At the Intersection of Moral Psychology and Belief Formation: An Investigation of Bias, Biasing Influences and Behavioural Correlates

Benjamin Michael Tappin

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology

Royal Holloway, University of London

2018

Abstract

This thesis presents an empirical investigation at the intersection of human *moral psychology*—people's perceptions of good and bad, right and wrong—and human *belief formation*—how people evaluate evidence and update their beliefs. The investigation is divided into two parts. In Part I, across 6 studies I investigate (i) people's beliefs about their own moral goodness relative to the average person—in particular, I ask whether and to what extent such beliefs are irrational—and (ii) people's beliefs about the moral goodness of their political in-party relative to their political out-party. Using economic games, I subsequently test whether people's beliefs regarding (i) and (ii) correlate with behavioural outcomes. Finally, I test whether people's motivation to "do the morally right thing" underpins their prosocial behaviour. In Part II, also comprising 6 studies, I investigate several factors that are purported to influence belief formation in the morally-charged context of contemporary US politics. In particular, I study (i) whether belief updating is biased by the beliefs people already hold or by their desired political outcomes (or both), and (ii) the extent to which cognitive sophistication facilitates biased information processing such that people who are more sophisticated are more likely to form factual beliefs that are favourable to their political identities. I draw four main conclusions from my investigation. First, people's beliefs about their own (vs. others') moral goodness, and about the moral goodness of their political in-party (vs. out-party), suggest a robust perception of "moral superiority". Second, said perceptions of moral superiority are not reliably related to the behavioural outcomes I examine; more reliably related to these behavioural outcomes is people's motivation to do what they perceive to be morally "right". Third, belief formation is possibly biased by people's desired outcomes and by their political identities, but the evidence is relatively undiagnostic on this front. The evidence is similarly undiagnostic as to whether cognitive sophistication *facilitates* such a bias. Fourth, and finally, following other scholars I conclude that reasonable inferences of "bias" in human belief formation demand evidence of systematic deviation from well-specified normative standards.

Acknowledgments

There are numerous people to whom I owe a debt of gratitude for guiding and supporting me during the past three years of work. First and foremost, I am grateful to my supervisor, Ryan. I feel very fortunate to have had such a positive supervisory experience: Ryan walked the line admirably between giving me the creative and intellectual freedom to explore ideas and develop my own lines of work, whilst also being passionate and engaged when we discussed research ideas and designs together. He was open to the idea that psychological science required changes in practice to create a more rigorous research culture and reproducible literature—and he supported me wholeheartedly in exploring preregistration and other such practices during the early stages of my PhD studies. Perhaps most importantly, though, Ryan came to be a valued and trusted confidant and friend; a friendship that (for my part) will continue long into the future.

The PhD experience would have been considerably less enjoyable were it not for the raucous scallywags for officemates that I had the fortune of befriending over the years. Together, we shared the ups and the downs (such as they are) of doctoral life; rendering the ups altogether more pleasurable, and the downs more bearable. Similar gratitude also goes to other colleagues and friends that I met along the way; including Robert Ross, Gary Lewis, Leslie van der Leer, Valerio Capraro, David Rand and Gordon Pennycook. Rob and Gary came to be good friends with whom I could chat about and discuss any topic, and be sure to receive cool and analytic insight. This was extremely helpful given that the topics of my thesis—moral psychology and politics—do not always lend themselves to reasonable or cool-headed discussion. I had the pleasure of meeting Leslie, Valerio, Dave and Gord during different stages of my studies, and we developed friendships and collaborations; some of which appear in this thesis, and others which are ongoing.

Lastly, I would like to extend a special thanks to the three most important people in my life: my parents, and my partner—and partner-in-crime (!)—Leysa Forrest. I owe my mum, Janet Hooper, and my Dad, Alan Tappin, more than I can put into words. Suffice to say that I could not have wished for better support or for more opportunities in life than they have given to me. The work in this thesis is as much theirs as it is mine. I feel tremendously lucky to have met my partner, Leysa, in the early stages of my doctoral studies. She has been a source of relentless

love and support throughout the PhD, and I am forever grateful to her for putting up with the long working hours, the spells of self-doubt and questioning, and the other baggage that comes with PhD life. She has been my rock.

Table of Contents

Bibliography of Work Completed During PhD

Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science, 8*, 623-631.

Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General, 146*, 1143-1149.

Tappin, B. M., McKay, R. T., & Abrams, D. (2017). Choosing the right level of analysis: Stereotypes shape social reality via collective action. *Behavioral and Brain Sciences, 40*. [Not in thesis].

Tappin, B. M., & Capraro, V. (2018). Doing good vs. avoiding bad in prosocial choice: A refined test and extension of the morality preference hypothesis. *Journal of Experimental Social Psychology, 79*, 64-70.

Tappin, B. M., & McKay, R. T. (2018). Investigating the relationship between self-perceived moral superiority and moral behavior using economic games. *Social Psychological and Personality Science*, 1-9.

Tappin, B. M., Ross, R. M., & McKay, R. (2018). Do the folk actually hold folk-economic beliefs? *Behavioral and Brain Sciences, 41*. [Not in thesis].

Tappin, B. M., & McKay, R. T. (submitted). Moral Polarization and Out-Party Hate in the US Political Context. Under review at *Journal of Social and Political Psychology*.

Tappin, B. M., Pennycook, G., & Rand, D. (submitted). Rethinking the link between cognitive sophistication and identity-protective bias in political belief formation. Under review at *Psychological Review*.

Introduction

Context of and Rationale for Research

Meta Context

The research reported in this thesis was conducted during the time period 2015 through 2018. During that period, there were prominent changes in the political and (psychological) scientific landscape; changes that had an impact on the content, direction and calibre of the research that I conducted. Regarding the political changes, the election of Donald Trump in the United States and the EU referendum in Britain—both occurring in 2016—contributed to the development of studies in which I investigated factors hypothesized to affect belief formation in politics; such as people's desired political outcomes, as well as their political identities and attachments. Furthermore, these events rendered in new light the importance of moral psychology in human perception, belief formation and decision-making. In the case of the British EU referendum, for example, people appeared willing to sustain material losses—in the form of probable damage to the British economy—for the principles of fairness, autonomy and self-governance (Crockett, 2016). Many people believed that their voting decision was right and good, while the decision of the other side was wrong and bad. My investigations of people's moral beliefs about themselves and about others, and about their political in-party and political out-party, were conducted against this very salient backdrop; as, too, was my examination of how these beliefs relate to behavioural outcomes.

The landscape of psychological science itself was also changing during this time—profoundly influencing the quality of the research that I conducted. In particular, there was an increasing awareness that previous research was not as robust or reliable as many perhaps thought (Camerer et al., 2018; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011), and growing calls for open scientific practices and better understanding and use of statistical methods in psychology and other sciences (e.g., Munafò et al., 2017; Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek et al., 2015). To be sure, many of these were not new problems in psychology (e.g., see Meehl, 1967, 1978) but they had now entered mainstream scientific consciousness, and were building up a head of steam as I began my doctoral studies in 2015. To my estimation, these events positively and concretely affected my

research in at least three ways. First, foregoing my earlier training, I learned to script and document all my data analyses from scratch in the statistical programming language *R* (R Core Team, 2017); second, I began to publicly archive all raw data and analysis scripts underlying the studies that I conducted and the manuscripts that I authored; third, and finally, I transitioned to preregistering all confirmatory hypothesis tests that I performed. In my opinion, the quality of the research in this thesis is much improved for these practices (but I defer to the readers' good judgment!), and I am tremendously grateful to all the people who directly or indirectly brought these issues to my awareness.

Theoretical Context and Rationale

This thesis lies at the intersection of two overlapping but ultimately distinct fields of research. The first, which I refer to here under the umbrella term "moral psychology", concerns the study of human understanding of right and wrong, good and bad, and how these constructs relate to and affect perception, judgment and decision-making. The second, which I refer to as "belief formation", concerns the study of how humans arrive at their beliefs about the world—through such processes as sampling information from their environments, reasoning about information, and incorporating new information into their existing beliefs (i.e., belief updating). As alluded to above, these distinct fields of research often find a common home in the domain of politics—where people's moral psychology appears very much interleaved with their consumption of information and the formation of their beliefs. For this reason, most—though not all—of the research that I report in this thesis was conducted in a political context. For example, in Part I of the thesis I report two studies which were conducted to test whether political partisans' beliefs about the moral character of their own party vs. that of the other party predicts their behavioural hostility towards the latter. In Part II, along similar lines, I report a series of studies that investigated whether people's belief formation is biased by their political identities—and, in particular, whether cognitive sophistication exacerbates said bias. In the following sections, I elaborate further on the theoretical linkages within each major part of my investigation, as well as highlighting briefly the rationale for the studies that I conducted.

Part I: Moral Psychology and Moral Behaviour

Part I of this thesis comprises 6 studies set against the theoretical backdrop of moral psychology and self- and social perception. The broad linking theme of these studies is their focus on human perceptions of good and bad, and how those perceptions relate to beliefs about oneself and others, the nature of those beliefs, and whether the beliefs predict behaviour. Furthermore, there are subthemes shared between particular studies.

For example, in Parts 1.1 and 1.2 I measure the same key outcome variable—people's beliefs about their own moral goodness relative to the average person—but test different focal hypotheses. In particular, in Part 1.1 I examine the *nature* of these beliefs; specifically, whether and to what extent they evince irrationality or "bias" on behalf of the believers. The prevalence of the tendency for people to believe themselves to be superior to the average person (for a review, see Alicke & Govorun, 2005; for phenomenon boundaries, see Heine & Hamamura, 2007; Kruger, 1999)—particularly on moral traits like "honesty" and "trustworthiness"—has been interpreted by some scholars to mean that these beliefs are *irrational*, but persist because they serve an adaptive function like protecting or enhancing mental wellbeing (Taylor & Brown, 1988). However, this interpretation is incommensurate with the empirical evidence, which is undiagnostic as to whether the beliefs are, in fact, irrational (Heck & Krueger, 2015; Krueger & Wright, 2011; see also Hahn & Harris, 2014). Thus, Part 1.1 takes up the challenge of better identifying (i) whether and to what extent these beliefs may be considered irrational, and (ii) whether they are correlated with indicators of positive mental health.

In Part 1.2, in contrast, I investigate whether these beliefs predict *behavioural* outcomes; in particular, behaviours that are commonly considered moral—like freely helping others, and reciprocating trust. There are several reasons why beliefs about one's moral superiority over the average person may predict such behaviour. On the one hand, for example, people with a strong sense of righteousness over others may be motivated to behave in ways that protect this positive social comparison—to maintain self-esteem and thus feel good about themselves (Sedikides & Strube, 1997; Wills, 1981). This generates the prediction that perceived moral superiority will be positively correlated with moral behaviour. On the other hand, insofar as a strong sense of righteousness over the average person reflects a *pessimistic* view of the morality of others—rather than a distinctly positive view of oneself (Van Damme et al., 2016)—there is reason to expect the opposite association; that is, a *negative* correlation. In particular, because evidence suggests that various types of moral behaviour (those that rest on interdependency) are *less* likely if people perceive that others will not behave in kind (Krueger & Acevedo,

2007). Given these (and other) competing predictions, in Part 1.2 I test the association between self-perceived moral superiority and moral behaviour.

In Part 1.3, I stay with the theme of measuring perceived moral superiority. In this case, however, I extend this measurement to people's beliefs about their political *in-party* and political *out-party* in the US context. Research in political science documents what is known as "affective polarization", the tendency for Democrats and Republicans to view co-partisans positively and out-partisans negatively. A recent review of this phenomenon (Iyengar et al., 2018) identifies significant gaps that remain in scholarly understanding; most notably, whether and to what extent affective polarization is related to *behavioural hostility* towards the out-party, and whether it reflects a *generalized* evaluative discrepancy between partisans vs. a more domain-*specific* disparity—such as the belief that in-partisans are more trustworthy than out-partisans. Drawing upon recent theoretical and empirical trends in psychology and political science, I hypothesize that *moral polarization*—the tendency for people to view co-partisans' *moral character* positively, and opposing partisans' *moral character* negatively—predicts behavioural hostility towards members of the out-party. I test this hypothesis using a behavioural economic game measure of out-party hostility and large samples of US partisans.

A shortcoming in the designs of the studies in Parts 1.1-1.3 is that they provide correlational evidence only—that is, I do not randomly assign a sense of moral superiority, nor the motivation to choose the "moral" behaviour. In Part 1.4, by contrast, I conduct an experiment that uses framing effects to manipulate people's moral behaviour. The rationale for this experiment was to conduct a refined test of the so-called *morality preference hypothesis*. This hypothesis challenges classic behavioural economic theory, which invokes outcome-based social preferences to explain anonymous, one-shot prosociality—such as giving money to a homeless person that one will never meet again. Famous examples of outcome-based social preferences include "inequity aversion" (Fehr & Schmidt, 1999); the notion that prosocial behaviour (of the kind above) is underpinned by people's preference to avoid inequitable social outcomes. Rejecting this explanation, the morality preference hypothesis says that prosocial behaviour is a function of people's preference for doing what they perceive to be the morally right *action*—not a preference for a particular social *outcome*. I identify significant confounds in recent work that purports to find evidence for this hypothesis (Capraro & Rand, 2018). In Part 1.4, I correct these confounds and extend the design to test whether the morality preference

is stronger for choosing the morally "right" behaviour or for avoiding the morally "wrong" behaviour.

Part II: Desires, Identities and Bias in Political Belief Formation

Part II of this thesis also comprises 6 studies, set against the theoretical backdrop of belief formation in politics. The theme that links these studies together is their focus on the various factors that are hypothesized to *bias* political belief formation. In particular, the formation of beliefs about politically-relevant *facts*—that is, what is true "out there" in the world—distinct from preferences, attitudes or behavioural intentions.

In Part 2.1, I investigate two distinct phenomena posited to bias belief updating. Here I refer to these phenomena as *confirmation bias* and *desirability bias*, respectively. The former phenomenon predicts that people's prior beliefs bias their belief updating such that they tend to incorporate new information into their posterior beliefs to a greater extent if it confirms (vs. disconfirms) their existing beliefs (all else being equal). Arguably the most famous example of this phenomenon is a study reported by Lord, Ross and Lepper (1979). These authors provided people with mixed evidence about the efficacy of capital punishment in deterring crime. They found that *proponents* of capital punishment—those who believed it was effective in deterring crime—became *more* convinced of its effectiveness following the mixed evidence; while *opponents*—those who believed that it was *in*effective—became more convinced of its *in*effectiveness in deterring crime. The authors took this result to imply that people incorporated the belief-confirming information into their posterior beliefs to a greater extent than the belief-disconfirming information. This interpretation has received significant criticism and qualification since publication of the original study, however (Guess & Coppock, 2018; Kuhn & Lao, 1996; Miller et al., 1993; Munro & Ditto, 1997); a point to which I return in the General Discussion section of this thesis.

The second phenomenon—referred to here as *desirability bias*—predicts that people's desired outcomes bias their belief updating such that they tend to incorporate new information to a greater extent if it is desirable vs. undesirable (all else being equal). The empirical evidence for this phenomenon was recently reviewed in Sharot and Garrett (2016). Though there exists a substantial body of evidence purporting to demonstrate such a bias, much of this evidence was

obtained using a methodological paradigm that contains problematic confounds (reported on at length in Shah, Harris, Bird, Catmur, & Hahn, 2016). Added to this, it remains unclear whether and to what extent people's desired outcomes bias their belief updating in the sense of causing systematic departures from *rational* updating, for example that prescribed by Bayes' theorem (Hahn & Harris, 2014). I also discuss these and related issues at length in the General Discussion section of this thesis.

The empirical predictions of these two phenomena—confirmation bias and desirability bias—are often confounded. This is because people often hold beliefs that they would also *prefer* to be true. For example, in the case of the Lord et al. (1979) study, it is reasonable to assume that proponents of capital punishment would also *prefer* a world in which capital punishment deters crime—perhaps because this would validate their worldview or political commitments. The upshot is that providing evidence to people and measuring their belief updating confounds an effect of desired outcome (desirability bias) with that of existing beliefs (confirmation bias). In Part 2.1, I thus attempt to tease these factors apart—and measure their associations with belief updating—by conducting an experiment in the context of the 2016 US presidential election.

The studies reported in Part 2.2 are related to the study in Part 2.1 insofar as they also investigate belief formation in the US political context. However, there are several key differences. For example, as well as the process of *belief updating*—changes in beliefs—I additionally study *evidence evaluation*; that is, how people judge the validity of the new evidence itself. Popular theoretical accounts of political belief formation assume that people's political identities—typically construed as the political parties with which they identify—*bias* their information processing such that they are prone to form beliefs that are favourable to those identities (e.g., Kahan, 2016a; Van Bavel & Pereira, 2018). A counterintuitive and rather troubling hypothesis advanced on the basis of these accounts is that cognitive *sophistication* tends to *exacerbate* this bias in information processing (Kahan, 2013; Kahan et al., 2017); polarizing the beliefs of cognitively sophisticated partisans who identify with opposing political groups. In effect, the claim is that the distinct proficiencies of these partisans are used to form identity-congruent assessments of new evidence (Kahan, 2017). Over five studies, I put this claim to the test—measuring diverse outcome variables spanning the distinct processes of belief updating and evidence evaluation. In addition, as alluded to above, a particular shortcoming of the study in Part 2.1 is the lack of a normative benchmark against which to

evaluate putative bias in belief updating (cf. Hahn & Harris, 2014). Thus, the relevant studies in Part 2.2 measure people's patterns of belief updating with respect to that of a Bayesian agent.

Methodology

Sampling Population and Strategy

The primary sampling population used in this thesis is *Amazon's Mechanical Turk* (MTurk). MTurk is a US-based online marketplace where "requesters" post Human Intelligence Tasks (HITs) for "workers" to complete. Typically, HITs are basic, monotonous tasks that machines struggle to complete—such as composite image recognition—but which humans find trivial (hence the moniker "human intelligence tasks"). In recent years, however, increasing numbers of behavioural scientists have turned to MTurk as a source of relatively inexpensive and rapidly available data (Buhrmester, Kwang, & Gosling, 2011; Rand, 2012; Stewart, Chandler, & Paolacci, 2017). Indeed, in all but one case (Study 4 in Part 2.2) I draw my study subjects from MTurk. In the following sections I thus outline in detail the strengths and weaknesses of MTurk as they relate to the research reported in this thesis.

Strengths

The explosion in the use of MTurk in behavioural science has prompted extensive study of its sample characteristics. Such studies have revealed that, while workers are not demographically representative of the general US population—they tend to be younger, more educated, more liberal (politically) and disproportionately White (Chandler & Shapiro, 2016)—they *are* more diverse than the average nonprobability sample used in behavioural science (Berinsky, Huber, & Lenz, 2012; Casler, Bickel, & Hackett, 2013). In other words, relative to undergraduate university students (the typical psychology study population), use of MTurk affords greater license to generalize results beyond the confines of the particular study in which they were obtained. Of course, engaging in confident generalization of results from any nonprobability sample of an unrepresentative population is fraught with difficulties, and, in general, should be avoided. I therefore consider the relative demographic diversity of MTurk—vs. the average behavioural science sample—a slightly-less-problematic *weakness*, than a strength *per se*. Indeed, I revisit sample diversity in the "Weaknesses" subsection below.

MTurk provides access to an enormous pool of potential research subjects. While Amazon claims that there are over 500,000 registered worker accounts (Chandler & Shapiro, 2016), this

number is likely to be a considerable *over*estimate of those completing HITs; impartial analyses estimate that there are approximately 15,000 unique US worker accounts active at any one time (Stewart et al., 2015). Despite this radically lower estimate, however, MTurk still provides access to a number of potential research subjects that far surpasses that of the average research pool previously available to researchers; which, on a liberal estimate, was probably somewhere between 150-250 undergraduate psychology students (in a given academic year). Accordingly, a significant advantage of MTurk is the ability for researchers to recruit a higher number of study subjects and thus increase *statistical power* when testing hypotheses.

This advantage is particularly acute in the case of between-subjects randomized experiment designs, where the required sample size to detect an effect size of $r = .21$ (the average effect size in social psychology, see Richard, Bond Jr., & Stokes-Zoota, 2003) with 80% statistical power is N = 174[1]. This sample size estimate reflects that required for a simple comparison between two groups; frequently, however, researchers (including myself in the studies in this thesis) desire to test second- and third-order interaction terms, meaning that N = 174 is likely to be a considerable *under*estimate of that required to test hypotheses common in psychology. Moreover, $r = .21$ is an estimate of the *average* effect size—implying that many effects that are studied in psychology will be smaller, and thus require even larger sample sizes to detect. Historically, most studies in psychology have had much smaller sample sizes than that required to detect the average effect size in the field; implying that numerous classic studies are severely underpowered. Combined with publication bias, optional stopping in data collection and other "questionable research practices" (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John, Loewenstein, & Prelec, 2012; Simmons et al., 2011), this may help explain the low rates of replication observed in the research literature (Camerer et al., 2018; Open Science Collaboration, 2015).

