed false acceptance concern, and separately false rejection concern, on condition, believer status (1: yes, 0: no), and the interaction of condition and believer status. As expected, self-attributions of moral bias on belief depended both on which finding the participant read (i.e., what risk accepting the finding would engender) and whether they believed the finding. We observed this significant interaction for both concern about the risk of false acceptance, $F(1, 796) = 57.73$, $p < .001$, $\eta^2_G = 0.07$, and concern about the risk of false rejection, $F(1, 796) = 150.74$, $p < .001$, $\eta^2_G = 0.16$ (Fig. 6).

Consider participants' concerns about mistakenly concluding that gender affirming care overall helps. Participants who particularly care about this error were more accepting of the finding that puberty suppression does not improve feelings of gender dysphoria, and they were less accepting of the finding that puberty suppressants help psychological functioning. We found that these participants seemed to be aware of the impact that this concern had on their judgment: Participants who accepted the null gender dysphoria finding cited this moral concern as a stronger influence than participants who did not accept this null finding, $t(370.67) = -4.70$, $p < .001$, 95% CI [−1.27, −0.52]. Likewise, participants who rejected the finding that psychological functioning improved cited this concern as a stronger influence on their reasoning than participants who accepted that finding, $t(363.69) = 5.94$, $p < .001$, 95% CI [0.75, 1.50].

Consistent with our predictions, we also observed an inverted pattern of self-attributions for the opposing moral concern. Recall that participants who were most worried about failing to believe accurately that affirming care is beneficial were more accepting of the finding that psychological functioning improved and less accepting of the finding that gender dysphoria did not improve. Consistent with the hypothesis that participants recognize the impact of these moral concerns in their reasoning, participants self-attributed these concerns more when they accepted (rather than rejected) the finding that psychological functioning improved, $t(378.36) = 7.86$, $p < .001$, 95% CI [1.05, 1.76].

[12] These judgments correlated sensibly with the item that directly asked participants to indicate which outcome (getting care and regretting it or needing care and not getting it) was worse. Participants who thought that getting care and regretting it was worse reported being more concerned about false acceptance, $r(796) = 0.57$, $p < .001$, while participants who thought that needing care and not getting it would be worse were more concerned about false rejection, $r(796) = -0.37$, $p < .001$.

[13] The results stated in this paragraph could have been written by focusing on participants with the opposing moral concern. Participants especially concerned with false rejection of gender affirming care were *more* likely to accept the Improvement / Psychological adjustment finding, but *less* likely to accept the Null / Gender dysphoria finding. It is easier to appreciate the impact of morality by looking at how it affects belief formation among participants who all share the same prior belief. We report a pre-registered analysis that does this in Appendix F.

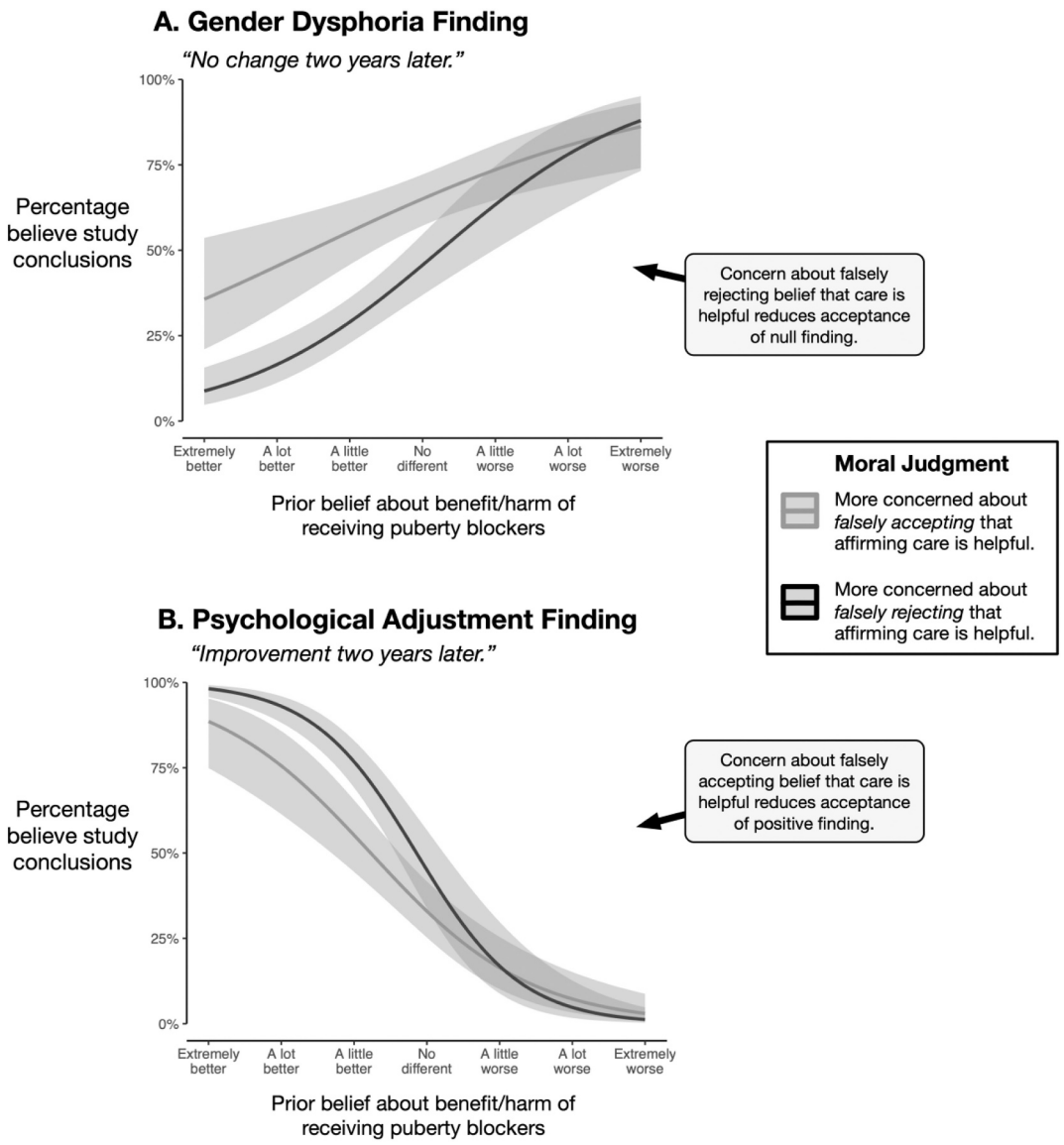

Fig. 5. In Study 4, the relationship between prior belief, relative concern about the risk of false acceptance vs false rejection that gender affirming care is helpful, and acceptance of the scientific study's findings. Trend line displays logistic fit, median split for visualization.