Much of the research reported in this thesis employs between-subjects randomized experiment designs—for example, the studies reported in Parts 1.4, 2.1 and 2.2. These studies involve focal statistical tests ranging from that on main effect terms (e.g., simple group comparisons, as above) to three- and even four-way interaction terms. The sample sizes range commensurately from N = 800 (Part 1.4) to N = 2000 (Part 2.2, Study 4), and, in all cases, are accompanied by

[1] This sample size estimate was generated in G*Power (v. 3.1.9.2, Faul, Erdfelder, Lang, & Buchner, 2009) using the following parameters for an independent t-test: Two-tailed, d = 0.4296 (equivalent to r = .21), $\alpha = .05$, $\beta = .80$, allocation ratio (i.e., the ratio of subjects in group 1 to group 2) = 1.

a priori power analyses or other principled justification for the sample size used (e.g., where I conduct replications of other scholars' work, I recruit *at least* the original sample size, if not 1.5-2 times greater). In most cases, without MTurk, these hypothesis tests would not have been feasible—or ethical—to conduct because of sample size constraints. I thus consider the size of the MTurk subject pool to be a substantial boon to the quality and content of the research in this thesis, and also to psychological science in general (whatever else may be said for MTurk).

A final strength that I would like to highlight is the rapidity of data collection afforded by MTurk (and, in fact, any *online* data collection). It is not uncommon that studies with a sample size of several hundred or more are completed within a few hours of initializing data collection. In my experience, this allowed me to redirect time that I would otherwise have spent in the lab supervising subjects—for hundreds of hours or more given the number of subjects required—to experiment design, reading articles and otherwise developing my research skills. Indeed, the knowledge that data collection would be so rapid meant that studies could remain in the design and analysis preregistration stage for considerably longer, and were thus more informative and of higher quality in the long run. Similarly, I was able to spend more of my time learning to use the statistical programming language *R*, preparing and documenting data and scripts for public archiving, and travelling to training courses and workshops to develop my research skills. In my estimation, these gains greatly improved the quality of the research reported in this thesis, and were in large part due to time saved not sitting in the lab supervising individual study subjects. Of course, not all research questions can be answered by collecting data online (and via MTurk in particular), but I was fortunate enough to work in a research area where this was—for the most part—the case.

Weaknesses

Given the politically left-leaning composition of MTurk, and the political focus of much of the research in this thesis, it is responsible to consider whether MTurk samples provide particular threats to inference in the case of my research. On the one hand, the relative lack of politically right-leaning workers in the MTurk subject pool clearly puts constraints on my ability to generalize some of my results to that population—for example, my results regarding political belief formation (e.g., Part II) (Kahan, 2013, 2016a). On the other hand, however, *none* of my primary hypothesis tests are focused on the political characteristics of right-leaning individuals

per se[2]. For example, whether they behave differently from left-leaning individuals (as in the line of research conducted by e.g., Jost and colleagues, 2003, 2017). Instead, my investigations focus on more general aspects of cognition and behaviour, and I use right- *and left*-leaning partisans—typically in conjunction—to test hypotheses of interest.

Added to this, recent work from Clifford and colleagues (2015) suggests that the political left and political right on MTurk share similar psychological characteristics to the left and right in nationally representative samples. In particular, these authors find that they score similarly on measures of personality and values related to political ideology (*ibid.*). This suggests that political partisans recruited via MTurk are not psychologically *incomparable* to the average US partisan. This suggestion converges with recent work on the generalizability of results from convenience samples. In particular, there appears to be minimal *heterogeneity* in average treatment effects (ATE) estimated on MTurk samples vs. national probability samples of US adults. Coppock (2018) recently replicated 12 political science experiments on MTurk and found a correlation of $r = .83$ between the ATEs and ATEs obtained in nationally representative samples (see also Mullinix et al., 2015). Moreover, this relative homogeneity in ATEs across samples has been demonstrated to occur because the *conditional* average treatment effects—that is, the ATEs among sample *sub*groups (e.g., Democrats and Republicans)—are strongly correlated across convenience and representative samples (Coppock, Leeper, & Mullinix, 2018). Put more simply, diverse subjects in both types of samples tend to respond similarly to treatment effects. It therefore seems reasonable to conclude that studies in which I estimate experimental effects are not strongly adversely affected by the political composition of MTurk.

Of course, when estimating *non*experimental effects or interaction terms with covariates, the scarcity of right-leaning MTurk workers may have more strongly affected inferences. For example, in Study 4 of Part 2.2 I conducted a direct replication of the immediately-preceding Study (3). Whereas the latter's sample was drawn from MTurk, the former drew from *Lucid*—a survey platform whose population is more representative of the general US population (Coppock & McClellan, 2017). It is telling that I did observe some key differences in results between these two studies, differences which may have been due to the change in sampling population. However, I also observed key *similarities*, and it is equally plausible that the

[2] In several cases, I test exploratory hypotheses that ask questions related to left-right differences in cognition/behaviour; for example, in Parts 1.3 and 2.1. These are clearly pointed out in the relevant studies.

aforementioned differences were due to minor changes in the design/procedure of the study and/or sampling variability (for more detail, I refer to Part 2.2 Study 4 Discussion).

The popularity of MTurk as a sampling population has generated problems of itself. Most acute is the problem of nonnaiveté among study subjects, and what has been dubbed an emerging "tragedy of the commons" in data quality (Stewart et al., 2017; cf. Hardin, 1968). The crux of the problem is that many research labs now draw their study subjects from MTurk, and, in some cases, have done so for many years. Popular experimental paradigms, tasks, and other measures have thus become very familiar to a subset of MTurk workers; typically, those who are most active on the platform. One possible consequence of this increase in familiarity is a *decrease* in effect sizes obtained in experimental tasks (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015). Another consequence may be inflation of task performance through practice effects.

Regarding the research reported in this thesis, studies in which task familiarity may pose a particular threat to inference include all of the studies in Part 2.2, as well as the studies in Parts 1.2 and 1.4. In the former studies (Part 2.2), I used a task known as the "Cognitive Reflection Test" (CRT)—a 3-item behavioural measure of the propensity to think analytically (Frederick, 2005). The 3-item CRT has been used extensively on MTurk, and many workers are now familiar with the items that comprise the test (Thomson & Oppenheimer, 2016); possibly upwardly biasing test scores (Stewart et al., 2017).

Fortunately, there are several factors that allay concerns regarding threat to inference in the case of Part 2.2. First, in all relevant studies I use an *updated* version of the CRT—which lengthens the test to 7-items, and rewords the original three items so that they are less familiar (Shenhav, Rand & Greene, 2012; Thomson & Oppenheimer, 2016). In fact, introduction of these new items—combined with some residual familiarity among MTurk workers—meant that the scores on the test were approximately normally distributed in most of my samples (see Part 2.2, Supplemental Material). This stands in contrast to scores in the general population where over a third of people get *none* of the answers correct (e.g., Kahan, 2013). Given that scores on the CRT are typically correlated with other measures, the distributions in my samples actually provide statistically- and theoretically-*favourable* conditions for associations of interest to emerge. A related point is that the CRT appears to retain its predictive validity even despite nonnaiveté (Bialek & Pennycook, 2017; Meyer et al., 2018). In other words, while the

average score in a sample may increase with familiarity, the between-subjects *ordering* of the scores—which is relevant to the predictive validity of the CRT—does not change much. Since the studies in Part 2.2 concern the predictive validity of the CRT, I consider some familiarity with the items to be minimally problematic for inference.

In Parts 1.2 and 1.4, I use classic behavioural economic games to measure prosocial behaviour. Specifically, Dictator Games (DG) and Trust Games (TG). Such games are commonly used on MTurk, and recent research has identified that subject behaviour tends toward the rational, game-theoretic optimum with experience (Rand et al., 2014). While I did not implement specific countermeasures in my studies to guard against prior experience with the DG and TG, I note that the results of the relevant studies are far from equivocal in any case. In particular, in the studies in Part 1.2, across a series of analytic techniques—including Bayesian analyses—I observe robust evidence in favour of the null hypothesis. Similarly, in Part 1.4 I observe relatively compelling *rejections* of the null hypothesis—the relevant |z| and |t| values exceeding 5 and 4, respectively. It strikes me as implausible that a (probably small) experience-driven shift in behaviour in the DG/TG would overturn such results in either of these cases (i.e., Parts 1.2 and 1.4). Nevertheless, I admit that I cannot rule it out completely.

A further potential threat to data quality on MTurk is highlighted by the recent discovery (in late summer 2018) of "server farms", where would-be study subjects concealed their physical locations using virtual private servers—circumventing traditional techniques that prevent the same worker from completing the same study more than once (Dennis, Goodson, & Pearson, 2018; TurkPrime, 2018). Moreover, the data quality provided by these subjects was found to be considerably lower than other subjects (*ibid.*). This episode illustrates the wider issue of data quality when data is collected online, and not under the supervision of the researcher. Fortunately, the latest data collection period for the research in this thesis was early summer 2018—which likely missed the majority of the server farms episode. Beyond this, however, my primary countermeasures to poor data quality via inattentive responding were to (i) specify only those workers who met certain quality criteria (such as completing a certain number of HITs in the past, cf. Chandler & Shapiro, 2016) and (ii) implement subtle "attention checks" and comprehension questions during data collection—excluding check failures for my primary analyses. Moreover, in most studies reported in the thesis I also conduct a range of exploratory *sensitivity* analyses and *robustness* checks—to ensure that results do not depend upon particular exclusion criteria (even those that are preregistered). For example, in the studies in Part 1.3,

after conducting such sensitivity analyses and robustness checks I conclude that the primary preregistered result is not particularly robust, and any effect is likely to be small/tenuous.

Measuring Outcomes: Self-Report and Behavioural Economic Games

The majority of studies in this thesis measure outcomes via self-report scales, most commonly some form of Likert scale or other scale-type response. However, as already mentioned, in several studies I also measure outcomes in the form of decisions made in financially-incentivized economic games. With respect to the content of research in this thesis—moral psychology and belief formation in politics—these alternative methods of measuring outcomes harbour different strengths and weaknesses, with implications for inference. I outline and discuss these below.

Using self-report as a format for measuring outcomes presents a particular problem in the case of my studies that focus on moral psychology and self- and social perception. That is, because of *socially desirable responding*. Moral traits and behaviours are highly desirable (Van Lange & Sedikides, 1998), and, given the confines of individual studies, difficult to check against reality (Alicke & Govorun, 2005). As a result, study subjects face little disincentive—and significant *in*centive—to exaggerate their possession of moral traits, and their willingness to engage in moral behaviours, when providing these via self-report. Indeed, a series of studies conducted by Epley and Dunning (2000) shows that individuals consistently over-report their moral qualities. For example, in one study, 84% of subjects reported that they would cooperate with their partner in a classic Prisoner's Dilemma game[3]; yet, when compared against another group of subjects (randomly assigned to actually play the game), only 61% of subjects chose to cooperate (*ibid.*). In another study from the same paper, individuals overestimated the likelihood that they would contribute to charity (through buying a daffodil) by a factor of almost 2-to-1 (84% predicted that they would, yet when the time came only 43% did so).

[3] While cooperation in the Prisoner's Dilemma may not seem a particularly representative—or ecologically valid—example of "moral" behaviour, research finds that people strongly and consistently recognize cooperation as morally superior to defection in this and similar economic games (e.g., Krueger & Acevedo, 2007; Krueger & DiDonato, 2010; Krueger et al., 2008). This indicates that cooperation in such games is considered a moral behaviour (at least, relative to defection) by the layperson.

This discrepancy between *self-reported* and *actual* behaviour in the study of moral psychology and self- and social perception is powerfully illustrated through the studies in Parts 1.2 and 1.3 of this thesis. There, I find fairly convincing evidence that patterns of self- and social perception obtained via self-report do not strongly—or even moderately—predict relevant behavioural outcomes; the latter measured using financially-incentivized economic games. Specifically, in Part 1.2 I find that subjects' self-perceptions of trait moral superiority over the average person do not predict their willingness to freely help others (in the Dictator Game), or reciprocate trust (in the Trust Game). I do find some evidence that subjects' perceptions of their *own* moral goodness *per se*—that is, independent of their perceived superiority over others—predicts behavioural outcomes. But, even then, the effects are almost trivial in size (e.g., $|r| < .15$), and the estimates are fairly imprecise despite large samples ($N > 400$).

In Part 1.3, similarly, I find considerable effects regarding the disparity between subjects' moral trait ratings of their political in-party vs. political out-party; including on traits such as "trustworthy" and "honest". Nevertheless, in the behavioural task (a modified version of the Intergroup Prisoner's Dilemma), I find limited evidence that the disparities on trait ratings predict outcomes. These studies, once again, were powered to detect effect sizes arguably approaching triviality (e.g., Odds Ratio = 1.4, roughly equivalent to $r = .09$); and, thus, are robust to concerns about lack of statistical power. In sum, the results of the studies in Parts 1.2 and 1.3 illustrate the value in not solely relying on self-report measures when investigating phenomena in the domain of moral psychology and self- and social perception. Furthermore, these behavioural results constrain possibly-wayward inferences that might have been made based on the self-report results alone—both here and by other scholars in the literature (e.g., see the Discussion in Part 1.3 for a particular example case).

Concerns about the validity of self-report extends to Part II of this thesis. Recall that the studies in this section investigate political belief formation—and, in fact, in these studies I make exclusive use of self-report to elicit the beliefs of subjects. Recent investigations have suggested that individuals are prone to report *insincere* beliefs about political issues—that is, beliefs they do not actually hold—or to exaggerate said beliefs in order to "cheer" for their political party (Bullock, Gerber, Hill, & Huber, 2015; Khanna & Sood, 2018; Prior, Sood, & Khanna, 2015; Schaffner & Luks, 2018). Though some scholars have contested its prevalence (e.g., Berinsky, 2018), the weight of evidence appears in favour of the notion that self-reported political beliefs are contaminated by such partisan cheerleading. For example, offering small

amounts of money for honest answers to politically-relevant factual questions (including "don't know" responses) diminishes estimates of partisan polarization over such questions (Bullock et al., 2015; Prior et al., 2015). This implies that "the apparent gulf in factual beliefs between members of different parties may be more illusory than real" (Bullock et al., 2015, abstract).

The precise extent of contamination by partisan cheerleading is an empirical question that (to my knowledge) remains up for debate, and, indeed, probably varies according to the beliefs under investigation. However, it must be acknowledged that the studies I present in Part II—in which I elicit the political beliefs of subjects via self-report—are likely to be contaminated to at least some degree. Given this, it is responsible to consider the implications for inference. On the one hand, a uniform effect of partisan cheerleading may serve to exaggerate effect sizes across the board. That is, insofar as the relevant hypothesis tests in Part II concern partisans' beliefs about politically-*favourable* vs. politically-*unfavourable* stimuli, uniform cheerleading can be expected to generate a straightforward overestimate of the disparity between the two beliefs. This is not ideal, of course, but does not represent a *fatal* threat to inference in those studies; primarily because studies that find partisan cheerleading never find *only* that. In other words, effects that are indicative of sincere discrepancies in beliefs still remain after accounting for cheerleading (e.g., see Khanna & Sood, 2018).

In any case, unlike in the aforementioned investigations (Bullock et al., 2015; Khanna & Sood, 2018; Prior et al., 2015), financially incentivizing the reporting of beliefs is impractical or impossible in many of my studies. This is because the beliefs are measured by questions that are not transparently or obviously verifiable. For example, in Studies 3 and 4 in Part 2.2 I ask people how "valid" they consider a psychological test of open-mindedness—to some extent an inherently subjective belief. Hence, paying subjects for correct responses seems a non-starter.

A final point is that—difficulties in implementation aside—from the perspective of ecological validity, financially-incentivizing responses is not necessarily preferable. For example, Kahan (2016b, p. 7) points out that, "In the real world, ordinary members of the public *do not get monetary rewards* for forming 'correct' beliefs about politically contested factual issues" (emphasis in the original). According to his theory, the only material stake individuals have in such beliefs is how those beliefs identify them vis-à-vis particular groups in society—the logic being that expressing the "wrong" belief amongst one's peers can induce material (and psychological) costs in the form of ostracism and exclusion (see also Kahan, 2016a; Kahan,

2017). Given that the large majority of studies presented in Part II concern a direct test of Kahan's theory, one could argue that the lack of financial incentives thus provides a fairer examination of that theory.

Observational and Experimental Research Designs

The studies reported in this thesis comprise a mix of observational and experimental research designs. From the perspective of causal inference, experimental designs are preferable (Gerber & Green, 2012). That is, because random assignment of treatments—provided said assignment meets several assumptions—allows the researcher to infer *causal* relationships between one variable and another. Often, however, random assignment is not feasible (or ethical). For example, in all of the studies in Part 2.2, a key measure that I employ is performance on the Cognitive Reflection Test—which I use as a proxy for individuals' cognitive ability. Clearly, random assignment of cognitive ability is not feasible, and so observational measurement of this variable was necessary. While observational research does not *altogether* preclude causal inference, it massively complicates it (e.g., Rohrer, 2018). Where I use observational research designs, therefore, I am careful to respect this fact—and to resist the tendency to "slip" into using causal language.

As a further remedy, where I use observational designs I try to point out and describe how subsequent research might improve on my design in order to test causal relationships between the variables of interest. For example, in Part 1.3 I use an observational design to test the hypothesis that *moral polarization*—that is, the tendency for people to view co-partisans' moral character positively, and opposing partisans' moral character negatively—predicts behavioural outcomes. In the Discussion section of that paper, I outline how future investigations might investigate whether such a relationship is causal. Specifically, I suggest that one might randomly assign moral characteristics to in-party and out-party members, and measure subsequent behaviour. Likewise, in the Discussion section of Part 2.2, I argue at length that many studies that report evidence of "motivated reasoning" do not randomly assign motivation—leaving the results of such studies open to various confounding factors, such as the prior beliefs of the reasoner (I refer to the Discussion of that paper for greater detail).

Preregistration

All studies that I report in this thesis contain at least one *a priori* hypothesis test. Accordingly, all studies (but one) in this thesis are *preregistered*. That is, the hypothesis, sampling strategy, study design and analysis plan were specified ahead of data collection. The primary benefits of preregistration are the prevention of (i) undisclosed flexibility in data analysis (conscious *or* unconscious) (Nosek et al., 2018; van't Veer & Giner-Sorolla, 2016), and (ii) hypothesizing after the results are known (e.g., Kerr, 1998). Both of these practices violate the assumptions of null hypothesis significance testing; which require that researchers do not condition their analytic strategy on, or derive their hypothesis from, the observed data.

There are now many routes and templates through which researchers can preregister their study designs and hypothesis tests (Nosek et al., 2018). The primary route I use in this thesis is the *Open Science Framework* (https://osf.io/), but I also use *AsPredicted* (https://aspredicted.org/); for shorter protocols, or protocols that were written early in my doctoral studies—before I transitioned to the OSF. In my experience, preregistration caused me to design better, more informative studies, and to think more clearly through each analysis plan—for example, exactly whether and how it provided a diagnostic test of the hypothesis in question. Of course, preregistration protocols can vary in quality, and even the most well-thought-out plans sometimes require deviations—given hindsight—once the data are in. Where such deviations are required, transparency in reporting is essential. As a case in point, in Studies 1-2 in Part 2.2 I mis-specified the random effects structure in a linear mixed effects modeling strategy written into the preregistration. This mis-specification was only discovered after the data were collected and analysed. Thus, in the write-up of these studies I fit and report the models with *both* random effects structures; the preregistered, incorrect structure, and the *correct* structure identified after-the-fact. Fortunately, in this case the results do not change much, but the example illustrates that preregistration is not perfect—and is certainly not a straitjacket to conducting additional (or different) analyses if the preregistered analyses are suboptimal. It is my opinion that preregistration considerably improved the quality of research in this thesis.

Part I: Perceived Moral Superiority and Moral Behaviour

Preface

In Part I, I investigate (i) people's beliefs about their own moral goodness relative to the average person—in particular, I ask whether and to what extent such beliefs are irrational—and (ii) people's beliefs about the moral goodness of their political in-party relative to their political out-party. Using economic games, I subsequently test whether people's beliefs regarding (i) and (ii) correlate with behavioural outcomes. Finally, I test whether people's motivation to "do the morally right thing" underpins their prosocial behaviour in economic games.

Specifically, in Part 1.1 I report a study in which people were asked to rate themselves and the average person on various moral and nonmoral traits. A positive disparity between self-ratings and ratings of the average person has been referred to as "self-enhancement", and such positive disparities are widespread. On this basis, previous researchers have inferred irrationality on the part of self-enhancing individuals. Following others, I argue that this inference is premature because it does not respect the uncertainty inherent in social perception, nor does it attempt to take account of this uncertainty. Adapting a recently-developed method, I attempt to account for this uncertainty and to isolate residual, "irrational" self-enhancement. My central finding is that such irrational self-enhancement is largest in the *moral* domain. However, contrary to influential theory, magnitude of moral self-enhancement was not correlated with self-esteem.

In Part 1.2, I report two studies in which I investigated whether the magnitude of moral self-enhancement identified in Part 1.1 predicts prosocial behaviour. Specifically, whether it predicts behaviours commonly considered moral—like freely helping others, and reciprocating trust. I model these behaviours using economic games; the Dictator Game and Trust Game, respectively. Through a combination of analytic approaches, I find a convincing *lack* of evidence that moral self-enhancement predicts behaviour in these games. However, I find some evidence that individuals' beliefs about their moral goodness *per se*—irrespective of their perceived *superiority* over others—weakly predicts behavioural outcomes. I conclude that moral self-enhancement (superiority) lacks predictive validity because of its dual constituent parts (i.e., self *and* other evaluation).

In Part 1.3, I report two studies in which I extended the investigation of moral superiority to Democratic and Republican partisans in the US. In particular, in these studies I tested whether the disparity in moral evaluation between the in-party and out-party—referred to here as *moral polarization*—predicts behavioural hostility toward the out-party. I model this behaviour using a variant of the Intergroup Prisoner's Dilemma. Though I find that moral polarization *per se* is large—larger than polarization in nonmoral domains of evaluation—I observe *un*convincing evidence that it predicts behavioural expressions of out-party hostility. These results strike an optimistic chord and converge with recent work in political science on the limits of partisan prejudice.

Finally, in Part 1.4 I report a study in which I conducted an improved and extended test of the *morality preference hypothesis*. According to this hypothesis, prosocial behaviour is motivated by people's preference for "doing the morally right thing". I identify important confounds and unresolved questions in recent work that purports to find support for this hypothesis. In my study, I correct these confounds and extend the design to answer the unresolved questions. I convincingly replicate support for the morality preference hypothesis. However, through my extension of the original design I find evidence contrary to popular social psychological theory.

1.1. The Illusion of Moral Superiority[4]

Abstract

Most people strongly believe they are just, virtuous, and moral; yet regard the average person as distinctly less so. This invites accusation of irrationality in moral judgment and perception—but direct evidence of irrationality is absent. Here, we quantify this irrationality, and compare it against the irrationality in other domains of positive self-evaluation. Participants (N=270) judged themselves and the average person on traits reflecting the core dimensions of social perception: morality, agency, and sociability. Adapting new methods, we reveal that virtually all individuals irrationally inflated their moral qualities, and the absolute and relative magnitude of this irrationality was greater than that in the other domains of positive self-evaluation. Inconsistent with prevailing theories of overly positive self-belief, irrational moral superiority was not associated with self-esteem. Taken together, these findings suggest that moral superiority is a uniquely strong and prevalent form of "positive illusion", but the underlying function remains unknown.

[4] The work presented in this section was conducted in collaboration with Ryan McKay (supervisor) and is published in *Social Psychological and Personality Science*: http://journals.sagepub.com/doi/abs/10.1177/1948550616673878

Introduction

Most people believe they are just, virtuous, and moral. These beliefs demand scientific attention for several reasons. For one, in contrast to other domains of positive self-belief, they likely contribute to the severity of human conflict. When opposing sides are convinced of their own righteousness, escalation of violence is more probable, and the odds of resolution are ominously low (Pinker, 2011; Skitka, Bauman, & Sargis, 2005). Moreover, self-righteousness is not confined to conflict situations; a substantial majority of individuals believe themselves to be morally superior to the average person. Compared to "above average" beliefs in other domains (e.g., see Alicke & Govorun, 2005), distinct lines of evidence suggest widespread moral superiority may be particularly irrational—yet direct empirical support for this is absent. In the present study, we quantify this irrationality. We find that moral superiority represents a uniquely strong and prevalent instance of "positive illusion" (Taylor & Brown, 1988).

Moral Superiority

In their seminal review, Taylor and Brown (1988) advanced the case for a triad of positive illusions—the first of which was overly positive self-evaluation. They regarded these phenomena as reflecting inaccuracies in social perception, persisting as a result of their beneficial effect upon human wellbeing. The common means of inferring the presence of positive illusions is to ask individuals how they compare with respect to the average person along some dimension. This method consistently reveals that an implausibly high number of people believe that they are above average—a phenomenon dubbed the "better-than-average effect" (Alicke & Govorun, 2005; or "self-enhancement", Sedikides & Gregg, 2008). Though this phenomenon emerges across a range of characteristics, the magnitude of self-enhancement is strongest for moral qualities.

Across four studies, Alicke and colleagues (2001) reported evidence that desirable moral traits, such as honesty and trustworthiness, are associated with the largest difference between judgments of the self and the average person. A similar pattern has been found for undesirable traits—clearly moralized terms such as "liar" produce the strongest asymmetries in self-other judgment (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995). Multiple studies converge on the same conclusion: magnitude of self-enhancement is stronger for moral

characteristics—like honesty—than for other desirable but non-moral characteristics, such as competence (Brown, 2012; Möller & Savyon, 2003), wisdom (Zell & Alicke, 2011), ambition (Alicke et al., 2001) and intelligence (Van Lange & Sedikides, 1998). Moreover, whereas self-enhancement of various non-moral traits may diminish with age, the self-other asymmetry for moral traits remains consistently large throughout the lifespan (Zell & Alicke, 2011). Such is the extent of this phenomenon that violent criminals consider themselves more moral than law-abiding citizens living in the community (Sedikides, Meek, Alicke, & Taylor, 2014).

The Ubiquity of Virtue

To compound the paradox of widespread moral superiority, most individuals appear assured of the loftiness of their virtuous qualities (relative to their other qualities). Desirable moral traits are perceived to be highly descriptive of oneself—more so than other desirable but non-moral traits. For example, in a vast cross-national sample of 187,957 participants spanning 11 European countries, Gebauer and colleagues (2013) reported that, of two distinct trait dimensions, the one comprising desirable moral terms such as "faithful" and "honest" was judged as more self-descriptive than the one which included non-moral terms such as "clever" and "wise". Additional cross-cultural data converge on the same conclusion. In a similarly sized sample comprising participants from 54 countries and all fifty US states, the moral characteristics of "honesty" and "fairness" were ranked consistently highly in individuals' self-description (Park, Peterson, & Seligman, 2006). Wojciszke and Bialobrzeska (2014) compared two distinct trait dimensions, one including desirable moral characteristics—such as "fair, "honest", and "loyal"—and the other including desirable non-moral characteristics such as "intelligent", "knowledgeable", and "logical". Across six diverse cultures, they found that the traits in the former dimension were judged to be more descriptive of the self (also see Wojciszke, Baryla, Parzuchowski, Szymkow, & Abele, 2011). Indeed, numerous studies have shown that individuals believe they possess, on average, more honesty and trustworthiness than any other characteristic, including intelligence, modesty, friendliness, determination, and independence (e.g., Alicke et al., 2001; Brown, 2012; Sedikides, 1993). Finally, individuals anticipate that, whereas desirable non-moral traits will come and go throughout the course of one's life, they will always possess desirable moral traits (Ybarra, Park, Stanik, & Lee, 2012).

The Paradox

Taken together, the preceding lines of evidence present a striking asymmetry. Most people consider themselves paragons of virtue; yet few individuals perceive this abundance of virtue in others. As a descriptive phenomenon, this pattern is perhaps unsurprising. Previous research indicates that self-enhancement emerges most strongly for traits that are both desirable and ambiguous (e.g., Dunning, Meyerowitz, & Holzberg, 1989)—a product of the increased degrees of freedom for self-favouring construal of the traits in question. That self-enhancement is strongest in the moral domain is directly consistent with this evidence. Morality traits are highly desirable (Van Lange & Sedikides, 1998) yet difficult to check against reality (Alicke & Govorun, 2005), and there is significant variability in the behaviours considered indicative of a "moral" person (Graham, Meindl, Beall, Johnson, & Zhang, 2016).

However, *normatively* speaking, moral superiority may reflect significant incoherence in social judgment and perception. To illustrate why, consider a typical individual, *Jane*, tasked with judging the morality of herself and the average person. The reviewed evidence suggests Jane construes her morality in very positive terms—in part by capitalizing on trait ambiguity. In contrast, her judgment of the average person is decidedly less positive. This suggests that Jane foregoes the corollary that high trait ambiguity permits a majority of others to be equally as moral as she, albeit in their own idiosyncratic ways (Dunning et al., 1989). Unfortunately, Jane's double standard incurs a cost to her judgment accuracy. Self-judgments act as valid cues to what the average person is like—justified by the fact that most people are in the majority most of the time. Indeed, appropriately gauging the prototypicality of one's own characteristics improves accuracy in judgments of ill-defined others (e.g., Hoch, 1987; Krueger & Chen, 2014); neglecting this prototypicality may thus amount to a failure of inductive reasoning (Krueger, Freestone, & MacInnis, 2013). Consequently, given that most people consider themselves highly moral, if Jane strongly self-enhances her morality—as the evidence indicates she will—this may also compromise the accuracy of her social perception.

The Present Study

There is mounting support for the idea that moral superiority is an especially potent positive illusion. However, the term "illusion" specifically implies irrationality in belief—an accusation that lacks decisive evidence (Krueger & Wright, 2011). Prevailing measures of self-enhancement do not discriminate between the rational (i.e., defensible) and irrational (indefensible) components of self-enhancement (Heck & Krueger, 2015). To our knowledge,

there have been no attempts to quantify and compare the irrationality in moral self-enhancement with that in other domains of self-enhancement. The present study addresses this lacuna. We adapt a novel method (Heck & Krueger, 2015) to isolate and quantify the irrational component of moral superiority, and compare it against the irrationality in other domains of self-enhancement. We also examine whether the irrational component of moral superiority is associated with wellbeing, as the prevailing conception of positive illusions (Taylor & Brown, 1988), and previous research (e.g., Campbell, Rudich, & Sedikides, 2002), would predict.

Methods

Participants

We sought to recruit 265 participants via Amazon's Mechanical Turk (www.mturk.com) to achieve greater than 90% power to detect a small effect of d=0.2 (at $\alpha = .05$) in our primary analyses of variance and paired samples t-tests. We over-recruited by 15% to account for data exclusions; bringing our collected sample size to 308 participants (153 male; $M_{age} = 37.81$, SD = 11.77). Data from 20 participants were excluded from all analyses due to failing at least one attention check (8 participants) and/or providing incomplete responses (15 participants). A further 18 participants were excluded from the final regression analyses due to lack of variation in judgments of the self, the average person, and/or trait desirability—leaving a sample size of 270 for the primary analyses.

Procedure & Materials

Procedure. After providing online consent, participants were presented with a list of 30 traits, comprising ten trait terms each for the dimensions' morality, agency, and sociability. They were asked to judge the extent to which each trait described (a) themselves, (b) the average person, and (c) the social desirability of each trait. Participants rated all 30 traits according to either (a), (b), or (c), before moving onto the next set of ratings, and the order of these three sets of judgments was counterbalanced across participants (any order effects were presumed to be trivial). The presentation order of the traits themselves was randomized across each rating set and participant. Rating judgments for the self and the average person were provided on a scale from 1 (*Not at all*) to 7 (*Very much so*). Social desirability judgments were also provided

on a seven-point scale, ranging from -3 (*Very undesirable*) to +3 (*Very desirable*). Following the trait judgments, participants completed four other measures (counterbalanced, detailed below) and provided simple demographic information.

Traits. The core dimensions of social perception, *communion/warmth* and *agency/competence*, are associated with traits related to benevolence and ability, respectively (Fiske, Cuddy, & Glick, 2007). Recently it has been empirically demonstrated that the communion/warmth dimension is comprised of distinct *morality* and *sociability* components—the former describing honesty, trustworthiness, and sincerity, the latter warmth, friendliness, and likeability (e.g., Goodwin, Piazza, & Rozin, 2014). Thus, drawing upon a comprehensive norming study of trait adjectives (Goodwin et al., 2014, Experiment 1), we selected 5 positive (desirable) and 5 negative (undesirable) traits for each of the three dimensions' morality, agency, and sociability; providing a total of 30 traits for the present study. Traits were carefully chosen to minimize dimension overlap (see the Supplemental Material for details of the trait selection procedure).

Other Measures. Self-esteem was measured using the 10-item Rosenberg Self-Esteem scale (Rosenberg, 1965). Three additional measures were included, but were not part of the primary analyses and are thus not reported further in the main text (see the Supplemental Material for relevant analyses). These were the 16-item Narcissistic Personality Inventory (Ames, Rose, & Anderson, 2006), and the Moral Identity (Aquino & Reed, 2002) and Need-To-Belong (Leary, Kelly, Cottrell, & Schreindorfer, 2013) scales.

Projection-Based Index of Self-Enhancement

Prevailing measures of self-enhancement conflate defensible (or "rational") self-enhancement with indefensible (or "irrational") self-enhancement (Krueger & Wright, 2011). For instance, given that individuals have more information about themselves than about others, they will be relatively less certain about what the average person is like. As a consequence, judgments of the average person are likely to be less extreme than self-judgments; with the former tending towards the midpoint of the judgment scale (Moore & Small, 2007). The corollary is that observed self-other differences in trait judgment—ostensibly indicative of self-enhancement—may actually reflect rationally cautious judgments made under uncertainty. In order to estimate the irrational component of self-enhancement, it is therefore necessary to first account for this

rational component. To this end we adapted the Social Projection Index of self-enhancement (SPI, Heck & Krueger, 2015).

To estimate what proportion of conventional self-enhancement may be considered rational, the SPI first asks how an individual might infer the characteristics of the average person. One strategy is to draw upon a relative abundance of self-knowledge. Indeed, in the absence of salient diagnostic information about others, one's own characteristics act as cues to what others are like. Decades of research has shown that individuals readily project their own characteristics onto others, and that this process—termed social projection—typically increases accuracy in judgments of what unknown others are like (for reviews see Krueger, 2007; Robbins & Krueger, 2005). However, projection may be too weak or too strong—individuals may under-perceive or over-perceive the similarity between themselves and others, respectively. Somewhere in between is the optimal amount of projection, which tracks the *actual* similarity among people. While individuals are unlikely to perceive this similarity with complete precision, the researcher can. The similarity may be quantified as the correlation between individual self-judgments and the average of all self-judgments in the group (Hoch, 1987; Krueger et al., 2013). This correlation "coefficient of similarity" thus describes how typical of the average a particular individual is, and, importantly, a fully rational perceiver may weight their self-judgments by this coefficient to maximize accuracy in their judgments of what the average person is like (Hoch, 1987; Krueger & Chen, 2014). Computation of the coefficients of similarity therefore provides a rational benchmark against which to evaluate the observed self-enhancement of individuals (i.e., the difference between their self-judgments and their judgments of the average person).

To illustrate, consider the following example. An individual whose self-judgments are highly typical of the average should project more, as their self-judgments are highly diagnostic of what the average person is like. In contrast, an individual whose self-judgments are highly *a*typical of the average should project less, as their self-judgments are only weakly diagnostic of what the average person is like. In the former case, judgments of the average person are expected to be minimally regressive with respect to self-judgments—resulting in a smaller latitude for defensible (rational) self-enhancement. In contrast, in the latter case, judgments of the average person are expected to be relatively more regressive with respect to self-judgments—resulting in a larger latitude for defensible (rational) self-enhancement. Crucially, in either case, the SPI explicitly models the fact that informational uncertainty mandates that a

proportion of conventional self-enhancement be considered rational—avoiding the pitfall of earlier measures.

Computationally speaking, the logic outlined above allows researchers to generate rational *predicted judgments of the average person*, by weighting individuals' self-judgments by their respective coefficient of similarity. As Heck and Krueger (2015) point out however, this conceptualizes self-enhancement as diminishment of others. Alternatively, it is possible to reverse-predict what self-judgments *should* have been, if rationally projecting individuals derived their empirically observed judgments of the average person from their self-judgments. These reverse-predicted self-judgements may be labelled *inferred self-judgments*, and they yield a more conceptually appropriate interpretation of self-enhancement as positive self-inflation.

Determining the rational and irrational components of conventional self-enhancement thus requires self-judgments, judgments of the average person, and computed inferred self-judgments. Then, in a final step, the SPI exploits the empirical observation that most individuals possess a positive self-image; ascribing positive traits more readily to themselves than to others (and vice versa for negative traits). Accordingly, self-enhancement is modelled as the relationship between trait desirability and trait judgment[5], meaning that (a) conventional self-enhancement is given as the difference between how well trait desirability predicts self-judgments compared to how well it predicts judgments of the average person; (b) the rational component of self-enhancement is the difference between how well trait desirability predicts inferred self-judgments compared to how well it predicts judgments of the average person; and, finally, (c) the irrational component of self-enhancement is the difference between how well trait desirability predicts inferred self-judgments compared to how well it predicts actual self-judgments.

Results

Descriptives

[5] This index of self-enhancement is psychometrically equivalent to a conventional difference-score index (see Heck & Krueger, 2015).

Table 1 displays the list of 30 traits, their mean self (S), average person (or "Other", O), and desirability (D) judgments, as well as the respective domain reliability coefficients. Table 2 displays the zero-order correlations among mean self, other and desirability judgments for each trait domain.

Table 1. Mean Self, Other and Desirability Trait Judgments and Domain Reliability Coefficients.

Trait	Self	Other	Desirability
Agency	-	-	-
Hard-working	5.71 (1.24)	4.44 (1.09)	6.54 (0.77)
Knowledgeable	5.66 (1.03)	4.27 (1.08)	6.37 (0.91)
Competent	5.89 (1.04)	4.49 (1.07)	6.48 (0.89)
Creative	4.87 (1.63)	3.94 (1.11)	5.90 (1.02)
Determined	5.66 (1.32)	4.54 (1.16)	6.18 (0.94)
Lazy	2.59 (1.54)	3.56 (1.29)	1.64 (1.04)
Undedicated	2.08 (1.36)	3.20 (1.21)	1.63 (0.91)
Unintelligent	1.63 (1.00)	3.34 (1.26)	1.56 (0.93)
Unmotivated	2.42 (1.50)	3.30 (1.22)	1.57 (0.93)
Illogical	2.02 (1.24)	3.56 (1.40)	1.72 (1.11)
M:	3.85 (1.83)	3.87 (0.53)	3.96 (2.47)
Reliability (α):	.88	.93	.88
Sociability	-	-	-
Sociable	4.31 (1.72)	4.99 (0.91)	6.25 (0.94)
Cooperative	5.50 (1.29)	4.64 (1.09)	6.41 (0.79)
Warm	5.13 (1.49)	4.48 (1.09)	6.41 (0.85)
Family-orientated	4.98 (1.89)	4.83 (1.12)	5.87 (1.17)
Easy-going	5.31 (1.45)	4.31 (1.02)	6.01 (1.02)
Cold	2.54 (1.57)	3.13 (1.15)	1.60 (1.02)
Disagreeable	2.38 (1.34)	3.32 (1.24)	1.49 (0.91)
Rude	2.14 (1.33)	3.34 (1.28)	1.29 (0.75)
Humorless	1.88 (1.24)	3.01 (1.12)	1.68 (1.02)
Uptight	2.47 (1.44)	3.47 (1.23)	1.83 (1.10)
M:	3.66 (1.50)	3.95 (0.77)	3.88 (2.44)
Reliability (α):	.89	.88	.82
Morality	-	-	-
Honest	5.93 (1.06)	4.44 (1.19)	6.55 (0.81)
Trustworthy	6.10 (0.98)	4.30 (1.26)	6.67 (0.78)
Fair	5.94 (0.99)	4.51 (1.13)	6.51 (0.85)
Respectful	5.88 (1.12)	4.55 (1.15)	6.52 (0.75)
Principled	5.63 (1.23)	4.26 (1.15)	6.16 (0.98)
Insincere	1.80 (1.02)	3.32 (1.31)	1.49 (0.90)
Prejudiced	2.12 (1.32)	3.78 (1.38)	1.51 (1.00)
Disloyal	1.65 (0.89)	3.06 (1.23)	1.31 (0.69)
Manipulative	2.10 (1.26)	3.39 (1.28)	1.60 (1.06)
Deceptive	2.07 (1.30)	3.34 (1.29)	1.44 (0.85)
M:	3.92 (2.09)	3.89 (0.58)	3.98 (2.65)
Reliability (α):	.88	.93	.88
M (total):	3.81 (1.76)	3.90 (0.61)	3.94 (2.43)

Note: For desirability judgments, the -3 to +3 scale was converted to 1 to 7. Standard deviations are given in parentheses. N=288.

Table 2. Zero-Order Correlations among Mean Self, Other and Desirability Judgments for Each Trait Domain.