Likewise, they self-attributed this concern more when they rejected (rather than accepted) the finding that feelings of gender dysphoria did not improve, $t(393.91) = -9.54$, $p < .001$, 95% CI [−1.93, −1.27].

5.2.4. Evaluation of moral bias

What did participants think about the role that their moral concerns played in their acceptance or rejection of the study's conclusions? As expected, participants by-and-large judged their reasoning to be ideal. 76% reported that they weighed a concern about false acceptance "the right amount" in their judgment, and likewise, 78% reported that they weighed a concern about false rejection "the right amount." We expected there to be little variation in evaluations of one's biases because participants should evaluate their moral bias as ideal whether they weighed it a little or a lot. To test this, we regressed evaluation of moral bias on condition, belief (1: yes, 0: no), and the interaction of condition and belief. Our prediction was borne out for concerns about false acceptance ($ps > .20$). However, we observed a significant condition × belief interaction in participants' evaluation of their false rejection concern, $F(1, 796) = 10.26$, $p = .001$: Participants who were more likely to cite false rejection risks were slightly more likely to say that this concern affected them too much compared to participants who did not cite this concern. We followed up these analyses by testing whether, across condition and believer status, variation in endorsement of one's reasoning was greater than variation in self-evaluations of one's reasoning. As expected, we observed a significant 3-way interaction between believer status, condition, and judgment type (bias attribution vs bias evaluation) for both concern about false acceptance, $F(1, 796) = 59.12$, $p < .001$, and concern about false rejection, $F(1, 796) = 124.59$, $p < .001$. Accordingly, participants varied a lot in their sense that certain moral biases affected their reasoning, but they varied much less in their sense that they had reasoned as they ought to.[14]

5.2.5. Robustness/replicability

We replicated the findings above using several different preregistered exclusion criteria. One concern we had was that our single-item

[14] Following the procedure in Study 2, we also restricted our analyses to only participants who reported weighing the concerns of error in their judgment "the right amount." All of our findings replicate, demonstrating motivated reasoning and awareness of motivated reasoning among the subset of participants who report the influence of moral concern on their reasoning as ideal.

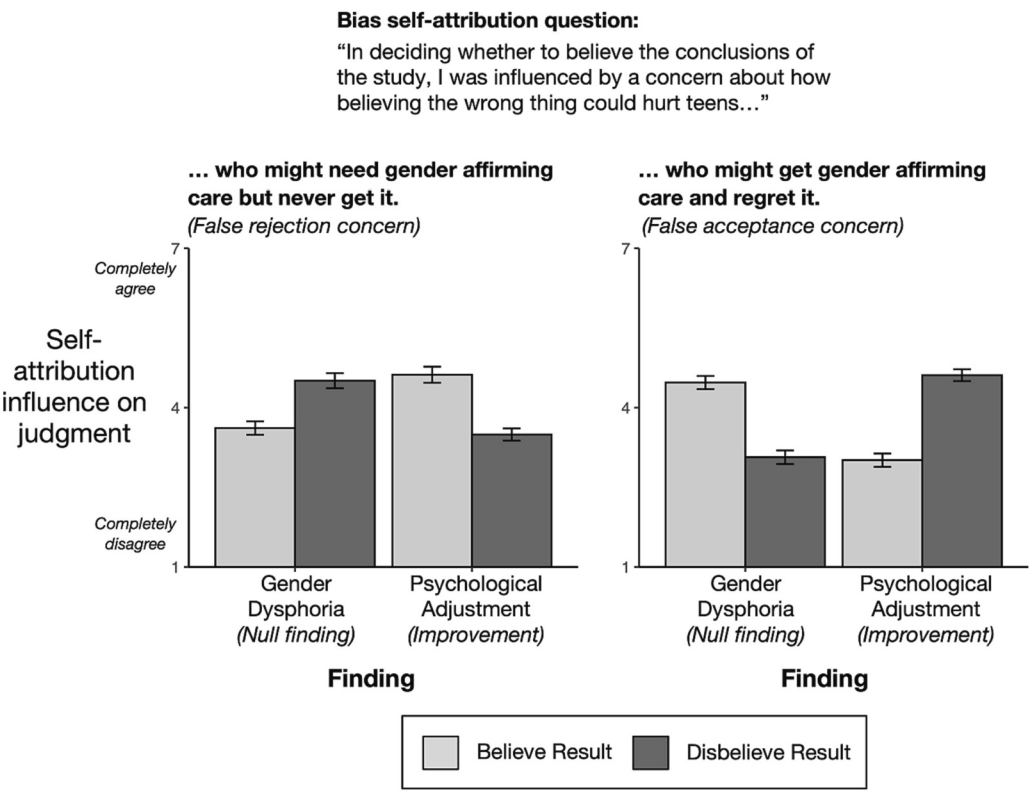

Fig. 6. For Study 4, the mean (and standard error) of moral concern attribution as an influence on one's judgment across concern type, condition, and believer status.

measure that records participants' prior beliefs about gender affirming care would not fully capture the nuance or complexity of their prior beliefs. In turn, our demonstration of morally motivated reasoning could be confounded if participants have some un-measured private information that licenses them to reject incongruent new information and that correlates with their moral judgments. To address this concern, we replicated our analyses after removing all participants who plausibly had prior information that would rationally license them to discount new information. We did this in two ways. First, we replicated our results among the subset of participants who self-reported being unconfident in their prior guesses. And second, we replicated our results among participants who reported having no experience with the topic (i.e., they answered "no" to all four background questions). Independent of this specific conern, we have also replicated these findings in two additional studies ($N = 600$; $N = 1000$; Supplemental Materials #7) that used different methods for measuring moral judgments and approval of morally motivated reasoning. In short: The results reported above are highly replicable and robust to different exclusion criteria and question wordings.