Mean judgment	1	2	3	4	5	6	7	8	9
1. Agency, self	-								
2. Agency, other	.30	-							
3. Agency, desirability	.39	.19	-						
4. Sociability, self	.63	.47	.30	-					
5. Sociability, other	.24	.83	.15	.38	-				
6. Sociability, desirability	.36	.20	.66	.40	.21	-			
7. Morality, self	.65	.32	.38	.69	.27	.43	-		
8. Morality, other	.23	.87	.16	.43	.89	.20	.30	-	
9. Morality, desirability	.39	.20	.74	.35	.16	.74	.46	.20	-

Note: For meaningful interpretation, the coefficients are based upon means calculated after reverse-coding negative traits. All ps<.05. N=288.

Rational and Irrational Self-Enhancement

To compute the rational and irrational components of self-enhancement, we first calculated the similarity between individual self-judgments and the average of all self-judgments in the group (i.e., the coefficients of similarity for each participant). Thus, for each participant, we regressed the average self-judgments made by all participants for the traits in a given dimension onto the self-judgments of the focal participant for the traits in that dimension. Estimating unique coefficients for each dimension acknowledges that individuals may be more similar on some dimensions compared to others. We thus obtained three coefficients of similarity and their corresponding intercepts for each participant. As outlined above, the SPI posits that these coefficients may be used to weight S judgments to generate rational predicted O judgments, *P*. P is thus given as:

$$P = coefficient\ of\ similarity * S\ judgment + intercept$$

However, following Heck and Krueger (2015), we instead computed inferred self-judgments, *I*, by rewriting the regression equation:

$$I = \frac{O\ judgment}{coefficient\ of\ similarity} + intercept$$

We computed I judgments for each trait over all participants, using the mean[6] coefficient of similarity and intercept corresponding to the traits' respective dimension:

$$I = \frac{O}{0.85} + 0.61 \quad [Morality\ traits]$$

$$I = \frac{O}{0.73} + 1.04 \quad [Agency\ traits]$$

$$I = \frac{O}{0.52} + 1.77 \quad [Sociability\ traits]$$

Thus, at this stage, each participant had four sets of judgments for the 30 traits: their empirically observed S, O, and D judgments, and the new I judgments computed according to the method outline above. For each dimension, we then regressed S, O, and I on D judgments over each participant. We also regressed O on S judgments to cross-check the assumption of social projection (a positive association constitutes evidence for social projection; Krueger, 2007). Table 3 displays the relevant means. We first draw attention to the evidence of social projection; across all trait dimensions, S judgments positively predicted O judgments (*mean* b_{SO} = .23-.30). We then examined whether social projection increased judgment accuracy. For each dimension, we computed an accuracy index by correlating other judgments with average self-judgments over all participants and traits corresponding to that dimension. Correlating this index with magnitude of social projection revealed the expected pattern. Accuracy of other judgments was positively associated with social projection across all trait dimensions, r (268) = .63-.85, all ps<.001. As projection increased, accuracy of other judgments improved. This is

[6] While a more fine-grained approach is to generate inferred-self judgments using the participants' idiosyncratic coefficients of similarity—rather than the mean—these different approaches yield equivalent results overall (see e.g., Heck & Krueger, 2015).

consistent with previous research (Hoch, 1987; Krueger & Chen, 2014) and confirms the validity of the methodological approach taken in the present study.

Table 3. Mean Slopes and Intercepts from Primary Regression Analyses.

	Unstandardized				Standardized		
	Slope (b)		Intercept		Beta	95% CI	
	M	SE	M	SE	M	LL	UL
Agency							
R_{SD}	0.70	0.02	1.06	0.09	.80	0.76	0.83
R_{OD}	0.21	0.02	3.04	0.10	.38	0.30	0.45
R_{SO}	0.23	0.02	3.01	0.10	.36	0.29	0.44
R_{ID}	0.28	0.03	5.17	0.14	.38	0.30	0.45
Sociability							
R_{SD}	0.56	0.02	1.48	0.09	.67	0.61	0.72
R_{OD}	0.30	0.02	2.78	0.09	.55	0.49	0.61
R_{SO}	0.30	0.02	2.81	0.10	.45	0.39	0.51
R_{ID}	0.57	0.04	7.13	0.17	.55	0.49	0.61
Morality							
R_{SD}	0.76	0.02	0.92	0.07	.88	0.85	0.90
R_{OD}	0.21	0.02	3.02	0.09	.41	0.33	0.48
R_{SO}	0.25	0.02	2.90	0.10	.41	0.34	0.48
R_{ID}	0.25	0.03	4.17	0.11	.41	0.33	0.48

Note: R = regression; S = self; O = other; D = desirability; I = inferred-self. SE = standard error; CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit. Unstandardized slopes (b) are used in analyses. N=270.

Next, we note the expected observation of conventional self-enhancement. Across all dimensions, trait desirability predicted self-judgments (*mean* b_{SD}) better than it predicted other judgments (*mean* b_{OD}); for morality (.76 vs .21), t (269) = 22.08, p<.001, d=1.34 95% CI [1.18, 1.51], agency (.70 vs .21), t (269) = 18.75, p<.001, d=1.14 [0.99, 1.29], and sociability (.56 vs .30), t (269) = 10.21, p<.001, d=0.62 [0.49, 0.75]. To compare across dimensions, we computed a difference measure of conventional self-enhancement as $b_{SD} - b_{OD}$ and conducted a repeated measures analysis of variance with Dimension as the single factor: F (2, 538) = 71.70, p<.001,

η^2 = .21. Conventional self-enhancement was greater for morality (.54) than agency (.49), t (269) = 2.01, p=.045, d=0.12 [0.00, 0.24], and sociability (.26), t (269) = 11.04, p<.001, d=0.67 [0.54, 0.80]. Agency was also greater than sociability, t (269) = 8.42, p<.001, d=0.51 [0.39, 0.64].

What proportion of conventional self-enhancement is accounted for by rational and irrational components? To determine the rational component, we examined how well trait desirability predicted inferred-self judgments (*mean* b_{ID}) compared to other judgments (*mean* b_{OD}). Sociability had the largest magnitude of rationally defensible self-enhancement: .57 vs .30, t (269) = 14.33, p<.001, d=0.87 [0.73, 1.01]. The magnitude for agency was substantially smaller (.28 vs. 21), but still non-trivial, t (269) = 3.19, p=.002, d=0.19 [0.07, 0.31]. In contrast, the rational component of self-enhancement in the moral domain was trivial in size; .25 vs .21, t (269) = 1.72, p=.087, d=0.10 [-0.02, 0.22].

This indicates that irrational self-enhancement is strongest in the moral domain. Indeed, this was the case. Examining how well trait desirability predicted actual self-judgments (*mean* b_{SD}) compared to inferred-self judgments (*mean* b_{ID}) revealed the largest discrepancy for morality; .76 vs .25, t (269) = 18.17, p<.001, d=1.11 [0.95, 1.26]. Though smaller, agency comprised a substantial magnitude of irrational self-enhancement; .70 vs .28, t (269) = 12.97, p<.001, d=0.79 [0.65, 0.93]. In stark contrast, there was no evidence for irrationality in self-enhancement of sociability traits—desirability predicted self and inferred-self judgments equally well; .56 vs .57, t (269) = -0.32, p=.750, d=-0.02 [-0.14, 0.10]. In other words, the average magnitude of conventional self-enhancement along the dimension of sociability was fully accounted for by rational projection-based other judgment. As before, we compared irrationality across dimensions by computing a difference measure, $b_{SD} - b_{ID}$, and conducting a repeated measures analysis of variance with Dimension as the single factor: F (2, 538) = 187.19, p<.001, η^2=.41. Morality (.50) comprised the greatest magnitude of irrational self-enhancement, compared with agency (.42), t (269) = 3.33, p=.001, d=0.20 [0.08, 0.32], and sociability (-.01), t (269) = 18.04, p<.001, d=1.10 [0.95, 1.25]. Agency was also greater than sociability, t (269) = 13.25, p<.001, d=0.81 [0.67, 0.94]. The complete pattern is displayed in Figure 1 as the percentage of conventional self-enhancement magnitude accounted for by rational and irrational components.

Corroborating the analysis of magnitude, the data also show that more individuals irrationally self-enhanced ($b_{SD} > b_{ID}$) for moral traits (243, 90% of sample), compared with agency traits (218, 81%), χ^2 (270) = 14.05, p<.001, and sociability traits (134, 50%), χ^2 (270) = 103.22, p<.001. Individuals were also more likely to irrationally self-enhance for agency traits than for sociability traits, χ^2 (270) = 73.29, p<.001.

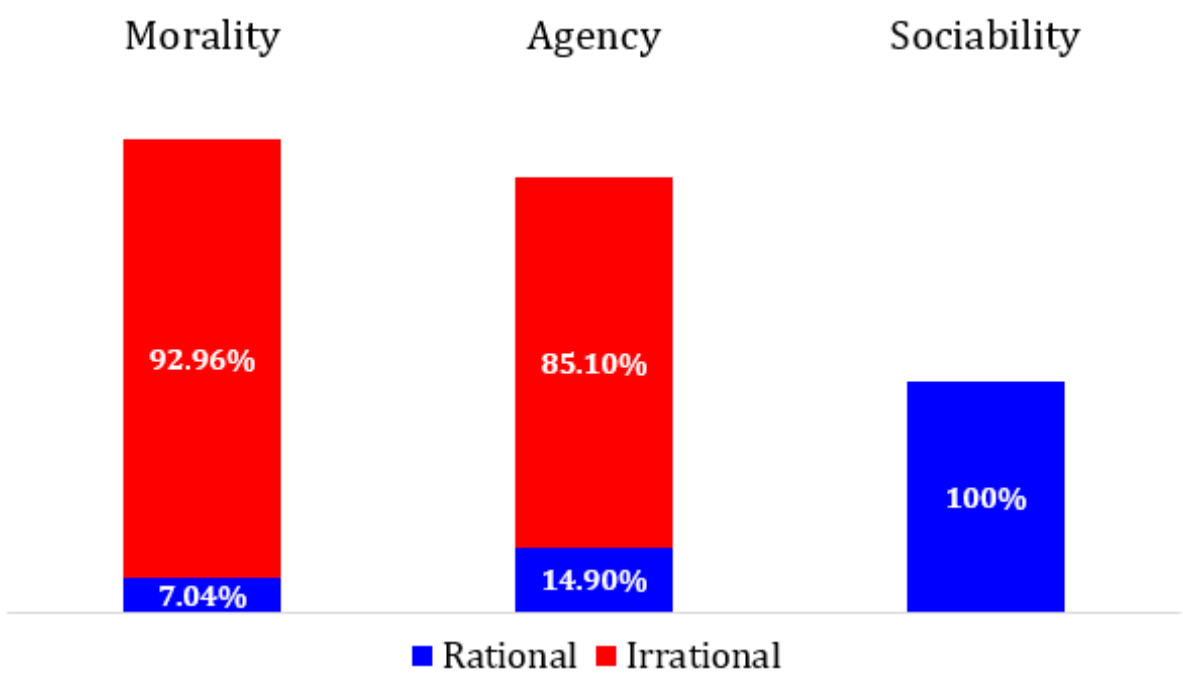


Figure 1. *Percentage of conventional self-enhancement ($b_{SD} – b_{OD}$) magnitude accounted for by rational ($b_{ID} – b_{OD}$) and irrational ($b_{SD} – b_{ID}$) components as a function of trait dimension. N=270.*

Ubiquity of Virtue

What accounts for the strength and prevalence of irrationality in moral self-enhancement? Consistent with the *ubiquity of virtue*, prototypicality of individual self-judgments was highest in the moral domain—self-judgments tracked average self-judgments better for morality traits (*mean b* = .85) than agency traits (.73), t (269) = 5.96, p<.001, d=0.36 [0.24, 0.49], and

sociability traits (.52), t (269) = 12.94, p<.001, d=0.79 [0.65, 0.92]. In other words, as expected, the strongest consensus in self-judgment emerged in the moral domain. Importantly however, strength of projection to others was not adequately adjusted to reflect this consensus. Projection from self to other was no stronger in the moral domain (*mean* b_{SO} = .25) than in the agency domain (.23), t (269) = 1.51, p=.133, d= 0.09 [-0.03, 0.21], and was in fact *weaker* than projection in the domain of sociability (.30), t (269) = 3.07, p=.002, d=0.19 [0.07, 0.31]. Thus, trait desirability predicted moral self-judgments to a much greater extent than it predicted moral other-judgments, and, taken in conjunction with the ubiquity of virtue, this discrepancy is classified as largely irrational.

Inaccuracy and Wellbeing

The prevailing conception of positive illusions (Taylor & Brown, 1988) encompasses two central claims about strongly positive self-evaluations. The first is that they reflect an inaccurate perception of reality (i), and the second is that this inaccuracy contributes to wellbeing (ii).

To examine (i), for each dimension we correlated the previously-computed accuracy index with magnitude of irrational self-enhancement in that dimension. As expected, accuracy was negatively associated with irrational self-enhancement; for morality, r (268) = -.73, p<.001, agency, r (268) = -.71, p<.001, and sociability r (268) = -.65, p<.001. The aforementioned accuracy index denotes *discrimination* accuracy only; thus, we also examined *bias*—i.e., absolute discrepancies in judgment (Epley & Dunning, 2006). For each dimension, we computed a discrepancy index of accuracy as the average difference between other judgments and the average of all self-judgments over traits and participants. We entered these scores into a repeated measures analysis of variance with Dimension as the single factor: F (2, 538) = 122.67, p<.001, η^2=.31. Judgments of others' morality (1.56) were the most discrepant; compared with agency (1.40), t (269) = 5.55, p<.001, d=0.34 [0.22, 0.46], and sociability (1.10), t (269) = 15.11, p<.001, d=0.92 [0.78, 1.06], judgments. The most inaccurate judgments of others occurred in the domain with the strongest irrationality in self-enhancement. Thus, both accuracy analyses support claim (i).

Moving onto (ii), we conducted partial correlations between irrational self-enhancement and self-esteem—controlling for the confounding influence of the corresponding rational, and the

other dimensions' rational and irrational, components of self-enhancement. Magnitude of irrational self-enhancement in the moral domain was *not* associated with self-esteem; r (263) = -.02, p=.701, whereas irrational self-enhancement in the agency and sociability domains was positively correlated with self-esteem; r (263) = .30, p<.001, and r (263) = .25, p<.001, respectively.

Discussion

The present study revealed two key findings. The first was that moral superiority comprised a substantial irrational component; the absolute and relative magnitude of which was greater than that observed in other domains of self-enhancement. Indeed, virtually all individuals irrationally inflated their moral qualities. The second key finding was that, unlike the other domains of self-enhancement, irrational moral superiority was not associated with self-esteem. Taken together, these results suggest a uniquely strong and prevalent illusion of moral superiority, and raise intriguing questions about the function of this phenomenon.

The irrationality of moral superiority was borne out of the ubiquity of virtue—almost everyone reported a strong positive moral self-image—and individuals' ignorance of this ubiquity when making judgments of the average person. Indeed, neglecting the prototypicality—and thus cue validity—of one's own self-judgments may signal an error in inductive reasoning (Krueger et al., 2013). Of course, self-judgments themselves may not accurately reflect genuine moral character—for example, compared to behaviour (Back & Vazire, 2012). However, given the substantial degrees of freedom in what constitutes "moral" behaviour (Alicke & Govorun, 2005; Graham et al., 2016), it seems probable that claims of positive moral character are equally legitimate (or illegitimate) for a large majority of people. In most cases, it would be difficult to make the argument that one moral self-image is more genuine than another. A fallacy thus arises when individuals do not apply to others the same degrees of freedom they invoke in their moral evaluation of themselves (Dunning et al., 1989). Insofar as this fallacy compromises the accuracy of social judgment and perception, it may be deemed erroneous (Heck & Krueger, 2015).

Despite finding strong support for the illusory nature of moral superiority, we found that the irrational component of moral self-enhancement was not correlated with self-esteem. This is

inconsistent with the prevailing conception of positive illusions (Taylor & Brown, 1988), and is especially pronounced given that self-esteem was positively associated with magnitude of irrational superiority in both the agency and sociability domains. Furthermore, our result is at odds with previous evidence that high self-esteem individuals possess a stronger belief in moral superiority (Campbell et al., 2002). However, the latter inconsistency may be accounted for by measurement differences. Campbell and colleagues assessed superiority using a "comparative" measure; that is, they directly asked individuals how much better than average they are. These measures correlate most strongly with self-judgments, and only weakly with judgments of others (Klar & Giladi, 1999; Krueger & Wright, 2011)—the corollary being that the measure used by Campbell and colleagues may have assessed (absolute) moral self-image rather than (relative) moral superiority. Support for this proposition is recovered from our own data; self-esteem *did* positively correlate with moral self-image (*morality* b_{SD}), $r(268) = .34$, $p<.001$.

As an indicator of wellbeing, self-reported self-esteem is far from exhaustive; it is necessary to measure wellbeing by more objective means to decisively test the theory of positive illusions (Heck & Krueger, 2015). Nevertheless, the lack of a relationship between self-esteem and the irrational component of moral superiority invites speculation as to why this illusion is so pervasive (cf. Taylor & Brown, 1988). Though a full discussion is beyond the scope of this article, we note that, from other perspectives, moral superiority may not be considered irrational at all (cf. Boudry, Vlerick & McKay, 2015). For example, error management theorists (e.g., Haselton & Buss, 2000) might view underestimating the morality of others as quite rational. Mistaking another person as trustworthy, when in fact they are not, may be associated with greater fitness costs than the reverse error. Under such conditions, individuals may tolerate decreased judgment accuracy for gains made elsewhere (Fetchenhauer & Dunning, 2006; but see McKay & Efferson, 2010). On this account, moral superiority may persist, in part, as a function of the adaptive value of presuming modest morality in unknown others.

The findings of the present study are limited in that they do not reveal the behavioural consequences of an illusion of moral superiority. While we advance the case that moral superiority is dubious partly because "morality" may be defined by many different behaviours (Alicke & Govorun, 2005; Graham et al., 2016), it would be practically useful to know whether the illusion of moral superiority predicts certain types of moral behaviour—for example, dishonesty for monetary gain. On the basis of existing research there is scope for competing predictions. Given the evidence that affirmation of moral image "licenses" subsequent immoral

behaviour (Blanken, van de Ven, & Zeelenberg, 2015), feeling morally superior may promote greater dishonesty. Alternatively, to the extent that people value belief-behaviour consistency (Festinger, 1962), moral superiority may be associated with a greater likelihood of honest behaviour. We defer to future research to test these hypotheses.

The belief that one is morally superior to the average person appears robust and widespread. Our examination of this belief revealed substantial irrationality; beyond that observed in other domains of positive self-evaluation. On this basis, moral superiority represents a uniquely strong and prevalent form of positive illusion.

Supplemental Material

Methods

Procedure & Materials

Traits. Goodwin et al. (2014, Experiment 1) asked 1,048 respondents how useful trait adjectives (170 total) were in providing information about higher-level person characteristics—such as "ability", "morality", and "character". Participants also rated the valence of the traits. For the present study, we selected traits from this dataset according to the following procedure. First, we averaged across relevant characteristics to obtain composite mean ratings for *morality* ("morality/immorality", "character"), *agency* ("ability", "agency"), and *sociability* ("warmth", "communion") dimensions. Next, for each dimension we identified 10 traits with high composite mean ratings for that dimension; 5 positively valenced (i.e., desirable) and 5 negatively valenced (undesirable). We selected both desirable and undesirable traits for the present study because the regression-based index of self-enhancement that we adapted (outlined in the main text) model's self-enhancement as the relationship between trait desirability and trait judgment. Thus, including desirable and undesirable traits generates the required variability in trait desirability judgments. We avoided traits that had a strong overlap across dimensions—for example, "compassionate" garnered an equally high composite rating for *morality* (7.55) and *sociability* (7.71), and was therefore not selected for either dimension. We also avoided direct antonyms (e.g., inclusion of both "honest" and "dishonest") and synonyms (e.g., "unintelligent", "stupid"). Two traits, "manipulative" and "deceptive", were not included in Goodwin et al. (2014), but were added to our final list of morality traits.

Results

Self-Centrality Breeds Self-Enhancement

Gebauer and colleagues (2013) reported evidence that domain self-centrality predicts stronger magnitude of conventional self-enhancement in that domain. Thus, we explored whether the self-centrality of a given trait domain would predict greater magnitude of *irrational* self-enhancement in that domain. Based upon the rationale behind the constructs, we included the

Narcissistic Personality Inventory (NPI-16, Ames et al., 2006), Need-to-belong (NTB, Leary et al., 2013) and Moral Identity (Aquino & Reed, 2002) scales as measures of the self-centrality of agency, sociability, and morality, respectively. All scale items loaded onto their predicted number of factors (all factor loadings ≥.36.), and the scales demonstrated acceptable reliabilities (α's = .86-.92). We thus computed sum scores for each scale following the authors' respective instructions.

We then conducted partial correlations[7] between construct scores and their respective irrational self-enhancement domain. Neither the *internalization* subscale nor overall moral identity scores were related to magnitude of irrational self-enhancement in the moral domain, r (263) = .08, p=.174, and r (263) = -.08, p=.208, respectively. Interestingly, the *symbolization* subscale demonstrated a small negative association with irrational moral superiority; r (263) = -.17, p=.005. We also found that scores on the NPI positively correlated with irrational self-enhancement in the domain of agency; r (263) = .35, p<.001. This is consistent with previous research which reported that narcissists most strongly self-enhance their agency characteristics (e.g., Campbell et al., 2002). Finally, irrational self-enhancement of sociability traits did not relate to NTB scores, r (263) = .03, p=.645. Thus, we found limited evidence that self-centrality—as construed by the included construct scales—breeds irrational self-enhancement.

[7] Controlling for the corresponding rational, and other dimensions' rational and irrational, components of self-enhancement.

1.2. Investigating the Relationship between Self-Perceived Moral Superiority and Moral Behaviour Using Economic Games[8]

Abstract

Most people report that they are superior to the average person on various moral traits. The psychological causes and social consequences of this phenomenon have received considerable empirical attention. The behavioral correlates of self-perceived moral superiority, however, remain unknown. We present the results of two preregistered studies (Study 1, N=827; Study 2, N=825) in which we indirectly assessed participants' self-perceived moral superiority, and used two incentivized economic games to measure their engagement in moral behavior. Across studies, self-perceived moral superiority was unrelated to trust in others and to trustworthiness, as measured by the Trust Game; and unrelated to fairness, as measured by the Dictator Game. This pattern of findings was robust to a range of analyses, and, in both studies, Bayesian analyses indicated moderate support for the null over the alternative hypotheses. We interpret and discuss these findings, and highlight interesting avenues for future research on this topic.

[8] The work presented in this section was conducted in collaboration with Ryan McKay (supervisor) and is published in *Social Psychological and Personality Science*: http://journals.sagepub.com/doi/full/10.1177/1948550617750736

Introduction

Self-perceptions of moral superiority appear robust and relatively widespread. In numerous studies, majorities of people rate themselves as fairer, more trustworthy, more honest—more *moral*—than the average person (Epley & Dunning, 2000; Fetchenhauer & Dunning, 2006; Klein & Epley, 2016, 2017; Tappin & McKay, 2017; Van Lange & Sedikides, 1998). Under the broader phenomenon of "self-enhancement" (Alicke & Sedikides, 2011), past work has investigated (i) psychological explanations for (Sedikides et al., 2014; Tappin & McKay, 2017; Van Lange & Sedikides, 1998), and (ii) interpersonal consequences of (Barranti et al., 2016; Heck & Krueger, 2016) self-perceived moral superiority. There is a conspicuous lack of evidence, however, for how these perceptions relate to engagement in behaviors commonly considered moral—such as freely helping others, or reciprocating trust. In the present article, we report an initial investigation of this relationship.

Self-Perceived Moral Superiority and Engagement in Moral Behavior

There exists much debate over whether the prevalence of self-superiority phenomena is best explained by motivational or non-motivational processes (Brown, 2012; Chambers & Windschitl, 2004; Taylor & Brown, 1988). This offers a useful framework for speculating on how self-perceived moral superiority may relate to engagement in moral behavior.

Consider people who perceive themselves to be *strongly* morally superior to the average person. As a function of their strong sense of righteousness relative to other people, these individuals may be motivated to behave in (moral) ways to protect this positive social comparison. According to various reviews, self-protection is a fundamental human motivation (Sedikides et al., 2015), and social comparison a common process by which people derive positive self-evaluation (Sedikides & Strube, 1997; Wills, 1981). Moral traits, moreover, are held in high regard (Van Lange & Sedikides, 1998), and morality appears to be central to notions of identity (Strohminger & Nichols, 2014, 2015). Individuals who possess a *weaker* sense of righteousness over the average person, then, may accordingly possess a relatively weaker motivation to protect the (less positive) social comparison. This implies that self-perceived moral superiority may be positively associated with engagement in moral behavior.

Another motivational process that might predict a positive association is sensitivity to the charge of hypocrisy. Hypocrites are loathed—more so than people who are honest about their moral limitations (Jordan et al., 2017)—and especially so when the hypocrite considers themselves to be superior to others (Alicke et al., 2013). Heck and Krueger (2016) recently reported evidence that agents who made inaccurate claims of moral self-superiority received the strongest moral condemnation from observers; stronger, even, than agents who *accurately* reported being *less* moral than the average person. Put another way, observers punished people most when their self-reported moral superiority was shown to be false by their behavior. These findings imply added motivation for such people to behave morally, so to avoid harsh social censure. Consistent with this suggestion is evidence that individuals behave more prosocially after criticizing another person (Simpson et al., 2013).

Some non-motivational processes, on the other hand, may lead us to expect a negative association between self-perceived moral superiority and engagement in moral behavior. Fetchenhauer and Dunning (2009, 2010) provide evidence that individuals underestimate the moral goodness (specifically, trustworthiness) of other people due to an informational asymmetry in the social environment. If person A decides to trust person B, this occasionally results in surprising and costly betrayal by person B. In contrast, when person A decides *not* to trust person B, this necessarily *precludes* person A learning that person B was, in fact, trustworthy. The implication is that individuals learn asymmetrically about the trustworthiness of other people; an asymmetry which may underlie cynicism about the moral goodness of others more generally (Fetchenhauer & Dunning, 2010; Miller, 1999).

Such a mechanism could help explain the prevalence of self-perceived moral superiority. Specifically, because the lion's share of the variance in self-perceived moral superiority likely derives from variance in how people perceive the moral goodness of *others*, rather than themselves. There is relatively limited variance in the latter—people seem largely in agreement that they *themselves* are morally virtuous (for a brief review, see Tappin & McKay, 2017). Taking this in conjunction with evidence that—in interdependent contexts—individuals' moral behavior is conditional on whether they think *others* will behave in kind (Krueger & Acevedo, 2007) implies that greater cynicism—and, thus, greater self-perceived moral superiority—may be associated with *less* moral behavior.

Overview

Given the uncertainty over how self-perceived moral superiority relates to engagement in moral behavior, we set out to investigate this relationship. Specifically, across two studies, we used canonical economic games as measures of moral behavior, and indirectly assessed how moral individuals perceived themselves to be relative to the average person.

Methods

The preregistered protocols, analysis scripts and data for both studies are available on the Open Science Framework (OSF): https://osf.io/p42mp/. Because of their similarity, we present the methods and results of these studies together.

Engagement in Moral Behavior

To measure engagement in moral behavior, we used two incentivized, one-shot, anonymous economic games (with no deception); the Trust Game (TG, Study 1) and Dictator Game (DG, Study 2). These games are typically taken as providing measures of trust in others and trustworthiness, and fairness[9], respectively (see below for descriptions of the games).

While a general prosocial preference is likely to underpin behavior in both the TG and DG (Peysakhovich et al., 2014), past work suggests that trusting behavior in the TG is distinct from giving in the DG (Brülhart & Usunier, 2012), and, indeed, a recent large investigation reported that the shared variance between trusting behavior in the TG, and behavior in the DG, was relatively modest at 12% (Peysakhovich et al., 2014). The relationship between DG behavior and trust*worthy* behavior in the TG was estimated to be somewhat higher—at 25% shared variance with behavior in the DG. In both cases, however, there was evidence of unique variance between the games. This suggests that inclusion of both the TG and DG provided us with three somewhat overlapping but distinct measures of behavior.

We used economic games to measure engagement in moral behavior because numerous studies indicate that people subjectively imbue choices in these games with moral weight. For

[9] We refer to the DG as measuring "fairness" throughout, but note that giving in the DG is also consistent with altruism (Rand et al., 2016). In analyses, we find little difference in the results depending on how the DG measure is construed.

example, recent evidence suggests that prosocial behavior in economic games is driven by an explicit preference for behaving morally (Capraro & Rand, 2017), and behaving prosocially in such games is consistently and strongly judged to be morally superior to behaving self-interestedly (Krueger & Acevedo, 2007; Krueger & DiDonato, 2010; Krueger et al., 2008). The inclusion of the TG and DG thus provided a straightforward decision environment with a recognizable "moral" behavior.

Trust Game

In our TG, participants are anonymously paired and assigned the role of either "Trustor" or "Trustee". Both participants are given $0.20 as a starting endowment, and the Trustor has the option to transfer any amount of their endowment to the Trustee (from $0.00 to $0.20 in increments of $0.01). Any amount they transfer is tripled on its way to the Trustee, and the Trustee is then able to decide how much, if any, of this tripled amount they would like to transfer back to the Trustor (from 0 to 100%). Since the Trustor takes a risk by sending money to the Trustee, their decision is usually taken as a measure of trust. The Trustee, on the other hand, has the option to reciprocate the trust placed in them by the Trustor, by sending some amount of money back to the Trustor. The Trustee decision is thus usually taken as measure of trustworthiness (e.g., Berg et al., 1995).

Dictator Game

In our DG, participants are anonymously paired and assigned the role of either "Dictator" or "Receiver". The Dictator is given $0.30 as a starting endowment, whereas the Receiver starts with nothing. The Dictator then has the option to transfer any amount of their endowment to the Receiver (from $0.00 to $0.30 in increments of $0.01). Since the Dictator's decision is unilateral, with no possibility of reciprocation (or punishment) from the Receiver, they have no financial incentive to share the money. As such, the Dictator's decision to share money is usually taken as a measure of fairness (more technically, *inequity aversion*, see Fehr & Schmidt, 1999).

Self-Perceived Moral Superiority

To measure self-perceived moral superiority[10], we used a regression-based index of trait self-superiority developed and described in detail elsewhere (Heck & Krueger, 2015; Tappin & McKay, 2017). In brief, participants are asked to judge the extent to which 10 moral traits describe (i) themselves and (ii) the average person. They also rate (iii) the social desirability of the traits. The moral traits are presented in Table 1. Conventional measures of self-superiority typically compare how positive self-judgments are with respect to judgments of the average person. However, this overestimates the magnitude and frequency of people who harbor perceptions of self-superiority. The current measure accounts for this overestimation by estimating—and allowing the researcher to remove—a component of self-superiority that may be deemed "defensible" because of the uncertainty people face when making judgments of the average person. Below we describe the computational steps of the measure only (for more detail, see Heck & Krueger, 2015; Tappin & McKay, 2017).

Table 1. Positive and Negative Moral Traits Used in Studies 1 and 2.

Positive moral traits	Negative moral traits
Honest	Insincere
Trustworthy	Prejudiced
Fair	Disloyal
Respectful	Manipulative
Principled	Deceptive

Note. We used the five positive and five negative moral traits from Tappin and McKay (2017).

Step 1. We first estimate how similar each participant's moral self-judgments are to those of the average participant in the sample. To do so, we calculate the average self-judgment for each moral trait over all participants, and then regress these averages on the moral self-judgments made by each individual participant. This provides a moral "coefficient of similarity" (unstandardized slope, *b*) and intercept for each participant. Higher coefficient values indicate that the participant is more like the average participant in the sample. We then compute the mean moral coefficient of similarity and intercept across participants.

[10] In both preregistrations, this construct is referred to as "self-righteousness". This was relabelled to "self-perceived moral superiority" during the review process for better linguistic and conceptual clarity. The measure is identical to that described in the preregistrations.

Step 2. Next, we generate *inferred* moral self-judgments (I) by weighting participants' empirically-observed moral judgments of the average person (O) by the mean coefficient of similarity and intercept, using the formula:

$$I = \frac{O}{mean\ coefficient\ of\ similarity} + mean\ intercept$$

Inferred self-judgments represent self-judgments an *ideal* judge would have made. That is, a judge who perceives how morally similar people are, and uses this information to weight their judgment of the average person to make a more accurate self-judgment. (The basic rationale is this: the more similar people are—defined here by the mean *coefficient of similarity*—the less participants' self-judgments are expected to deviate from their judgments of the average person; see Heck & Krueger, 2015; Tappin & McKay, 2017). At this stage, then, each participant has four sets of judgments for the 10 moral traits. Their empirically observed self-judgments (S), judgments of the average person (O), and social desirability judgments (D), and the new inferred self-judgments (I) computed according to the preceding method.

Step 3. In the final step, we regress S, O, and I on D judgments for each participant. This produces three unstandardized slopes per participant. These slopes express how well moral trait desirability predicts their (i) moral self-judgments (b_{SD}), (ii) judgments of the average person (b_{OD}), and (iii) inferred self-judgments (b_{ID}). In other words, b_{SD} describes the positivity of participants' moral self-perception, b_{OD} describes the positivity of participants' perception of the average person's morality, and b_{ID} describes the positivity of the participants' moral self-perception presupposing they were an ideal judge.

The index of self-perceived moral superiority is computed as the difference between b_{SD} and b_{ID} (specifically, b_{SD} - b_{ID}). This index represents self-perceived moral superiority, but is more conservative than conventional measures because it partitions out a "defensible" component of self-superiority (which is defined by the difference between b_{ID} and b_{OD})[11].

Samples

[11] We report correlations between the "defensible" component of self-perceived moral superiority and economic game behavior in the SM (section 5).

We sought to recruit 824 participants in each study, providing approximately N=412 in each role. Participants were recruited via Amazon's Mechanical Turk (Amir & Rand, 2012; Arechar et al., 2018; Chandler & Shapiro, 2016; Rand, 2012). Sample sizes were determined via power analyses: Our smallest effect size of interest was r=.15, which we required N=343 to achieve 80% power (α=.05) to detect in each of our three primary linear regression analyses (Faul et al., 2009). We deliberately oversampled by approximately 20% to guard against power loss due to planned data exclusions. Sample sizes after data collection were: Study 1 N=827 (50.18% female, M_{age}=38.35 SD_{age}=12.97; Trustor N=413, Trustee N=414), Study 2 N=825 (55.27% female, M_{age}=37.50 SD_{age}=12.62; Dictator N=413, Receiver N=412).

Procedure

The procedure in both studies was substantively identical, and we recruited separate samples in each case (Study 1 participants were identified via their unique Mechanical Turk ID and blocked from participating in Study 2). All participants provided informed consent, before being assigned their role in their respective economic game (Study 1: TG, trustor or trustee, Study 2: DG, dictator or receiver, role assignments were counterbalanced). All participants then completed (i) the trait judgment task, and (ii) the economic game (counterbalanced), except for those assigned the role of receiver in the DG. These participants always completed the DG first, and then completed an unrelated task (receivers are entirely passive and so collecting their trait judgments was unnecessary).

In the trait judgment task, participants were presented with the list of 10 moral traits alongside 20 additional, nonmoral filler traits (inclusion of the nonmoral traits allowed us to replicate the primary results reported by Tappin & McKay, 2017; see SM section 6). Participants were asked to judge (i) the extent to which each trait described themselves, (ii) the extent to which each trait described the average person, and (iii) the social desirability of each trait. Participants rated all 30 traits according to either (i), (ii), or (iii), before moving onto the next set of ratings, and the order of these three sets of judgments was counterbalanced across participants. The presentation order of the traits themselves was randomized in each rating set and for each participant. Rating judgments for the self and the average person were provided on a seven-point scale, ranging from 1 (*Not at all*) to 7 (*Very much so*). Social desirability judgments were

also provided on a seven-point scale, ranging from -3 (*Very undesirable*) to +3 (*Very desirable*).

In the economic games, participants read instructions and completed three comprehension questions assessing their understanding of the payoff structure. Failure to answer all three comprehension questions correctly after two attempts resulted in participants being prevented from completing the survey. After these questions, we revealed which role the participant had been assigned, and they made their decision. We informed them that pairs of decisions would be combined and their bonus calculated and awarded after the survey had concluded (which was true). In addition to bonuses, all participants received a base fee of $0.50 for taking part. At the end of the survey, participants completed simple demographic questions, provided feedback on their experience, and were asked whether they had previously taken part in a similar decision task.

Results

All analyses were conducted in the R environment (R Core Team, 2016). Only dictators are used in Study 2 analyses. Table 2 displays descriptive statistics.

Table 2. Descriptive Statistics from Studies 1 and 2.

	Study 1 (TG)				Study 2 (DG)			
	Slope (b)		Intercept		Slope (b)		Intercept	
	M	SD	M	SD	M	SD	M	SD
Components								
R_{SD}	0.74	0.27	0.98	1.15	0.76	0.26	0.90	1.09
R_{OD}	0.19	0.34	3.13	1.47	0.18	0.36	3.15	1.50
R_{ID}	0.22	0.41	4.38	1.75	0.21	0.41	4.14	1.72
	M		SD		M		SD	
Index of SPMS	0.52		0.41		0.55		0.44	
Transfer amount								
Trustors (c)	13.38		7.28		-		-	
Trustees (%)	35.24		24.44		-		-	
Dictators (c)	-		-		10.39		6.96	

Note. Components are within-participant regressions involved in computing the index of self-perceived moral superiority, according to the procedure outlined in the methods section. TG = Trust Game, DG = Dictator Game; R = regression; S = self-judgments; D = desirability judgments; O = other (average person) judgments; I = inferred-self judgments; SPMS = self-perceived moral superiority (i.e., $b_{SD} - b_{ID}$); M = mean; c = cents. Study 1 N = 736, Study 2 N = 369.

Data Exclusions

All data exclusions were preregistered. Before computing the self-perceived moral superiority index, we excluded responses that contained duplicate IP addresses (Study 1: n=8, 0.97%, Study 2: n=2, 0.48%) and/or one or more failed attention checks (there were three embedded in the trait judgment task) (Study 1: n=28, 3.39%, Study 2: n=22, 5.33%). We then proceeded to compute the index as outlined in Steps 1-3 in the methods section. During Step 1, those participants who responded uniformly on moral self-judgments were excluded (Study 1: n=0, 0%, Study 2: n=4, 0.97%), because the regression analyses in this step require at least *some* variation. During Step 3, for the same reason, we additionally excluded participants who responded uniformly on moral judgments of the average person (Study 1: n=54, 6.53%, Study 2: n=19, 4.60%), and/or social desirability judgments (Study 1: n=1, 0.12%, Study 2: n=4, 0.97%). Sample sizes for the primary analyses were thus, Study 1: Trustors N=369, Trustees N=367, Study 2: Dictators N=369.

Self-Perceived Moral Superiority and Trust in Others

Preregistered analyses. We first regressed trustor decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to transfer amount, model summary: F (1, 367) = 0.14, p=.706, R=.02 [predictor summary: b=-0.34, se=0.89, t=-0.38]. Because the decision data were non-normally distributed, we also conducted a Spearman's rank correlation with the same two variables. The results mirrored the parametric analysis: r_s=-.05, p=.326. Magnitude of self-perceived moral superiority was not meaningfully associated with trusting behavior in the TG.

Exploratory analyses. We conducted several exploratory analyses to test the robustness of this conclusion. First, we dichotomized the trustor decisions by assigning them a value of 1 if they were greater than the median transfer amount of 15c, and a value of 0 if they were equal to or less than this amount. A total of 179 (48.51%) participants transferred greater than the median amount of 15c. A binary logistic regression predicting the probability of an above median transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: Odds Ratio (OR) = 0.90, 95% CI [0.55, 1.46], p=.669 (Figure 1). That is, self-perceived moral superiority was not meaningfully associated with the probability of transferring greater than the median transfer amount. For all DVs, we also explored whether prior experience with the games was masking an association between self-perceived moral superiority and decision behavior in our sample (it wasn't) (cf. Chandler et al., 2015; see SM section 2 for these analyses).

Given these results, we sought to quantify the relative strength of evidence in favor of the null hypothesis. We conducted a Kendall's tau Bayesian correlation analysis using JASP software (JASP Team, 2017). Under a uniformly distributed prior, we obtained a Bayes Factor (BF) of 8.23 in favor of the null hypothesis. That is, the BF indicated moderate support for the null over the alternative hypothesis. The BF in favor of the null remained moderate-to-strong over a wide range of priors (see SM section 1). The results of the exploratory analyses support those of the preregistered analyses.

Self-Perceived Moral Superiority and Trustworthiness

Preregistered analyses. As before, we began by regressing trustee decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to back-transfer amount: F (1, 365) = 0.04, p=.851, R=.01 [b=0.60, se=3.17, t=0.19]. Because the

decision data were again non-normally distributed, we followed with a Spearman's rank correlation. The results mirrored the parametric analysis: r_s=-.004, p=.931. Magnitude of self-perceived moral superiority was not meaningfully associated with trustworthiness behavior in the TG.

Exploratory analyses. Once again, we conducted several exploratory analyses to investigate the robustness of this conclusion. We first dichotomized the trustee decisions by assigning them a value of 1 if they were greater than the median back-transfer amount of 50%, and a value of 0 if they were less than this amount. A total of 55 (14.99%) participants back-transferred greater than the median amount of 50%. A binary logistic regression predicting the probability of an above median back-transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: OR=0.64 [0.31, 1.32], p=.233 (Figure 1). That is, self-perceived moral superiority was not meaningfully associated with the probability of an above median back-transfer. As before, a Kendall's tau Bayesian correlation analysis conducted in JASP (uniformly distributed prior) returned a BF of 14.58 in favor of the null hypothesis. That is, the BF indicated strong support for the null over the alternative hypothesis. The BF remained moderate-to-strong over a wide range of priors (see SM section 1). The results of the exploratory analyses thus support those of the preregistered analyses.