5.3. Discussion

Study 4 replicated our findings from Studies 2 and 3 in a new context while further illuminating one way in which people might feel "Biased and Proud." Our starting point was the observation that belief formation is sensitive to the relative risks of harm that people associate with different errors in judgment. Thus, one source of bias in everyday reasoning comes from people accepting or withholding belief based on whether doing so is risky in light of their goals and values. We documented a moralized version of this bias, wherein people's concerns for the well-being of adolescents who might regret care (or might never get it) tempered their acceptance of new evidence that such care is beneficial or not. We found that participants who were relatively more concerned about wrongly assuming that puberty suppressants were helpful were less accepting of new evidence that suggests such care is helpful (compared to participants were relatively more concerned about making the opposite error in judgment). These participants were also more accepting of (equally strong) new evidence that such care is not helpful (compared to participants relatively more concerned with making the opposite error in judgment). As in Studies 2 and 3, participants' biased acceptance of new evidence could not be fully explained by appeal to their prior beliefs.

Moreover, participants exhibited awareness and acceptance of their morally biased reasoning. Consider, for instance, participants who rejected the scientific study when it reported that gender affirming care improved psychological functioning. These participants tended to worry about the risk of harm that would befall teens who might get gender affirming care and regret it. This is consistent with the model of biased hypothesis testing, described above, wherein people seem to hold risky beliefs to stricter standards. And indeed, these participants were also likely to cite this moral concern as a reason why they rejected the study's finding. We found the same pattern of judgments, except for the opposing moral concern, for participants who rejected the study when it showed evidence that gender affirming care was not beneficial. Finally, participants by and large accepted their biased reasoning as ideal. Not only did most participants say that they weighed each moral concern "the right amount," but these findings replicated when we restricted our sample to only participants who thought they correctly weighed moral concerns.

6. General discussion

Our beliefs about our beliefs – including whether they are biased or justified – play a crucial role in guiding inquiry, shaping belief revision, and navigating disagreement. One line of research suggests that these judgments are almost universally characterized by an *illusion of objectivity* such that people consciously reason with the goal of being objective and basing their beliefs on evidence, and because of this, people

nearly always assume that their current beliefs meet those standards. Another line of work suggests that people sometimes think that values legitimately bear on whether someone is justified to hold a belief (Cusimano & Lombrozo, 2021b). These findings raise the possibility, consistent with some prior theoretical proposals (Cusimano & Lombrozo, 2021a; Tetlock, 2002), that people will knowingly violate norms of impartiality, or knowingly maintain beliefs that lack evidential support, when doing so advances what they consider to be morally laudable goals. Two predictions follow. First, people should evaluate their beliefs in part based on their perceived moral value. And second, in situations in which people engage in morally motivated reasoning, they should recognize that they have done so and should evaluate their morally motivated reasoning as appropriate. We document support for these predictions across four studies (Table 1).

In Studies 1a and 1b, participants reported their beliefs about a variety of topics and indicated how justified they felt in their beliefs. According to the Objectivity Illusion, people should only hold beliefs that they think impartially reflect their evidence. However, we found that participants' beliefs and belief evaluations reflected a combination of their appraisals of their evidence and their appraisals of how morally good the belief was to hold. For instance, participants' beliefs that God exists, and that animals experience emotions just like humans do, outstripped self-assessments of evidential strength and were partly explained by how morally good they thought those beliefs were. Indeed, participants were most likely to self-attribute an evidentially irrational belief – a belief that they thought lacked evidential support – when they associated the belief with moral value. As a result, in a substantive proportion of trials, participants reported believing something, lacking evidence for it, and feeling justified in their belief for moral reasons. At least for some common and moralized factual beliefs, people seem to think that their beliefs are morally biased and justified in part by their moral biases.

Studies 2–4 then demonstrated that people are some sometimes sensitive to the presence of moral bias in their reasoning. In Studies 2 and 3, participants read about a scientific study that offered evidence either that Black people tipped less than White people or that Black and White people tipped the same. Participants largely thought that the former proposition was morally risky to believe whereas the latter was not. And as expected, participants exhibited morally motivated skepticism by evaluating the study making the morally risky claim more harshly (Study 2), and endorsed the morally safe belief at a higher rate (Study 3). Contrary to the Objectivity Illusion, however, we found that participants who engaged in morally motivated reasoning self-attributed the moral bias that motivated their reasoning. Indeed, self-attributions of moral bias were sensitive to the presence of bias in reasoning: Participants who did not engage in morally motivated reasoning self-attributed bias at lower rates. And in Study 3, participants who engaged in morally motivated reasoning, and recognized that they did so, assumed that other people in the study were as biased as they were. This latter finding eliminates the well-documented "bias blind spot," which is a tendency for people to assume that others are more biased than they are. Finally, participants who recognized their moral biases thought that they were reasoning just as they ought to. Participants did not naively think they were objective, but instead felt a sense of justifiable self-respect about their biased reasoning – they were "Biased and Proud."

Study 4 replicated the recognition and endorsement of moral bias in the context of learning about the impact of gender affirming care on adolescents with gender dysphoria. Study 4 took advantage of the fact that, in this context, people vary in the relative concern they attach to a "false positive" (such as wrongly assuming that gender affirming care is helpful) and a "false negative" (such as wrongly dismissing the idea that gender affirming care is helpful). A common bias in reasoning is a tendency to be skeptical of new evidence that risks an error that one is especially concerned about. And indeed, when participants learned about a study on gender affirming care, whether they accepted the study's results depended on whether the results they read about recommended a belief they thought was risky to form. For instance, participants who were concerned about the harm of falsely concluding that puberty blockers are helpful were less likely to accept new evidence that puberty suppressants improve psychological adjustment. However, participants with this same profile of concern were more likely to accept (equally strong) evidence for the (subjectively less risky) claim that puberty blockers do not improve feelings of gender dysphoria. And just as in Studies 2 and 3, participants were aware, and thought it appropriate, that they tempered their beliefs in light of these concerns. Study 4 also expanded on Studies 2 and 3 by studying a context that elicited competing, and stereotypically liberal and conservative, moral concerns. Both sets of participants – those motivated to disregard evidence that puberty blockers are beneficial and those motivated to disregard evidence that they are not – were shown to be motivated reasoners, aware of their motivated reasoning, and accepting of their motivated reasoning. Being "Biased and Proud" does not appear to be uniquely liberal or conservative.

This work builds on the recent discovery that people sometimes positively evaluate bias in other people's beliefs (Cusimano & Lombrozo, 2021a; Tenney et al., 2015). But prior to the current investigation, there were serious reasons to doubt that positive evaluations of others' biases would extend to the self. For instance, even though people believe that others sometimes ought to hold overly optimistic beliefs (Tenney et al., 2015), it is plausible that people would never say of one of their own beliefs that it is overly optimistic. Indeed, one of the most important claims of the Objectivity Illusion is that reasoning is constrained such that people only hold beliefs that they can think are unbiased (Kruglanski, 1996; Kunda, 1990; Pyszczynski and Greenberg, 1987). Accordingly, if someone realized that their belief was overly optimistic, they would immediately correct it (Pronin et al., 2004). If motivated reasoning is constrained in this way, then the standards that people apply to others' beliefs that condone bias would be inert in their own reasoning. Thus, in extending the importance of moral evaluations of belief formation into people's own reasoning, the present studies challenge the notion of the Objectivity Illusion as a near-universal description of metacognition. This result entails that one of the most important implications of the Objectivity Illusion, namely that reasoning is constrained by people's need to think of themselves as unbiased, is wrong. Below we discuss some of the implications that follow for understanding disagreement and debiasing belief.

6.1. *Disagreement and belief-based conflict*

Our findings offer a new way of thinking about why disagreements between parties may be so difficult to resolve. Prior explanations appeal to a tendency for conflicting parties to erroneously judge themselves as objective and their opponents as biased. This explanation leads to the prediction that, if one side viewed the other as objective and themselves as biased, then the conflict would dissipate. Our results suggest otherwise. Consider for instance results from Study 4, and in particular the analysis we report in Appendix F. Here we observed two groups of people who shared the same prior belief but varied in their moral concerns about which errors in judgment were more important. As a result of their moral biases, these two groups come to different conclusions based on the same new information. But pointing out to these two groups that their disagreement reflects their different moral concerns would not change their mind or resolve their disagreement! They already know that their moral concerns influenced their judgment, they just think that they are right to be influenced in the way that they are. It is easy to see how this dynamic could play out in conflicts of even higher stakes: An atheist and a fundamentalist (who associates their belief with moral value) are unlikely to resolve their debate by both agreeing that there is little evidence for God. These kinds of disagreements may persist because people hold incompatible views regarding what beliefs are *morally* recommended. Accordingly, resolving conflicts between parties

may require coordinating not only on what is warranted on the basis of evidence, but also what is warranted on the basis of moral considerations.

6.2. Implications for interventions to reduce motivated reasoning

The observation that people sometimes regard their motivated reasoning as justified raises practical challenges for interventions aiming to limit motivated reasoning. Namely, insofar as people believe they are justified in being biased, they will dismiss and resist interventions that aim to debias them. At the same time, if people are sometimes biased because they think that they ought to be, then psychologists may be able to debias people by changing what norms people use to evaluate their beliefs (for instance, by affirming the value of impartiality). While psychologists have pursued many strategies for debiasing reasoning (e.g., educating people about biases they are unaware of, see, e.g., Baron, 2008; Milkman, Chugh, & Bazerman, 2009), we are not aware of any work that has attempted to debias others by affirming the value of impartiality. In light of our findings, one promising direction for future work is to test whether shifting people's relative valuation of impartiality and evidentialism changes beliefs (Cusimano & Lombrozo, 2021a).

That said, it is important to note that motivated reasoning – despite its connotations – is only a description of reasoning, not an evaluation of it. Motivated reasoning is only a bad form of reasoning in all cases if objective, evidence-based reasoning is an inviolable standard for belief formation. However, while there are some arguments that favor such a view, these arguments are controversial (Cusimano & Lombrozo, 2021a). For instance, there is an on-going debate about whether morality demands that people temper their judgment when making inferences about others on the basis of sex- or race-based statistics (Appiah, 1995; Basu, 2018; Bolinger, 2018; Moss, 2018). Thus, it may be that participants who tempered their beliefs about race-based differences in Studies 2 and 3 were reasoning as they ought to. Indeed, it would be consistent with the normative doctrine of *inductive risk* to hold beliefs that pose a risk to society to higher standards of evidence compared to non-risky beliefs (see Cusimano & Lombrozo, 2021a; Douglas, 2021, for discussion). Put simply, it does not follow from the observation that people engage in morally motivated reasoning that they are reasoning poorly. The normative status of motivated reasoning is important to consider because, unless people are unjustified when they engage in motivated reasoning, psychologists ought not intervene to "correct" that reasoning. We do not take our findings to bear on these normative questions one way or the other. Instead, we urge caution in claiming that participants are or are not making an error.