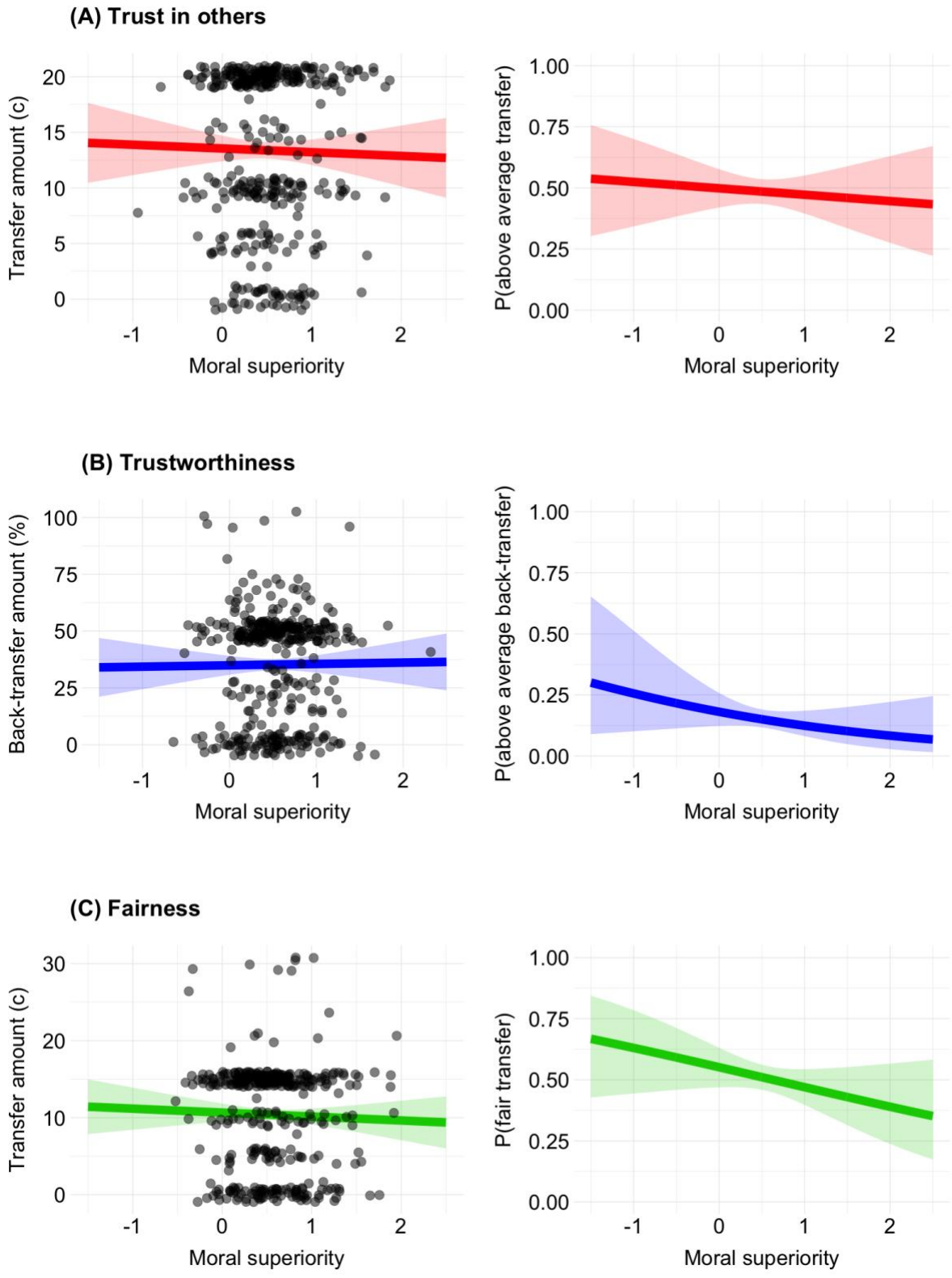


Figure 1. Relationship between self-perceived moral superiority and transfer amounts in studies 1 (A, B) and 2 (C). *Scatter points are raw data with slight jitter for visibility and the shaded regions denote 95% confidence intervals. (A) Left panel: Preregistered analysis*

regressing trustor transfer amount on self-perceived moral superiority (b = -0.34, se = 0.89). Right panel: Exploratory binary logistic regression analysis of the probability that trustor transfer was greater than the median transfer amount (15c), based on self-perceived moral superiority scores (OR = 0.90 [0.55, 1.46]). N = 369. (B) Left panel: Preregistered analysis regressing trustee back-transfer amount on self-perceived moral superiority (b = 0.60, se = 3.17). Right panel: Exploratory binary logistic regression analysis of the probability that trustee back-transfer was greater than the median back-transfer amount (50%), based on self-perceived moral superiority scores (OR = 0.64 [0.31, 1.32]). N = 367. (C) Left panel: Preregistered analysis regressing dictator transfer amount on self-perceived moral superiority (b = -0.62, se = 0.83). Right panel: Exploratory binary logistic regression analysis of the probability that dictator transfer was fair (15c), based on self-perceived moral superiority scores (OR = 0.72 [0.45, 1.15]). N = 369.

Self-Perceived Moral Superiority and Fairness

Preregistered analyses. We began by regressing dictator decisions on self-perceived moral superiority scores (Figure 1). Self-perceived moral superiority was trivially related to transfer amount: F (1, 367) = 0.57, p=.452, R=.04 [b=-0.62, se=0.83, t=-0.75]. Because the decision data were non-normally distributed, we conducted a Spearman's rank correlation with the same two variables. The results mirrored the parametric analysis: r_s=-.05, p=.345. We quantified the relative strength of evidence in favor of the null by conducting a Bayesian correlation analysis in JASP. We preregistered our intention to conduct a *Pearson's rho* Bayesian correlation, but, given the severe non-normality of the decision data, a Kendall's tau Bayesian correlation is more appropriate. For transparency, we report both. The BF_{rho} was 11.57, and BF_{tau} was 8.38 in favor of the null hypothesis (uniformly distributed priors). Both indicated moderate-to-strong support for the null over the alternative hypothesis. In SM section 1, we report BF_{tau} over a wide range of priors (it remained moderate-to-strong in favor of the null).

Next, to account for the fact that transfer amounts of greater than 15c—that is, greater than half the dictator's endowment—are technically "unfair" (Fehr & Schmidt, 1999), rather reflecting altruism or "hyper-fairness" (Henrich et al., 2006; Rand et al., 2016), we repeated the above analyses with a truncated sample of dictators—excluding those who transferred greater than

15c ($N_{excluded}$=15, 4.07%). The truncated analyses thus tested whether self-perceived moral superiority was associated with fairness behavior, where unfair behavior was defined as inequity in favor of oneself (i.e., the dictator). The pattern of results was the same as in the full sample, regression: F (1, 352) = 0.87, p=.351, R=.05 [b=-0.73, se=0.78, t=-0.93], Spearman's rank correlation: r_s=-.06, p=.235, Bayesian correlation: BF_{rho}=9.75 and BF_{tau}=5.88 in favor of the null (uniform priors; BF_{tau} robust over a range of priors, see SM section 1). Magnitude of self-perceived moral superiority was not meaningfully associated with fairness behavior in the DG.

Exploratory analyses. To check robustness, we dichotomized the dictator decisions by assigning them a value of 1 if they were equal to 15c, and a value of 0 if they were greater than or less than this amount. Fairness was thus strictly defined as rejection of inequity in favor of either oneself (dictator) *or* the other person (receiver). A total of 187 (50.68%) participants split the money fairly, transferring exactly 15c. A binary logistic regression predicting the probability of fair transfer, based on self-perceived moral superiority scores, corroborated the preregistered analyses: OR=0.72 [0.45, 1.15], p=.174 (Figure 1). That is, self-perceived moral superiority did not meaningfully predict the probability of a fair transfer. The results of the exploratory analyses are consistent with those of the preregistered analyses.

Discussion

We investigated how self-perceived moral superiority related to behavior in two canonical economic games—the Trust Game (TG) and Dictator Game (DG). Across two studies, self-perceived moral superiority was not associated with magnitude of trust in others, trustworthiness, or fairness, as these behaviors are measured in the games. This pattern of results was robust to a variety of analyses, and, for each of the three dependent variables, Bayesian analyses indicated relatively strong support for the null vs. alternative hypothesis.

The findings are inconsistent with our hypotheses: that self-perceived moral superiority would be associated with (i) more, or with (ii) less, moral behavior. Whereas some evidence suggests that perceptions of *non*moral self-superiority are associated with (Blanton et al., 1999; Heck & Krueger, 2015), and possibly facilitate (O'Mara & Gaertner, 2017) behavioral performance, we found that self-perceived moral superiority was not associated with behavior in canonical

economic games—in which moral motivation appears reliably engaged (Capraro & Rand, 2017), and where morally superior decisions are readily discerned (Krueger & Acevedo, 2007; Krueger & DiDonato, 2010; Krueger et al., 2008).

Why was self-perceived moral superiority unrelated to behavior in the games? One explanation is that our measure was *domain general*. That is, participants provided judgments for a range of moral traits, which fed into a single score indexing their self-perceived moral superiority. It is possible that superiority perceived on specific moral traits *is* associated with behavior representative of those traits, but that our domain general measure obscured these relationships. We examined this possibility by computing raw difference scores between participants' self-judgments and their judgments of the average person for the traits "trustworthy" and "fair" only, and correlating these scores with trustee decisions, and dictator decisions, respectively (SM section 3). These coefficients were also trivial in size ($|r_s| < .03$)—suggesting that the domain-generality of our measure does not account for the current pattern of results.

An interesting and related question is whether individuals' moral *self*-perception—not their perceived *superiority* over others—was associated with absolute magnitude of monetary transfer in the games. Exploratory correlations suggested a small but consistently positive association between moral self-perception (b_{SD}) and transfer amount across dependent variables; trust in others (r_s=.12), trustworthiness (r_s=.15), and fairness (r_s=.06). We observed some evidence for self-knowledge—those people who had a more positive view of their own morality tended to transfer more money to their partners. This is consistent with prior evidence that self-perceptions are at least somewhat diagnostic of behavior/reality (Epley & Dunning, 2000; Vazire & Carlson, 2010), and that self-reported traits correlate with prosociality in economic games (Hillbig et al., 2013). This raises the question of what role moral judgments of the average person had in participants' behavior.

It is plausible that the magnitude of self-perceived moral superiority is driven primarily by variance in how people view the morality of *other* people, not themselves (cf. Tappin & McKay, 2017), and that greater moral cynicism about others is associated with lower engagement in certain types of moral behavior (Krueger & Acevedo, 2007). This provides one explanation for why the above positive associations between moral self-perception and behavior did not emerge for self-perceived moral superiority. Specifically, because they were cancelled out by the cynicism disproportionately driving the latter.

We subjected this speculation to the data. First, comparing the shared variance between self-perceived moral superiority scores and both (i) moral self-perceptions (b_{SD}), and (ii) perceptions of the average person's morality (b_{OD}), revealed that the latter explained, on average, 64% variance in the scores, whereas the former accounted for less than a quarter of this amount (SM section 4). Second, perceptions of the average person's morality were weakly but consistently positively related to transfer amount across dependent variables; trust in others (r_s=.11), trustworthiness (r_s=.12), and fairness (r_s=.08). In other words, self-perceived moral superiority was mainly driven by how individuals viewed the morality of other people, not themselves, and greater moral cynicism about these others tended to be associated with lower monetary transfers. This supports our speculation on both counts, and is consistent with two areas of prior work: the first, that observers interpret expressions of self-superiority as condemnation of others, rather than egregious self-flattery (Van Damme et al., 2016; Van Damme et al., 2017), and, the second, that individuals condition their behavior in these games on whether they think *others* will behave in kind (Krueger & Acevedo, 2007).

Based on this, we suggest that, despite the robust observation that most people rate themselves as morally superior to the average person, this phenomenon has limited predictive validity due to the seemingly opposed behavioral influences of self- and other-perception that comprise its measurement. That said, we note there is mixed evidence over whether economic games are valid analogues of behavior in the real world (Benz & Meier, 2008; Fehr & Leibbrandt, 2011; Franzen & Pointner, 2013; Galizzi & Navarro-Martinez, 2017). It is thus reasonable to ask whether our results would generalize to more ecologically valid cases of moral behavior. This represents an interesting avenue for future research. Furthermore, there is evidence that East Asian samples do not report self-superiority perceptions to the same extent as Western samples (Heine & Hamamura, 2007); indicating our results may differ along these specific cultural lines.

We do expect, however, that our results will be robust to variations in the economic game environment—in particular, changes to the size of the monetary stakes. Indeed, meta-analytic reviews indicate that game behavior tends to differ rather minimally over variance in stake size (Engel, 2011; Johnson & Mislin, 2011). In addition, both our measure of self-perceived moral superiority, and our analytic approach, were comprehensive—comprising a variety of validated moral traits (see Tappin & McKay, 2017), and a range of robustness checks, respectively. We

expect conceptual replications that use alternative measures of moral superiority and alternative analytic approaches to produce similar results to those we observed here. We have no reason to believe that the results depend on other characteristics of the participants, materials, or contexts (Simons et al., 2017).

Here, we investigated how self-perceived moral superiority related to moral behavior as measured in canonical economic games. We observed robust evidence that self-perceived moral superiority is not associated with magnitude of trust in others, trustworthiness, or fairness, as defined by the games; a result seemingly produced by the opposite behavioral manifestations of (i) self-knowledge and (ii) cynicism about the morality of the average person.

Supplemental Material

Section 1

Bayesian Analyses

All Bayesian analyses were Kendall's tau Bayesian correlation pairs conducted in JASP. Below we report the robustness checks (displayed graphically) for each Bayesian analysis reported in the main text.

Self-Perceived Moral Superiority and Trust in Others

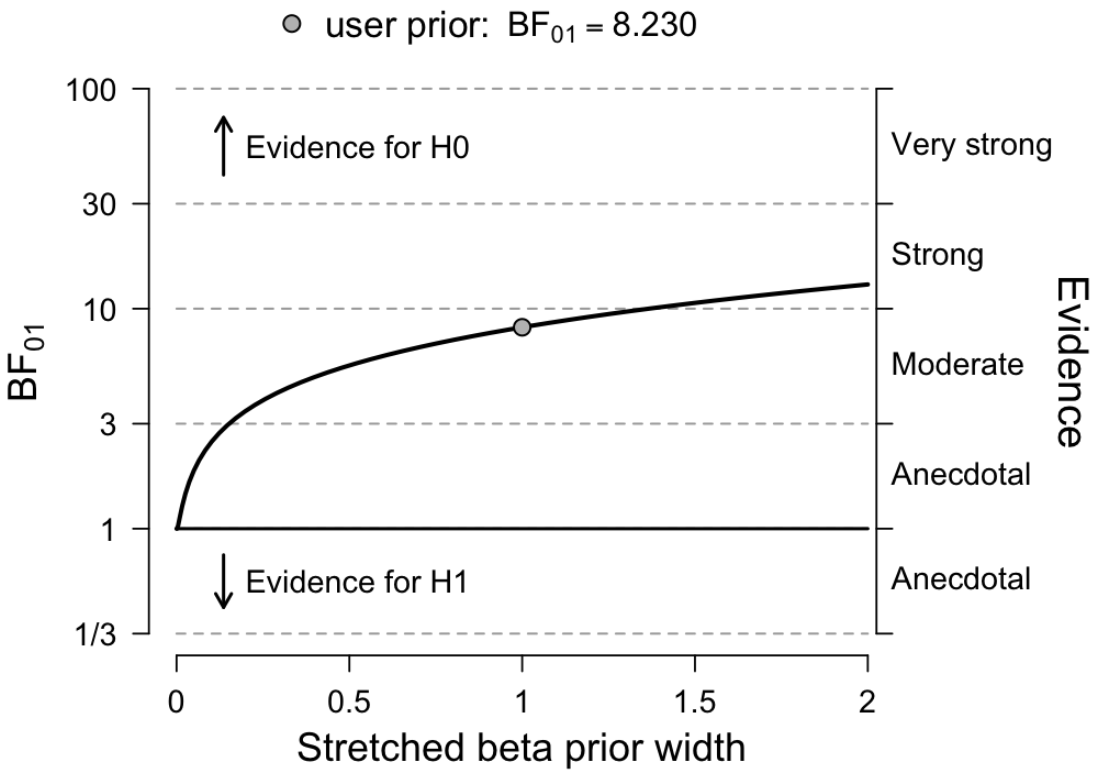


Figure S1. Bayes Factor robustness check: Trustors. *Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.*

Self-Perceived Moral Superiority and Trustworthiness

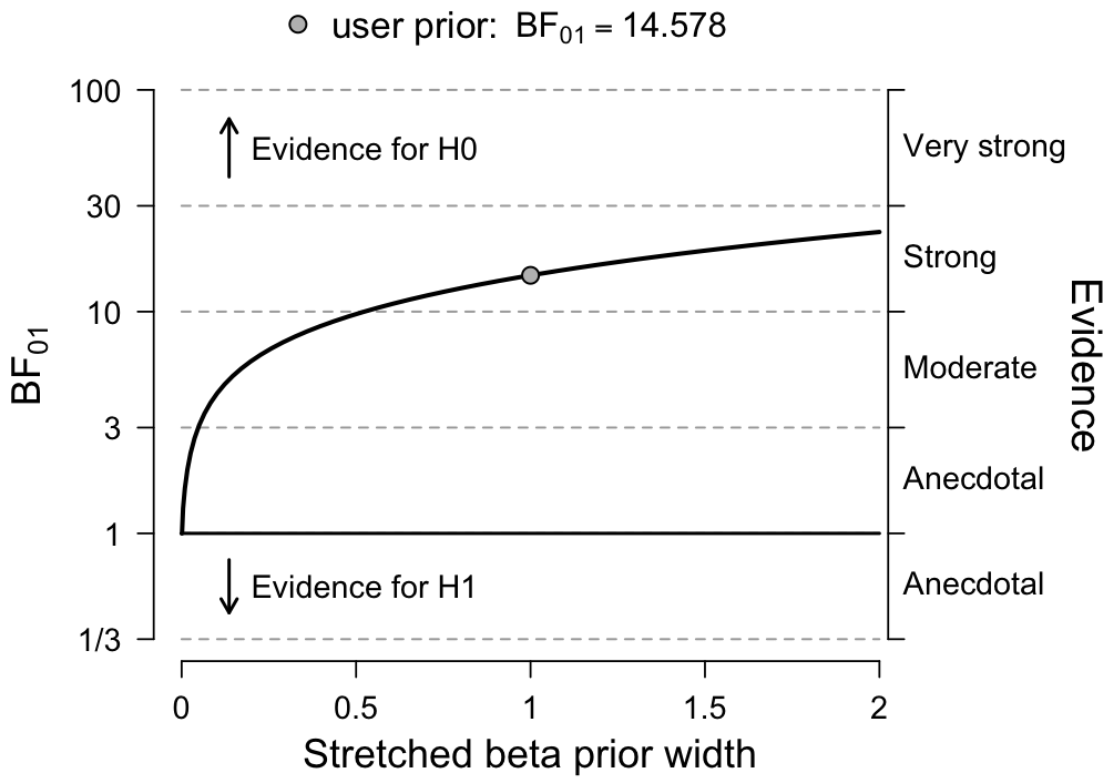


Figure S2. Bayes Factor robustness check: Trustees. *Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.*

Self-Perceived Moral Superiority and Fairness

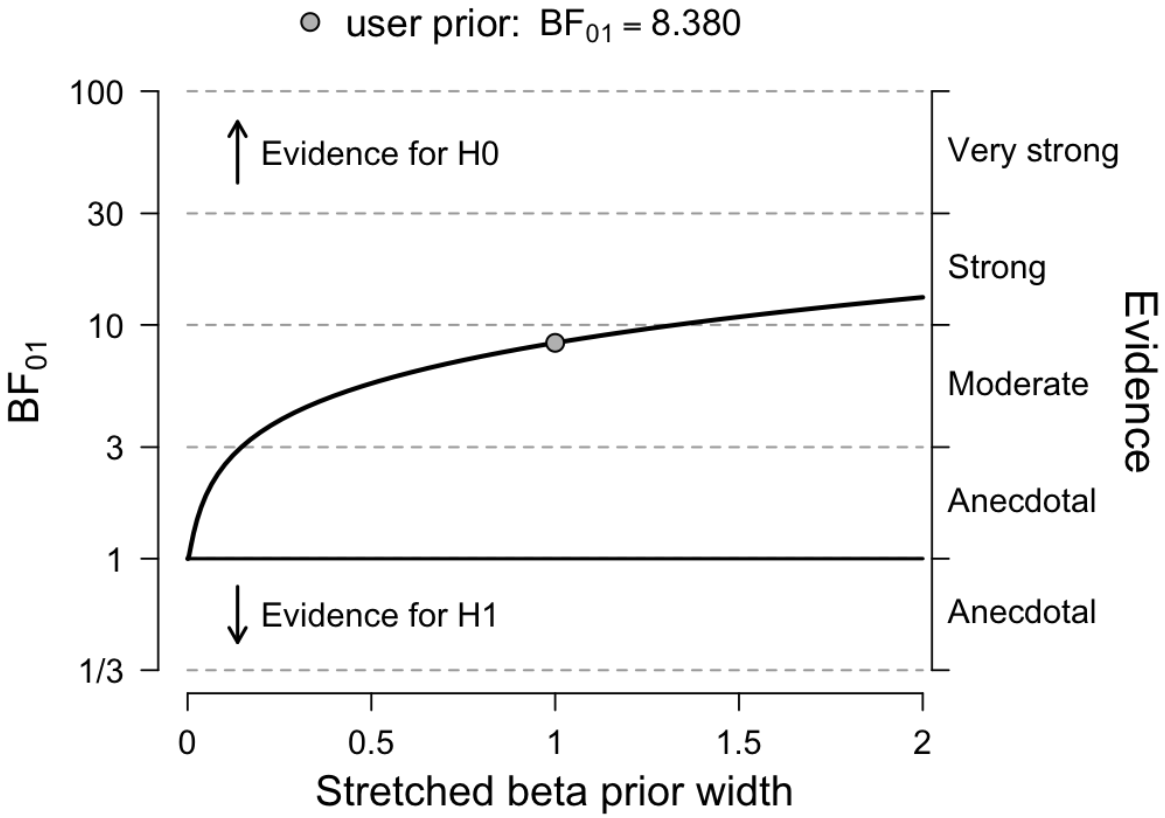


Figure S3. Bayes Factor robustness check: Dictators. *Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.*

Self-Perceived Moral Superiority and Fairness (Truncated Sample)

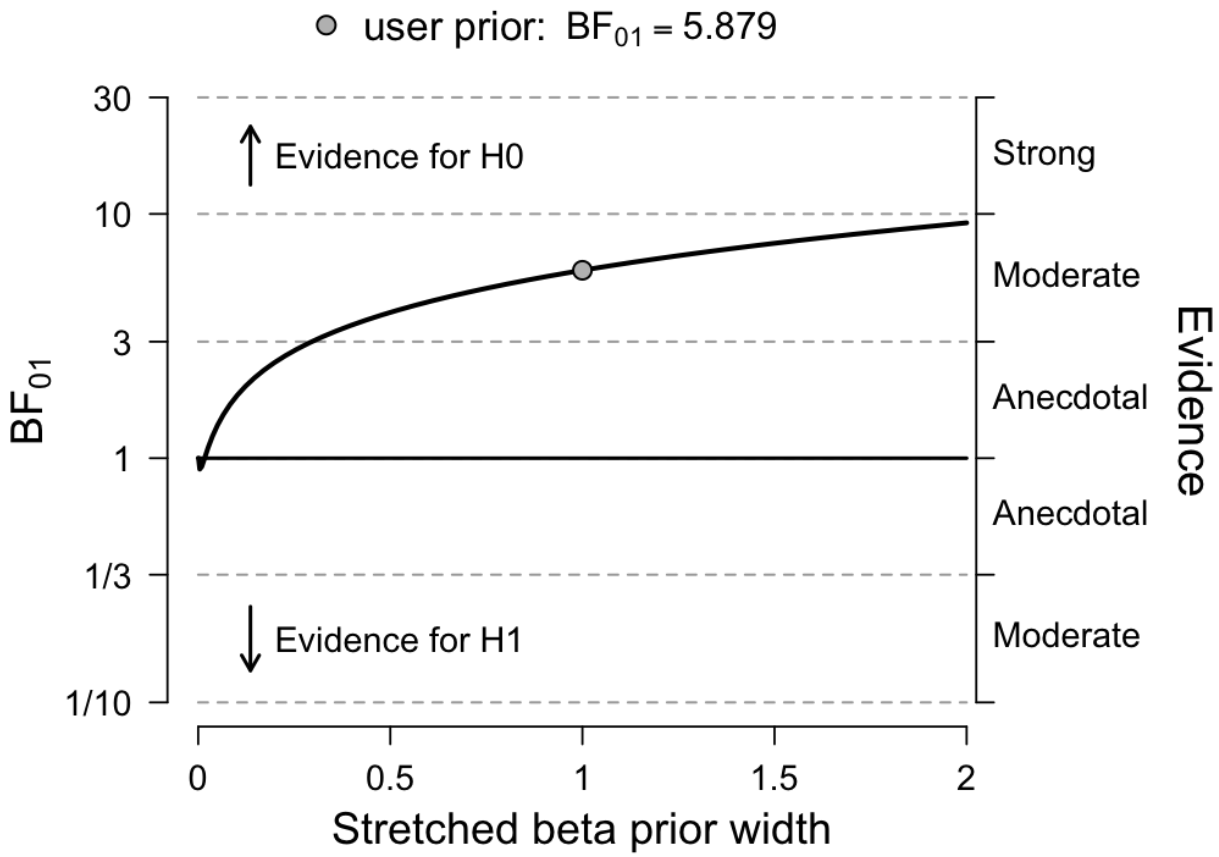


Figure S4. Bayes Factor robustness check: Dictators (truncated). *Relative support for the null over alternative hypothesis as a function of prior width. The BF indicates moderate to strong support over a range of priors.*

Section 2

Previous Experience with Economic Games

For all DVs (trustor transfers, trustee transfers, dictator transfers), we explored whether prior experience with the games was masking an association between self-perceived moral superiority and decision behavior (cf. Chandler et al., 2015). Below we present these analyses in the order of DVs present in the main text.

Excluding trustors who reported having previously seen the TG ($N_{excluded}$=125, 33.88%), and repeating the preregistered regression analysis, corroborated the preregistered and other exploratory analyses: F (1, 242) = 0.41, p=.522, R=.04 [b=-0.66, se=1.03, t=-0.64]. Similarly, excluding trustees who reported having previously seen the TG ($N_{excluded}$=114, 31.06%), and repeating the preregistered regression analysis revealed similar results to those in the full (nonnaive) sample: F (1, 251) = 0.01, p=.929, R=.01 [b=0.35, se=3.89, t=0.09]. Finally, excluding dictators who reported having previously seen the DG ($N_{excluded}$=140, 37.94%), and repeating the preregistered regression analysis, produced the same pattern of results as those in the full (nonnaive) sample: F (1, 227) = 0.35, p=.555, R=.04 [b=-0.64, se=1.08, t=-0.59]. All these exploratory analyses are consistent with the preregistered and exploratory analyses reported in the main text; specifically, indicating that prior experience with the economic games was not masking an association between our variables of interest.

Section 3

The Domain-Generality of Our Measure

To explore the possibility that the domain-generality of the self-perceived moral superiority measure was obscuring any relationship between perceived superiority on *specific* moral traits (e.g., trustworthiness, or fairness) and behavior representative of those traits, we first computed difference scores between participants' given self-judgments (s) and their judgments of the average person (o) for the traits "trustworthy" and "fair" only. We then conducted Spearman's rank correlations between these scores and trustee decisions (Study 1), and dictator decisions (Study 2), respectively. Perceived superiority on the trait "trustworthy" was trivially related to trustee back-transfer amount: r_s = -.03, p = .588. Similarly, perceived superiority on the trait "fair" was trivially related to dictator transfer amount: r_s = -.002, p = .973. These results mirror those of the preregistered analyses using the self-perceived moral superiority measure.

Section 4

Explaining Variance in Self-Perceived Moral Superiority

We explored whether self-perceived moral superiority scores were better explained by (i) moral self-perceptions (b_{SD}), or (ii) perceptions of the average person's morality (b_{OD}). Separately correlating (i) and (ii) with self-perceived moral superiority scores indicated that the latter explained, on average, 64.20% variance in these scores (Study 1: r = -.79, p <.001, 62.16% variance explained, Study 2: r = -.81, p <.001, 66.24% variance explained). Whereas, the former accounted for less than a quarter of this amount (average variance explained: 14.06%, Study 1: r = .35, p <.001, 12.35% variance explained, Study 2: r = .40, p <.001, 15.77% variance explained).

Section 5

Defensible Self-Perceived Moral Superiority and Economic Game Behavior

We explored whether the "defensible" component of self-perceived moral superiority—as given by the regression-based index—was associated with behavior in the economic games. This component is defined by $b_{ID} - b_{OD}$; or, the amount of self-superiority that may be justified by the fact that individuals have limited information about the average person (Heck & Krueger, 2015; Tappin & McKay, 2017). Defensible self-perceived moral superiority was weakly but positively correlated with transfer amount for those in the role of trustor [r_s = .11, p=.034], trustee [r_s = .12, p=.023], and dictator [r_s = .08, p=.110]. This provides some intuitive rationale for labeling the index "defensible", but we emphasize that caution must be used when interpreting the meaning of these results given the lack of a priori predictions we had about these associations.

Section 6

Replication of Tappin & McKay (2017) (The Illusion of Moral Superiority)

We included 20 nonmoral filler traits in the trait judgment task—10 of which pertained to the domain of agency, and 10 to the domain of sociability (also drawn from Tappin & McKay, 2017)—and we were thus able to replicate the primary results reported in Tappin and McKay (2017) (Table S1 displays the full list of traits). Specifically, in their study they found that self-perceived moral superiority—measured using the same regression-based index as in the current

studies—was larger in magnitude, and more frequent, than perceived superiority in *nonmoral* domains of social perception. Thus, in the following section we reproduce the primary analyses reported in Tappin and McKay (2017) (p.6, para beginning: "*This indicates that irrational self-enhancement is strongest in the moral domain.*")[12]. For consistency, we use their terminology to refer to perceived superiority (that is, "*irrational self-enhancement*").

Table S1. Full List of Traits Used in Studies 1 and 2.

Domain	Positive traits	Negative traits
Morality	Honest	Insincere
	Trustworthy	Prejudiced
	Fair	Disloyal
	Respectful	Manipulative
	Principled	Deceptive
Agency	Hard-working	Lazy
	Knowledgeable	Undedicated
	Competent	Unintelligent
	Creative	Unmotivated
	Determined	Illogical
Sociability	Sociable	Cold
	Playful	Disagreeable
	Warm	Rude
	Family-orientated	Humorless
	Easy-going	Uptight

Note. Traits were identical to those used in Tappin & McKay (2017), except that the trait "playful" replaced "cooperative".

Irrational Self-Enhancement

We first investigated the magnitude of irrational self-enhancement in each trait domain by examining how well trait desirability predicted actual self-judgments (mean b_{SD}) vs. inferred self-judgments (mean b_{ID}). In both studies, replicating Tappin and McKay (2017), paired t-tests revealed that magnitude of irrational self-enhancement was largest in the moral domain,

[12] Note: we replicate the *magnitude* and *frequency* analyses of Tappin and McKay (2017) only, not the analyses with self-esteem (since we did not collect self-esteem data in the current studies). Also, sample N's differ from those reported in the preregistered analyses because, prior to replicating Tappin & McKay (2017), we had to additionally exclude uniform responders in the nonmoral trait domains.

Study 1: Morality (0.74 vs. 0.23), t (727) = 33.52, p <.001, Cohens d = 1.24 95% Confidence Interval [1.15, 1.34], Agency (0.73 vs. 0.28), t (727) = 24.50, p <.001, d = 0.91 [0.82, 0.99], and Sociability (0.61 vs. 0.52), t (727) = 4.17, p <.001, d = 0.15 [0.08, 0.23]; *Study 2*: Morality (0.76 vs. 0.21), t (361) = 23.60, p <.001, d = 1.24 [1.10, 1.38], Agency (0.72 vs. 0.28), t (361) = 17.10, p <.001, d = 0.90 [0.78, 1.02], and Sociability (0.63 vs. 0.49), t (361) = 5.49, p <.001, d = 0.29 [0.18, 0.39].

We confirmed statistically that Morality was the largest by computing the difference measure ($b_{SD} - b_{ID}$) for each trait domain, and conducting paired t-tests *between* trait domains. In both studies, replicating Tappin and McKay (2017), the moral domain comprised the largest magnitude of irrational self-enhancement, *Study 1*: Morality (0.52) vs. Agency (0.46), t (727) = 4.00, p <.001, d = 0.15 [0.08, 0.22], and vs. Sociability (0.09), t (727) = 25.36, p <.001, d = 0.94 [0.85, 1.03]; *Study 2*: Morality (0.55) vs. Agency (0.44), t (361) = 5.04, p <.001, d = 0.27 [0.16, 0.37], and vs. Sociability (0.15), t (361) = 17.16, p <.001, d = 0.90 [0.78, 1.02].

Finally, corroborating the analysis of magnitude, and again replicating Tappin and McKay (2017), McNemar's Tests showed that more individuals irrationally self-enhanced ($b_{SD} > b_{ID}$) in the moral domain than in either of the nonmoral domains, *Study 1*: Morality (n = 659, 90.52%) vs. Agency (n = 611, 83.93%), χ^2 (df = 1, N = 728) = 26.94, p <.001, and vs. Sociability (n = 396, 54.40%), χ^2 (df = 1, N = 728) = 249.61, p <.001; *Study 2*: Morality (n = 331, 91.44%) vs. Agency (n = 302, 83.43%), χ^2 (df = 1, N = 362) = 16.68, p <.001, and vs. Sociability (n = 214, 59.12%), χ^2 (df = 1, N = 362) = 109.40, p <.001.

1.3. Moral Polarization and Out-Party Hate in the US Political Context[13]

Abstract

Affective polarization describes the phenomenon whereby people identifying as Republican or Democrat tend to view opposing partisans negatively and co-partisans positively. Though extensively studied, there remain important gaps in scholarly understanding of affective polarization. In particular, (i) how it relates to the distinct behavioural phenomena of in-party "love" vs. out-party "hate"; and (ii) to what extent it reflects a *generalized* evaluative disparity between partisans vs. a domain-*specific* disparity in evaluation. Here, we report the results of an investigation that bears on both of these questions. Specifically, drawing on recent theoretical and empirical trends in political science and psychology, we hypothesize that *moral* polarization—the tendency to view opposing partisans' *moral character* negatively, and co-partisans' *moral character* positively—is associated with behavioural expressions of *out-party hate*. We test this hypothesis in two preregistered studies comprising behavioural measures and large convenience samples of US partisans (total N=1354). Our results strike an optimistic chord: Taken together, they suggest that the hypothesized association is probably small and somewhat tenuous. Though moral polarization *per se* was large—likely exceeding prior estimates of generalized affective polarization—even the *most* morally polarized partisans appeared reluctant to engage in a mild form of out-party hate behaviour. These findings converge with recent evidence that polarization—moral or otherwise—has yet to translate into the average US partisan wanting to actively harm their out-party counterparts.

[13] The work presented in this section was conducted in collaboration with Ryan McKay (supervisor) and is posted as a preprint on *PsyArXiv*: https://psyarxiv.com/4fxb3/

Introduction

Animosity between Republicans and Democrats is a salient feature of American political life. This animosity has been dubbed *affective polarization*; that is, the "tendency for people identifying as Republican or Democrat to view opposing partisans negatively and co-partisans positively" (p. 691, Iyengar & Westwood, 2015). Affective polarization is in evidence across a range of measures, and has been increasing over time (Iyengar, Sood, & Lelkes, 2012; Iyengar et al., 2018). For example, time series data from the American National Election Study indicate that the disparity in "warmth" that Democrats and Republicans (hereafter, *partisans*) express for their own party vs. the other party was greater in 2012 than at any point during the 34 preceding years; almost doubling in size since 1978 (Iyengar et al., 2012). Indeed, according to more recent analysis using this measure, the average partisan now feels almost three times more positive about the in-party than out-party (cf. Figure 1 in Iyengar et al., 2018).

Of course, affective polarization inferred from placement on these "feeling thermometers" (or other self-report measures) says little about the behavioural manifestations and consequences of the phenomenon. In their recent review of affective polarization, Iyengar and colleagues (2018) cite and discuss evidence of such ramifications. In particular, partisans are liable to allocate more money to the in-party than to the out-party in behavioural economic games, (Carlin & Love, 2013; Iyengar & Westwood, 2015), and are more likely to pursue online dating opportunities with politically similar others (Huber & Malhotra, 2017). Furthermore, the magnitude of affective polarization positively correlates with avoidance of opposing partisans in a group problem-solving task (Lelkes & Westwood, 2017). Despite this evidence, Iyengar and colleagues (2018) note that it remains unclear (i) precisely *how* affective polarization relates to the distinct behavioural phenomena of "love" for the in-party (i.e., ingroup favouritism) vs. "hate" for the out-party (outgroup hostility) (cf. Brewer, 1999); and (ii) to what extent affective polarization reflects a *generalized* evaluative disparity between partisans vs. a more domain-*specific* disparity in evaluation (e.g., that out-partisans are less *trustworthy* than in-partisans).

In this paper, we report the results of an investigation that bears on both of these questions. Specifically, we draw on recent theoretical and empirical trends in psychology and political science to hypothesize that *moral* polarization—that is, the tendency for people to view

opposing partisans' *moral character* negatively, and co-partisans' *moral character* positively (cf. Iyengar & Westwood, 2015)—is associated with behavioural expressions of *out-party hate*. We test this hypothesis in two preregistered studies comprising behavioural economic game measures and large samples of US partisans.

1.1. Group Identity and Moral Psychology in American Politics

Generally speaking, political (or ideological) conflict entails disagreement over which set of shared beliefs, values and practices make for a good and desirable society, and how this can be achieved (Jost et al., 2009). On this basis, even mild political disagreement is likely to be characterized by the belief that the in-party is more "moral" than the out-party; in other words, moral polarization. While the average American is not particularly committed to one ideological viewpoint over another, they do appear committed to a partisan *group identity*; that is, for the average American voter, politics may be more a case of Us vs. Them, than "our policy" vs. "their policy" (Kinder & Kalmoe, 2017). This can be expected to exacerbate moral polarization insofar as the human mind is primed to distinguish between ingroups and outgroups, and to interpret the social world in moral terms (Brewer, 1999; Haidt, 2012). Indeed, this proposition is consistent with the putative importance of both group identity and moral psychology in contemporary American politics.

Mason (2016, 2018), for example, documents that party identity in the US is increasingly in alignment with various other group identities, including race-, ideological-, and religious-based identities. Such "social sorting" may facilitate identification with the in-party and reduce the tempering influence of cross-cutting identities on out-party hostility (Mason & Wronski, 2018; Roccas & Brewer, 2002). At the same time, Ryan (2014) and Koleva and colleagues (2012) report evidence to suggest that moral psychological factors play an important and distinct role in the political preferences and behaviour of US partisans. In particular, the latter report that endorsement of a small number of "moral foundations" explains variance in attitudes across a wide range of US political issues—including gun control, immigration, same-sex marriage and abortion—beyond other relevant factors such as age, gender, ideology and interest in politics (Koleva et al., 2012). Ryan (2014), corroborating these results, finds that moral conviction is common in partisans' policy attitudes—even for putatively *non*moral policy issues—and may undergird both political activism and political antagonism (see also Ryan, 2017; Skitka et al., 2005). Finally, recent work using data from Twitter suggests that posts about US political

issues spread over the (ingroup) network to a greater extent if they contain *moral*-emotional language (vs. nonmoral-emotional language) (Brady et al., 2017).

In summary, the suffusion of group identities and moral psychology in contemporary American politics suggests that moral polarization—the tendency for people to view opposing partisans' moral character negatively, and co-partisans' moral character positively—may be particularly prominent among US partisans. We now consider the possible behavioural manifestations and consequences of moral polarization.

1.2. Moral Polarization and Out-Party Hate

Though they are often conflated, ingroup "love" and outgroup "hate" are distinct phenomena (Brewer, 1999). Whereas the former represents adulation for—and favouritism towards—members of one's own group, the latter represents hatred of—and hostility towards—members of other groups. Thus, as the distinction makes clear, people can exhibit ingroup love *without* exhibiting hostility towards a relevant outgroup (we note that the reverse case—outgroup hate in the absence of ingroup love—appears less plausible a phenomenon). Encouragingly, where the two are appropriately disentangled, ingroup favouritism often takes psychological and behavioural primacy over outgroup hate. That is, most people—given the chance—opt only for behaviours that benefit the ingroup, rather than opting for behaviours that both benefit the ingroup *and harm* the outgroup (e.g., Brewer, 1999; Halevy et al., 2008; Weisel & Böhm, 2015). However, in some contexts individuals exhibit both love for the ingroup *and* hostility towards a relevant outgroup (extreme examples include suicide terrorism, war, etc.).