6.3. Claims that do not follow from the evidence we present

Before concluding, we want to draw attention to three claims that do not follow from the evidence that we have presented:

1) "*All beliefs on these topics reflect morally motivated reasoning.*" We have only shown that when being "impartial" and being "morally good" seem to conflict, enough people are prone to opt for the latter over the former that we can statistically detect their doing so. This finding does not entail that everyone who has formed a belief about, for instance, God, racial differences, or gender affirming care, was morally motivated when they did so.
2) *"People only/always recognize and condone their morally motivated reasoning."* We have only shown that people readily self-attribute bias and may be somewhat accurate when they do. It does not follow that people are always able to detect moral biases, nor that they are always biased when they say they are. We endorse a more modest view: People are sometimes ignorant of their biases, and they sometimes aren't. Future work should investigate what features of morally charged situations lead people to be aware of, or unaware of, their biases. Likewise, our studies do not demonstrate that moral biases are the only ones that people acknowledge and affirm. We investigated moral biases because they represented the best-case scenario for people being aware of their biases. Though we have some evidence against people acknowledging non-moral biases in their reasoning (see Studies 1a/1b and Study 3), there may be non-moral biases that we did not examine, but that people recognize and condone.
3) *"People can believe whatever they want."* Motivated reasoning is not constrained in the narrow sense that people can only adopt beliefs when they think they are unbiased. Additionally, people sometimes hold beliefs despite thinking that they cannot back those beliefs up with evidence. It does not follow from these findings that people can believe whatever they want. But these findings do raise the question of what cognitive processes constrain belief, given that they are not the simple ones assumed by prior theories. This is an important task for future research.

7. Conclusion

A great deal of work has assumed that people treat objectivity and evidence-based reasoning as cardinal norms governing their belief formation. This assumption has grown increasingly tenuous in light of recent work highlighting the importance of moral concerns in almost all facets of life. Consistent with this recent work, we find evidence that people's evaluations of the moral quality of a proposition predict their subjective confidence that it is true, their likelihood of claiming that they believe it and know it, and the extent to which they take their belief to be justified. Moreover, people exhibit metacognitive awareness of this fact and approve of morality's influence on their reasoning. People often want to be right, but they also want to be good – and they know it.

Author note

Corey Cusimano, School of Management, Yale University; Tania Lombrozo, Department of Psychology, Princeton University.

Materials, data, pre-registrations, and analyses can be accessed at ResearchBox link.

CRediT authorship contribution statement

Corey Cusimano: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Funding acquisition. **Tania Lombrozo:** Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Data availability

Data is available at https://researchbox.org/150

Acknowledgments

Jon Baron, Oleg Urminsky, and several anonymous reviewers provided especially valuable feedback on this work. This work also benefited from feedback provided by audience members at the Chicago Booth marketing seminar, Yale SOM marketing seminar, University of Kent social psychology seminar, Waterloo University brownbag seminar, and the Concepts and Cognition lab. We thank everyone at these events for their time and feedback.

Appendix A

Table A1
Index of supplementary materials (available from: https://researchbox.org/150).

Section	Pages
Supplement 1. Study 1a additional analyses	2–6
Supplement 2. Study 1b additional analyses	7–9
Supplement 3. Study 2 additional analyses	10–12
Supplement 4. Study 3 additional analyses	13
Supplement 5. Study 3 replication (Study S1)	14–23
Supplement 6. Prior confidence as an alternative explanation to morally motivated reasoning in Studies 2 and 3	24–27
Supplement 7. Study 4 Replications (Study S2 and Study S3)	28–33
Supplement 8. Response bias as an alternative explanation to findings from Studies 2–4	34–36

Appendix B. Study 1a/1b Stimuli

Study 1a propositions:

- God exists
- Genetically modified foods are safe to eat
- The universe makes sure that people ultimately get what they deserve
- People have free will
- Immigration is good for the United States economy
- The climate is warming due to human activity
- Black holes exist
- There are more than 35 million different species in tropical rainforests
- Social media (like Facebook) is bad for people's mental health

Study 1b propositions:

- God exists
- Genetically modified foods are safe to eat
- The universe makes sure that people ultimately get what they deserve
- People have free will
- Men tend to score higher than women on standardized math tests
- I will avoid getting a serious illness (like cancer) in my lifetime
- Women tend to be more neurotic (sad, moody, emotionally unstable), compared to men
- Heaven is real
- Scientists will discover the cure for cancer in the next 10 years
- There is still time to significantly reduce the effects of global climate change
- I have an implicit bias against minorities
- Animals (like pigs and cows) feel emotions just like humans do
- On Jan 6, Trump conspired to overturn the election
- Police in the United States tend to be biased against Black people

Appendix C. Study 2 stimuli

Race-different	Race-same
Here are the details of a study published in 1996 by Michael Lynn and Jeffrey Graves in the *Hospitality Research Journal.*	Here are the details of a study published in 1999 by Connie Mok and Sebastian Hansen in the *Journal of Restaurant & Foodservice Marketing*
Method for studying tipping behavior:	Method for studying tipping behavior:
The experimenters recruited interviewers from a hospitality service class and gave them all a questionnaire and standard set of interview questions to use.	One of the authors of the study (Hansen) approached all parties who dined at a restaurant for dinner over the course of 3 evenings. The restaurant was located in Houston, Texas, but the authors of the study did not report what specific restaurant they recruited participants from.
Interviewers stood outside the restaurants and intercepted customers who were leaving. They did this at two chain restaurants in Houston, Texas: Bennigan's and Olive Garden.	Customers were approached right after the bill was paid for and it was explained to them that the restaurant would like to obtain their feedback on their rating of service quality and to know more about them so that improvements could be made to better meet their needs.
Upon approaching the customers, the interviewers indicated that they were university students conducting a study for a class and asked the customers if they would answer several questions. Those dining parties that agreed to participate were asked to have the person(s) paying the bill answer the questions.	They had patrons self-report how much their meal cost and how much they tipped. They did nothing to rule out the possibility that customers could lie about how much they tipped.
They had customers report how satisfied they were with the food, and separately, how satisfied they were with the service. For instance, customer's rated food on portion size, taste, and price. And they rated service on appearance, knowledge, and friendliness. They used a rating scale from 1 "poor" to 5 "excellent".	They also had customers self-report how many people were in their group, their gender, age, income, and ethnicity.

(*continued on next page*)

(*continued*)

Race-different	Race-same
Customer's also reported their bill as well as how much they tipped. The authors of the study rejected data from people who reported a percentage instead of a dollar amount to weed out people who might have lied about how much they tipped. The interviewers then wrote down the customer's apparent sex, age, and ethnicity. They ended up with a total of 161 interviews for their analysis.	Lastly, they had customers report how satisfied they were with the food, and separately, how satisfied they were with the service. For instance, customer's rated food on portion size, taste, and price. And they rated service on appearance, knowledge, and friendliness. They used a rating scale from 1 "poor" to 5 "excellent". They ended up with a total of 112 interviews for their analysis.

Information condition	Race-different	Race-same
Goal	**Lynn and Graves** were interested in documenting how (i) customer service, (ii) food quality, and (iii) demographic factors (sex, age, and ethnicity) predict tipping behavior. On the next page, you will report how well suited you think the method of this study is to these goals. For instance, you will report what the study does well and what flaws you think it has.	**Mok and Hansen** were interested in documenting how (i) customer service, (ii) food quality, and (iii) demographic factors (sex, age, and ethnicity) predict tipping behavior. On the next page, you will report how well suited you think the method of this study is to these goals. For instance, you will report what the study does well and what flaws you think it has.
Results	Lynn and Graves made four claims based on the results of their study: • Customers left larger tips for more expensive meals. • Food quality did not predict tipping. • Customers who received better service left bigger tips. • White respondents left larger tips (even after controlling for the other variables) than did non-white respondents. Thus, "ethnic minorities leave smaller tips than do nonethnic customers" (p. 6). On the next page, you will report how well suited you think the method of this study is to providing evidence for these claims. For instance, you will report what the study does well and what flaws you think it has.	Mok and Hansen made four claims based on the results of their study: • Customers left larger tips for more expensive meals. • Food quality did not predict tipping. • Customers who received better service left bigger tips. • There was no variation across different demographic variables. On the next page, you will report how well suited you think the method of this study is to providing evidence for these claims. For instance, you will report what the study does well and what flaws you think it has.

Appendix D. Study 3 stimuli

Race-different	Race-same
In 1999, the researchers Connie Mok and Sebastian Hansen published a paper titled, "**A Study of Factors Affecting Tip Size in Restaurants**". We provide an accurate summary below. These researchers had students interview restaurant customers in Houston, Texas, as they were departing the restaurant. They collected information from customers about the sizes of their bills and tips, and recorded the customers' apparent ethnicities. This is a common and widely accepted technique for gathering data in this field. They found that better service was associated with larger tips. However, they also found that White and Black customers tipped, on average, similar percentages of their bill. That is, after controlling for other factors like the quality of service received, there were no racial differences in tipping. Even though these researchers recruited over a hundred people, this study has been criticized by others for having a relatively small sample size. Other scientists have also wondered to what extent this finding generalizes outside the location where these researchers conducted their study. This research topic is controversial. However, this study has not been retracted since its publication. This study has also been cited by many others who have built upon their findings and conclusions. By these standards, this study has been accepted within the scientific community. Screenshot of study:	In 1996, the researchers Michael Lynn and Jeffrey Graves published a paper titled, "**Tipping: An incentive/reward for service**". We provide an accurate summary below. These researchers had restaurant servers in Houston, Texas, record information about their customers. They collected information from servers about their customers' bill sizes, tips, and apparent ethnicities. This is a common and widely accepted technique for gathering data in this field. They found that better service was associated with larger tips. However, they also found that Black customers tipped, on average, a smaller percentage of their bill compared to White customers. That is, even after controlling for other factors like the quality of service received, they found racial differences in tipping. Even though these researchers recruited over a hundred people, this study has been criticized by others for having a relatively small sample size. Other scientists have also wondered to what extent this finding generalizes outside the location where these researchers conducted their study. This research topic is controversial. However, this study has not been retracted since its publication. This study has also been cited by many others who have built upon their findings and conclusions. By these standards, this study has been accepted within the scientific community. Screenshot of study:

Appendix E. Study 4 stimuli

All participants saw the following:
Here are basic facts about the how the scientists conducted their study:
Population:

- Sample size = 70 teens (ages 12–17) in the Netherlands. These were the first 70 eligible candidates for gender affirming medical care between the years 2000 and 2008.

Method:

- Measure teen mental health at two times:
  - Time 1: When the adolescent first attended the gender identity clinic, before the start of puberty suppression therapy. (Average age = 14 years old.)
  - Time 2: Two years after receiving suppression therapy, right before starting additional treatment. (Average age = 16 years old.)

Analyses:

- Compare average responses to measures at Time 1 and Time 2. Not all adolescents completed every measure, so most measures included data from only 57 of the 70 teens.

No Improvement / Gender Dysphoria Condition:
One of the main outcomes that scientists measured was *gender dysphoria*. They measured gender dysphoria using two scales:

- Gender dysphoria scale
  - Adolescents rated agreement with items like: "Puberty felt like a betrayal." and "I feel unhappy when I have to behave like my assigned sex."
  - This is a widely-accepted way to measure gender dysphoria in clinical psychology.
- Body Image Scale
  - The scale lists 30 body features. Adolescents rate their satisfaction with each body feature on a 5-point scale. Each of the 30 items falls into one of three basic groups based on its relative importance as a gender-defining body feature: (1) primary sex characteristics, (2) secondary sex characteristics, and (3) neutral body characteristics.

Here is what the scientists found:
Remember that Time 1 is before receiving puberty suppression and Time 2 is roughly two years after receiving puberty suppression.
Gender Dysphoria Scale:

- At Time 1: Adolescents reported severe gender dysphoria.
- At Time 2: There was *no change* in reported gender dysphoria compared to Time 1.

Body Image Scale:

- At Time 1: Adolescents reported overall low satisfaction with their primary and secondary sex characteristics, but high satisfaction with their neutral body characteristics.
- At Time 2: There was *no change* in their satisfaction ratings compared to Time 1.

Based on these findings, the authors wrote, "puberty suppression did not result in an amelioration of gender dysphoria" (p. 2281).
Improvement / Psychological Adjustment Condition:
One of the main outcomes that scientists measured was teens' *psychological functioning*. Here is how they measured psychological functioning:

- Child Behavioral Checklist
  - Parents evaluate their child's behavior across six topics including aggressive and rule-breaking behavior, other social problems, and emotional problems (e.g., "argues a lot").
  - (Commonly used in clinical psychology.)
- Youth Self-Report Scale
  - Adolescents rate their agreement or disagreement with 100 items that measure emotional or behavioral problems (e.g., "I disobey my parents", "I enjoy being with people").
  - (Commonly used in clinical psychology.)
- Depression Inventory
  - Adolescents rate their agreement with items such as: "I cry more now than I used to."
  - (Commonly used in clinical psychology.)

Here is what the scientists found:
Remember that Time 1 is before receiving puberty suppression and Time 2 is roughly two years after receiving puberty suppression.
Child Behavioral Checklist:

- At Time 1: Parents reported poor behavior (44% scored in the clinical range).
- At Time 2: Improvement in behavior (22% scored in the clinical range).

Youth Self Report Scale:

- At Time 1: Adolescents exhibited poor psychological adjustment.
- At Time 2: Adolescents reported better psychological adjustment compared to Time 1.

Depression Inventory:

- At Time 1: Adolescents were, on average, more depressed than the general population.
- At Time 2: Adolescents reported ***lower*** depression compared to Time 1.

Based on these results, the authors wrote, "psychological functioning of adolescents diagnosed with gender identity dysphoria had improved in many respects after an average of nearly 2 years of [puberty suppressants]" (p. 2281).

Appendix F. Additional study 4 analysis

In Study 4, the most common prior belief among participants was that gender affirming care would make someone "a little better off" ("3" on a 7-point rating scale) ($n = 223$, 28% of sample). We divided this sample into those who were more concerned about false acceptance ($n = 91$) and those who were more concerned about false rejection ($n = 132$). As seen in Fig. F1, acceptance of the study's findings among these participants depended on whether the study risked a judgment that the participant was particularly concerned about, $b = 2.20$, $SE = 0.58$, $z = 3.76$, $p < .001$. Acceptance of the Improvement (psychological functioning) finding, $b = -1.34$, $SE = 0.41$, $z = -3.27$, $p = .001$, and Null (gender dysphoria) finding, $b = 0.86$, $SE = 0.42$, $z = 2.07$, $p = .039$, depending on their moral concerns. Participants who were particularly concerned about wrongly believing that gender affirming care is helpful were equally likely to accept both new evidence that it would be helpful (consistent with their prior belief) and new evidence that it would not be (inconsistent with their prior belief) ($b = -0.07$, $SE = 0.43$, $z = -0.16$, $p = .874$). However, participants who were particularly concerned about wrongly believing that gender affirming care is not helpful were much more likely to accept the positive finding ($b = 2.13$, $SE = 0.40$, $z = 5.32$, $p < .001$).

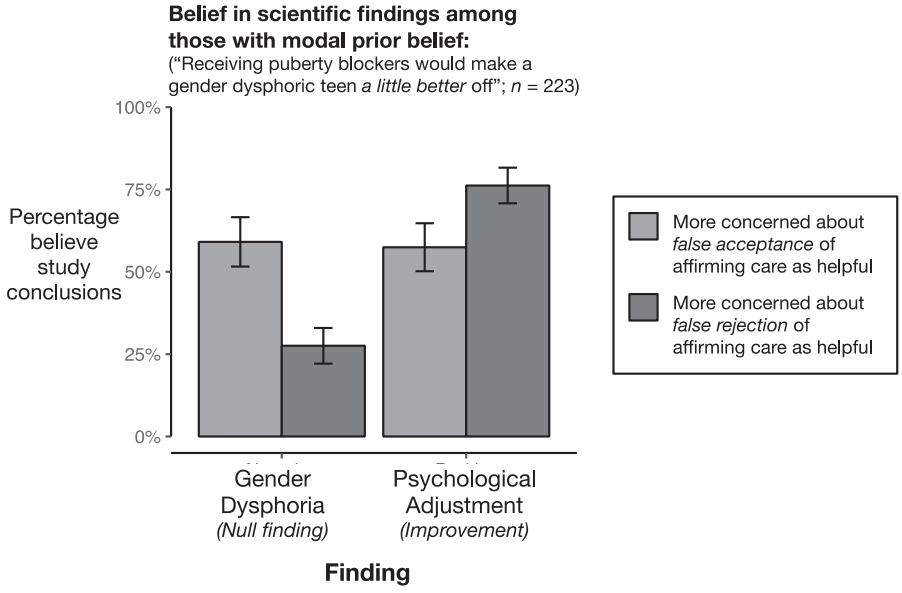

Fig. E.1. In Study 4, the percentage (and standard error) of belief in the study's finding, across conditions, among participants whose prior guess was that affirming care would make teens "a little better" off.

References

Appiah, K. A. (1995). The uncompleted argument: Dubois and the illusion of race. In A. G. Mosley (Ed.), *African philosophy: Selected readings* (pp. 199–215). Englewood Cliffs: Prentice-Hall.

Baron, J. (2008). *Thinking and deciding* (4 ed.). New York, NY: Cambridge University Press.

Baron, J., & Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the United States equally biased? *Perspectives on Psychological Science, 14*(2), 292–303. https://doi.org/10.1177/1745691618788876

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1–16. https://doi.org/10.1006/obhd.1997.2690

Basu, R. (2018). The wrongs of racist beliefs. *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1137-0

Basu, R. (2019). Radical moral encroachment: The moral stakes of racist beliefs. *Philosophical Issues, 29*(1), 9–23. https://doi.org/10.1111/phis.12137

Baumeister, R. F., & Newman, L. S. (1994). Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin, 20*(1), 3–19. https://doi.org/10.1177/0146167294201001

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88*(1), 1–45. https://doi.org/10.1037/0033-2909.88.1.1

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese.*. https://doi.org/10.1007/s11229-018-1809-5

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same bayesian judgment they criticize in others. *Psychological Science, 30*(1), 20–31. https://doi.org/10.1177/0956797618805750

Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology. General, 148*, 1701–1732. https://doi.org/10.1037/xge0000547

Cusimano, C., & Goodwin, G. P. (2020). People judge others to have more voluntary control over beliefs than they themselves do. *Journal of Personality and Social Psychology, 119*, 999–1029. https://doi.org/10.1037/pspa0000198

Cusimano, C., & Lombrozo, T. (2021a). Reconciling scientific and commonplace values to improve reasoning. *Trends in Cognitive Sciences, 25*, 937–949. https://doi.org/10.1016/j.tics.2021.06.004

Cusimano, C., & Lombrozo, T. (2021b). Morality justifies motivated reasoning in the folk ethics of belief. *Cognition, 209*, Article 104513. https://doi.org/10.1016/j.cognition.2020.104513

De Vries, A. L., Steensma, T. D., Doreleijers, T. A., & Cohen-Kettenis, P. T. (2011). Puberty suppression in adolescents with gender identity disorder: A prospective follow-up study. *The Journal of Sexual Medicine, 8*(8), 2276–2283. https://doi.org/10.1111/j.1743-6109.2010.01943.x

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Douglas, H. (2021). *The rightful place of science: Science, values, and democracy: The 2016 Descartes lectures*. Tempe, AZ: Consortium for Science, Policy & Outcomes.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*(1), 5–24. https://doi.org/10.1037/0022-3514.71.1.5

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin, 31*(5), 680–692. https://doi.org/10.1177/0146167204271570

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Frantz, C. (2006). I am being fair: The bias blind spot as a stumbling block to seeing both sides. *Basic and Applied Social Psychology, 28*(2), 157–167. https://doi.org/10.1207/s15324834basp2802_5

Gardiner, G. (2018). Evidentialism and moral encroachment. In K. McCain (Ed.), *Believing in accordance with the evidence: New essays on Evidentialism* (pp. 169–195). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-95993-1_11.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. Free Press.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168. https://doi.org/10.1037/a0034726

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People claim objectivity after knowingly using biased strategies. *Personality and Social Psychology Bulletin, 40*(6), 691–699. doi:10.1177%2F0146167214523476.

Hardy, S. A., & Carlo, G. (2005). Identity as a source of moral motivation. *Human Development, 48*(4), 232–256. https://doi.org/10.1159/000086859

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *The Behavioral and Brain Sciences, 34*, 1–16. https://doi.org/10.1017/S0140525X10001354

Jern, A., Chang, K. M., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review, 121*, 206–224.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin, 34*(6), 833–848. https://doi.org/10.1177/0146167208315158

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56*, 28–55.

Kruglanski, A. W. (1996). Motivated social cognition: Principles of the interface. In E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 493–520). The Guilford Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109.

Lynn, M., & Graves, J. (1996). Tipping: An incentive/reward for service? *Hospitality Research Journal, 20*(1), 1–14. doi:10.1177%2F109634809602000102.

Metz, S. E., Weisberg, D. S., & Weisberg, M. (2018). Non-scientific criteria for belief sustain counter-scientific beliefs. *Cognitive Science, 42*(5), 1477–1503. https://doi.org/10.1111/cogs.12584

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science, 4*, 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Mok, C., & Hansen, S. (1999). A study of factors affecting tip size in restaurants. *Journal of Restaurant and Foodservice Marketing, 3*(3–4), 49–64. https://doi.org/10.1300/J061v03n03_05

Moss, S. (2018). *Probabilistic knowledge*. Oxford University Press.

Plunkett, D., Buchak, L., & Lombrozo, T. (2020). When and why people think beliefs are "debunked" by scientific explanations of their origins. *Mind & Language, 35*(1), 3–28.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences, 11*(1), 37–43. https://doi.org/10.1016/j.tics.2006.11.001P

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 111*(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), *Advances in experimental social psychology: 20* (pp. 297–340). Academic Press. https://doi.org/10.1016/S0065-2601(08)60417-7.

Reeder, G. D., Pryor, J. B., Wohl, M. J. A., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin, 31*(11), 1498–1510. https://doi.org/10.1177/0146167205277093

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review, 123*(2), 182–207. https://doi.org/10.1037/rev0000017

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology, 68*(3), 404. https://doi.org/10.1037/0022-3514.68.3.404

Rogers, T., Moore, D. A., & Norton, M. I. (2017). The belief in a favorable future. *Psychological Science, 28*(9), 1290–1301. https://doi.org/10.1177/0956797617706706

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: My research journey. *Perspectives on Psychological Science, 13*(6), 750–769. doi:10.1177%2F1745691618769855.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *The Jean Piaget symposium series. Values and knowledge* (pp. 103–135). US: Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision making, 10*(4), 296–313.

Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour, 3*, 1055–1061. https://doi.org/10.1038/s41562-019-0666-7

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne social desirability scale. *Journal of Clinical Psychology, 28*(2), 191–193. https://doi.org/10.1002/1097-4679(197204)28:2<191::aid-jclp2270280220>3.0.co;2-g

Tenney, E. R., Logg, J. M., & Moore, D. A. (2015). (too) optimistic about optimism: The belief that optimism improves performance. *Journal of Personality and Social Psychology, 108*(3), 377–399. https://doi.org/10.1037/pspa0000018

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review, 109*, 451–471. https://doi.org/10.1037/0033-295x.109.3.451

Tetlock, P. E., Kristel, O. V., Beth, S., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*(5), 853–870. https://doi.org/10.1037/0022-3514.78.5.853

Trope, Y., & Liberman, A. (1996). Social hypothesis-testing: Cognitive and motivational mechanisms. In E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239–270). The Guilford Press.

Walco, D. K., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological Science, 28*(12), 1807–1820. https://doi.org/10.1177/0956797617723377

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology, 103*(3), 506. https://doi.org/10.1037/a0028857