Recent work suggests that outgroup hate behaviour of this kind (if not severity) is more common when group identities are defined—and the relevant groups divided—along *morality*-based lines (Parker & Janoff-Bulman, 2013; Weisel & Böhm, 2015). For example, using a novel behavioural economic game with subjects in Germany, Weisel and Böhm (2015) find that game decisions indicative of hostility toward the outgroup are more common towards supporters of the National Democratic Party (NPD)—considered neo-Nazi and widely morally opposed—than towards supporters of other political parties in Germany. While the NPD are arguably unique in their moral, cultural and historical significance in Germany, this result suggests that behavioral expressions of out-party hate may increase in conjunction with moral

polarization in the United States. In other words, as the perceived moral "gap" between in- and out-party widens, behavioural expressions of out-party hate increase in frequency.

This hypothesis is corroborated by the conjunction of two phenomena identified in analyses of real-world ideological conflict. First, analysis of the patterns of thinking in militant extremism (Giner-Sorolla et al., 2012; Saucier et al., 2009), violent political and religious conflict (Ginges et al., 2011), and genocidal regimes (Koonz, 2003; Reicher et al., 2008) identifies the *extolling of ingroup virtue* as a persistent theme. This is intuitive: To eschew self-interest and contribute to ingroup ends, would-be contributors must presumably feel sufficiently persuaded of the righteousness of their comrades, and of their cause. When the ingroup and its cause are perceived as just, personal costs may be tolerable or even desirable (Saucier et al., 2009). Second, *moral demonization of the outgroup* is another recurring theme in real-world ideological conflict (Giner-Sorolla et al., 2012; Halperin, 2008; Reicher et al., 2008; Saucier et al., 2009). In genocides, for example, propaganda depicting outgroup targets as nefarious agents with hostile intentions is reputedly commonplace, and is thought to be a deliberate strategy to rally public support for genocidal policy (Bilewicz & Vollhardt, 2012). Morally demonized outgroups may be perceived as an existential threat to the ingroup (Giner-Sorolla et al., 2012), and, in contexts where the latter is morally championed, this feeds a compelling Manichean survival narrative of good against evil (Reicher et al., 2008; Saucier et al., 2009). Under such conditions, expressions of outgroup hate can become morally *mandated* (Skitka & Mullen, 2002).

1.3. Overview of Studies

Following the preceding analyses, we hypothesized that moral polarization would be associated with behavioural expressions of out-party hate in the US political context. In particular, as the perceived moral "gap" between in- and out-party widens, behavioural expressions of out-party hate increase in frequency. We tested this hypothesis in two preregistered studies—an initial study and a close replication—comprising large samples of US partisans, and a behavioural economic game measure of outgroup hate (cf. Weisel and Böhm, 2015).

Methods

Both studies were preregistered on *AsPredicted*: https://aspredicted.org/e3hw9.pdf (link to Study 1 protocol); https://aspredicted.org/tiuw7.pdf (Study 2 protocol). To avoid unnecessary repetition—Study 2 was a close replication of Study 1—we present the methods and results of the studies together.

2.1. Samples

We sought to recruit 450 subjects in Study 1 and 900 subjects in Study 2. Subjects were supporters of the US Republican or Democratic Party, recruited via Amazon's Mechanical Turk (MTurk), an online labour market commonly used for psychological research (Arechar et al., 2018; Chandler & Shapiro, 2016; Rand, 2012). Subjects recruited on MTurk cannot be considered demographically representative of the wider US population; for example, they are more educated and more liberal/Democrat, among other demographic differences (Chandler & Shapiro, 2016). Despite this, there is evidence that political partisans recruited via MTurk are *psychologically* similar to partisans in nationally representative samples of US adults (Clifford et al., 2015). In particular, they score similarly on measures of personality and values related to political ideology (*ibid.*). Therefore, while there are documented constraints on the generalizability of results obtained from MTurk samples, the results of Clifford and colleagues (2015) suggest that subjects recruited via MTurk are not psychologically *incomparable* to the average US partisan.

The sample size for Study 1 was determined by power analysis (Faul et al., 2009), according to which we required N = 391 to detect an odds ratio of 1.4 in our primary binomial logistic regression analysis (key parameters: Two-tailed test; $\alpha = .05$; power = 0.9; Pr(Y=1|X=1) H0 = 0.5). We oversampled by approximately 15% to guard against power loss due to planned data exclusions. Sample size after data collection was $N_{S1} = 454$ in Study 1 (52.86% female, $M_{age} = 36.73$, $SD_{age} = 12.64$). Slight oversampling is the result of subjects not submitting their completion code on MTurk despite completing the study (i.e., meaning additional subjects were able to complete the study). In Study 2, we doubled the target sample size of Study 1 and recruited $N_{S2} = 900$ (59.33% female, $M_{age} = 36.28$, $SD_{age} = 11.40$).

2.2. Measures

2.2.1. Out-Party Hate

Subjects played a six-person behavioural economic game; the positive variant of the Intergroup Prisoner's Dilemma-Maximizing Difference (IPD-MD) game (Halevy et al., 2008; Weisel & Böhm, 2015). Each subject was assigned to a subgroup with two other supporters of the same party; that is, the *in-party*. This subgroup was matched with another subgroup of three supporters of the opposite party; the *out-party*. This formed a collective group of six players in total. Each player in the game was given three options about how to allocate money. Option one conferred 5 Monetary Units (MU, 1 MU = USD $1) to the focal subject, but nothing to any other players in the game. Option two conferred 2.5 MU to the subject and to each of their two in-party members, but nothing to the three out-party players. Option three conferred 2.5 MU to each player in the game. The decision options (as they were shown to Democratic Party subjects) are displayed in Figure 1. The decision option relevant to our hypothesis is Option 2. This decision option evinces a willingness to pay a personal cost to benefit the in-party (i.e., forsaking Option 1; self-interest), while simultaneously refusing to benefit members of the out-party at no extra cost to oneself or to members of one's in-party (forsaking Option 3; the collective interest). Following Weisel and Böhm (2015), we thus interpret decision Option 2 as an expression of out-party hate[14].

[14] In the preregistered protocols, we referred to this decision option as "parochial altruism". However, here we refer to it as "outgroup (out-party) hate" to clearly distinguish between *our* focus—which is simply those instances where ingroup "love" and outgroup "hate" appear in conjunction (such as in suicide terrorism and war)—and the parochial altruism *hypothesis*—which concerns the evolutionary origins of this conjunction. While the latter hypothesis has received recent criticism (e.g., Rusch et al., 2016; Yamagishi & Mifune, 2016), these criticisms do not contest the *existence* of ingroup love/outgroup hate, but, rather, the proposition that the conjunction of these behaviours manifests (i) consistently at the individual-level (i.e., as a within-individual correlation) and (ii) as a result of group-level selection pressure (for more detailed discussion, we refer to Rusch et al., 2016; Yamagishi & Mifune, 2016). We are grateful to an anonymous reviewer for emphasizing this point.

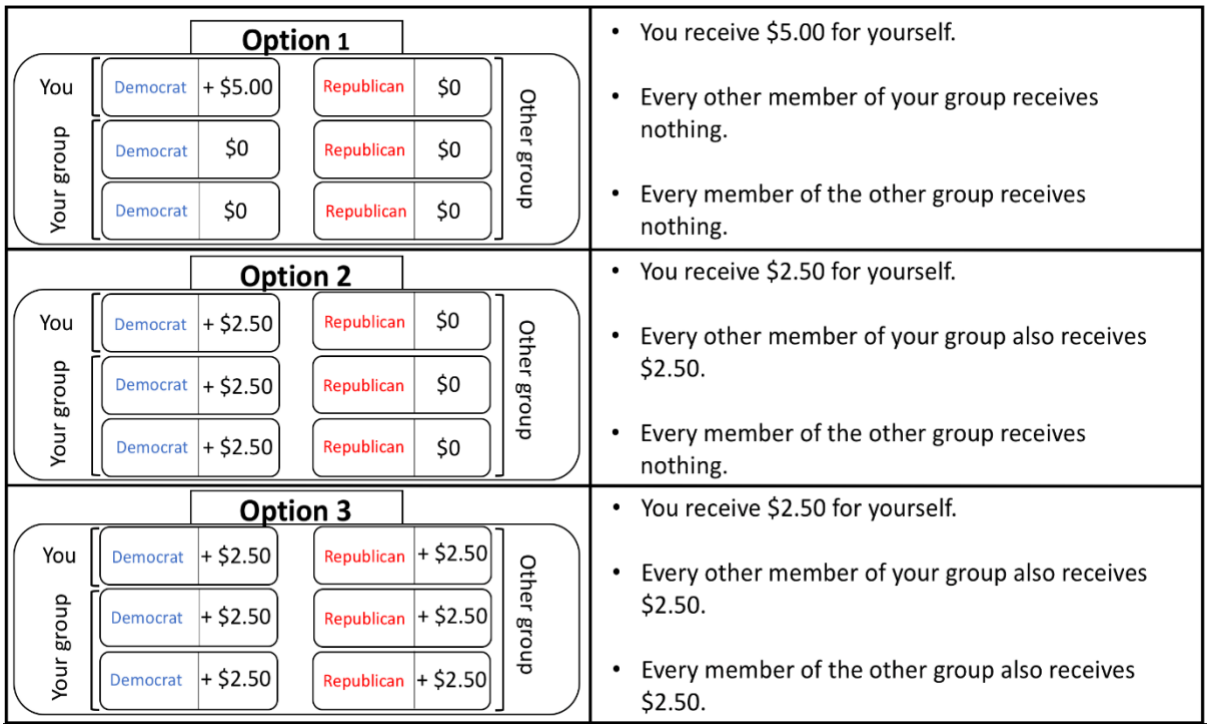


Figure 1. Decision Options in the Positive Variant of the Intergroup Prisoner's Dilemma Maximizing-Difference Game used in Studies 1 and 2. *Self-identified Democrats saw the displayed decision option screen. For subjects who identified as Republicans, the Republican and Democrat labels were reversed (i.e., Republicans were indicated as "your" group, and Democrats were indicated as the "other" group).*

2.2.2. Moral Polarization

To measure this variable, subjects completed a trait judgment task. Each subject was asked to judge the extent to which 5 positive and 5 negative moral traits described each of two targets: (i) the "average Democratic Party voter" and (ii) the "average Republican Party voter". Subjects also rated the social desirability of each trait. All trait ratings were provided on a 1-7-point scale. The target ratings were anchored from "Not at all" to "Very much so"; the desirability ratings were anchored "Very undesirable" to "Very desirable". The traits comprised personality descriptors such as *trustworthy*, *fair*, *manipulative* and *prejudiced*, and were embedded alongside a mix of 20 nonmoral traits. Table 1 displays the full list of traits

used in Studies 1 and 2. The traits used in Study 1 were taken from prior work (Tappin & McKay, 2017) and were chosen to represent the fundamental domains of social perception: morality, agency, and sociability (Leach et al., 2007). The traits used in Study 2 were slightly modified from the set used in Study 1. In particular, we replaced several of the traits with new traits from a large dataset of normed trait adjectives (Goodwin et al., 2014, Study 1). We did this to minimize any residual overlap between trait domains. That is, we wanted to *maximally* differentiate between the distinct trait domains of morality, agency and sociability.

Subjects' moral evaluation of each target (i.e., the average Democratic and Republican Party voter) was computed as the correlation between (i) their social desirability ratings for the moral traits, and (ii) their Democratic/Republican target ratings for the moral traits. Thus, each subject had two "coefficients of moral evaluation", describing the extent to which they ascribed desirable and undesirable moral traits to each target. Positive coefficient values indicate that the ascription of moral traits to the target *positively* correlated with the perceived desirability of those traits. In contrast, therefore, negative coefficient values indicate that the ascription of moral traits *negatively* correlated with the desirability of the traits. Because each subject rated the desirability of each trait, the coefficient values—representing subjects' moral evaluation of the targets—are sensitive to subjects' *idiosyncratic* beliefs about the desirability of the moral traits. This has the advantage of allowing for individual differences in which moral traits people consider more vs. less desirable when computing their coefficients of moral evaluation for each target. This is important because previous work suggests foundational differences in the moral preferences of Democrats (or "liberals") and Republicans (or "conservatives") (e.g., Graham et al., 2009).

Finally, whether the subject identified as Democrat or Republican informed which coefficient represented moral evaluation of the in-party (r_{inParty}) and out-party (r_{outParty}). For example, for a self-identified supporter of the Republican Party, r_{inParty} corresponded to the coefficient of moral evaluation for the Republican target, and r_{outParty} for the Democratic target (and vice versa for subjects who identified as supporters of the Democratic Party). The difference between these coefficients of moral evaluation ($r_{\text{inParty}} - r_{\text{outParty}}$) was taken as the discrepancy

in moral evaluation between the in-party and out-party; that is, the preregistered measure of *moral polarization*[15].

Table 1. Traits used in Studies 1 and 2.

Trait domain	Positive traits	Negative traits
Morality		
	Honest	Insincere
	Trustworthy	Prejudiced
	Fair	Disloyal
	Respectful (Just)	Manipulative (Violent)
	Principled	Deceptive (Greedy)
Agency		
	Hardworking (Intelligent)	Lazy
	Knowledgeable	Undedicated (Incompetent)
	Competent (Organized)	Unintelligent (Unproductive)
	Creative	Unmotivated
	Determined	Illogical (Weak)
Sociability		
	Sociable	Cold (Negative)
	Cooperative (Playful)	Disagreeable
	Warm (Happy)	Rude (Reckless)
	Family-orientated (Funny)	Humorless
	Easygoing	Uptight

Note. Traits outside of parentheses are used in Study 1. Traits inside parentheses replaced the preceding traits in Study 2.

2.2.3. Other Variables

We collected additional variables after the behavioural economic game and trait rating task. These variables were collected for the purpose of secondary preregistered and exploratory analyses. First, we asked each subject which of the three decision options they believed their two in-party members, and three out-party members, had chosen. Second, we asked subjects to rate the extent to which they believed that their out-party (i) threatened the "power, resources, or safety of the US and its citizens", and (ii) threatened the "values or identity of the

[15] In the preregistered protocols, we referred to this variable as "inframoralization". However, here we changed the label to "moral polarization" for descriptive clarity and consistency with concepts as defined in closely relevant work (Iyengar et al., 2012; Iyengar & Westwood, 2015). The variable is unchanged in all other respects. We are grateful to Mark Brandt for emphasizing the relevance of this work to the present investigation.

US and its citizens" (Stephan et al., 2011). Lastly, we asked subjects to rate (iii) the extent to which they believe the Democratic and Republican Party are in "direct competition". Ratings for (i), (ii), and (iii) were provided on 7-point Likert scales, anchored from 1 = "Not at all" to 7 = "Very much so".

2.3. Procedure

The procedure in both studies was substantively identical and we recruited unique samples in each (i.e., subjects who took part in Study 1 were prevented from taking part in Study 2). All subjects provided informed consent, before completing a brief screening questionnaire. This questionnaire identified whether the subject was a supporter of the Democratic Party or the Republican Party, and included other demographic questions such as age, gender, religious affiliation, and ethnicity. Importantly, subjects were not made aware of the specific purpose of the screening questionnaire (to minimize false responding). Subjects who identified with either the Democratic Party or Republican Party were eligible to continue with the study, whereas supporters of a political party other than these (including "none") were directed to an end-of-study message and were unable to continue. The Study 1 sample was skewed Democrat (Study 1 = 67.62% Democrat). In Study 2, we balanced the number of Democrats and Republicans by recruiting approximately equal numbers of each (Study 2 = 50.11% Democrat).

Eligible subjects then completed the trait judgment task. They judged the extent to which each of 30 traits (see Table 1) described (i) the "average Democratic Party voter" and (ii) the "average Republican Party voter". They also rated (iii) the social desirability of the traits. Subjects rated all 30 traits according to either (i), (ii), or (iii), before moving onto the next set of ratings, and the order of these three sets of judgments was counterbalanced across subjects. The presentation order of the traits themselves was randomized across each rating set and subject.

Following this task, subjects took part in the economic game. They read instructions detailing the structure of the game and were shown an example set of decisions (and the resultant pay offs). Those who identified as Republican were presented with instructions specifying two other Republicans as their subgroup members (and three Democrats as members of the other subgroup), and vice versa for Democrats. After these instructions, subjects made their decision about which option to choose (i.e., Option 1, 2, or 3). We informed them that six individual

decisions (three from Democrats, three from Republicans) would be combined, and the calculated bonuses paid out to one group of six—selected at random—after the survey had ended (which was true). After making their own decision, each subject indicated which decision they believed each of the other players had chosen, and responded to the threat and competition questions described above. Finally, at the end of the study, subjects were asked whether they had adequately understood the economic game before making their decision (yes/no), and they provided feedback on the study. In addition to any bonuses, all subjects were paid a base fee of $1 for taking part.

Results

All analyses were conducted in the R environment (v. 3.4.0, R Core Team, 2017), using R Studio (v. 1.1.423, RStudio Team, 2016). The R packages used in data analysis were: *scales* (v. 1.0.0, Wickham, 2018), *coin* (v. 1.2-2, Hothorn et al., 2008), *gridExtra* (v. 2.3, Auguie, 2017), *ggthemes* (v. 3.4.0, Arnold, 2017), *dplyr* (v. 0.7.7, Wickham et al., 2018), *ggplot2* (v. 3.0.0, Wickham, 2016), *reshape* (v. 0.8.7, Wickham, 2007), *plyr* (v. 1.8.4, Wickham, 2011), *metafor* (v. 2.0-0, Viechtbauer, 2010) and *datatable* (v. 1.10.4-3, Dowle & Srinivasan, 2017). The raw data and analysis scripts to reproduce the results and figures reported in this paper are available online via the project hub on the Open Science Framework: https://osf.io/mceqh/. Because Study 2 was a close replication of Study 1, after reporting each of the study-specific effect size estimates in the primary and sensitivity analyses, we also report the corresponding *meta-analytic* estimate (i.e., computed across studies). We note that (i) all meta-analytic estimates are fixed effects estimates, and (ii) the meta-analyses were not preregistered.

3.1. Key Descriptive Statistics

Table 2 displays the frequency and corresponding percentages of choices made in the IPD-MD game in Studies 1 and 2 (the full samples are displayed i.e., before any data exclusions—see section 3.2 below for details of data exclusions). Table 3 displays the median values of the coefficients of moral evaluation. In particular, we display the median coefficients pertaining to evaluation of the Democrat and Republican targets, separately for Democratic- and Republican-identifying subjects. Also displayed are the median coefficient values pertaining to in-party and out-party targets (that is, collapsing across Democratic and Republican

targets/subjects, as described in the Methods). We compare the coefficients using Wilcoxon signed-rank tests. The distributions of the coefficients are shown in Figure 2. The values in Table 3 and Figure 2 reveal a robust discrepancy in moral evaluation for the in-party vs. out-party—that is, strong evidence of moral polarization—among both Democratic- and Republican-identifying subjects.

Table 2. Economic Game Decisions in Studies 1 and 2.

Study	Subject Political Affiliation	Self-Interest (Option 1)	Out-Party Hate (Option 2)	Collective Interest (Option 3)	Total
Study 1					
	Republican	40 (27.2%)	22 (15.0%)	85 (57.8%)	147 (100%)
	Democrat	72 (23.6%)	54 (17.7%)	179 (58.7%)	305 (100%)
	Total	112 (24.8%)	76 (16.8%)	264 (58.4%)	452 (100%)
Study 2					
	Republican	113 (25.3%)	72 (16.1%)	262 (58.6%)	447 (100%)
	Democrat	131 (29.6%)	57 (12.9%)	255 (57.6%)	443 (100%)
	Total	244 (27.4%)	129 (14.5%)	517 (58.1%)	890 (100%)

Note. The numbers outside parentheses are frequencies and the numbers inside parentheses are row-wise percentages. N=2 observations are missing from Study 1 and N=10 observations from Study 2 due to missing values for choice decision in the economic game.

Table 3. Median Coefficients of Moral Evaluation in Studies 1 and 2.

Study	Subject Political Affiliation	Moral Evaluation					
		Dem. Target	Rep. Target	(Pseudo) Difference	In-Party Target	Out-Party Target	(Pseudo) Difference
Study 1							
	Republican	-.34	.89	-0.94***			
	Democrat	.88	-.50	1.10***			
	Combined				.88	-.46	1.04***
Study 2							
	Republican	-.34	.85	-0.91***			
	Democrat	.90	-.42	1.08***			
	Combined				.88	-.37	0.99***

*Note. Values for targets are the median correlations between ratings of trait desirability and trait ascription. The pseudo-difference is computed by Wilcoxon signed-rank test. Study 1 N = 442 (N = 12 subjects could not be included due to uniform responding on the trait judgment task); Study 2 N = 874 (N = 26 subjects were not included due to uniform responding and/or missing values on the trait judgment task). * p <.05; ** p <.01; *** p <.001.*

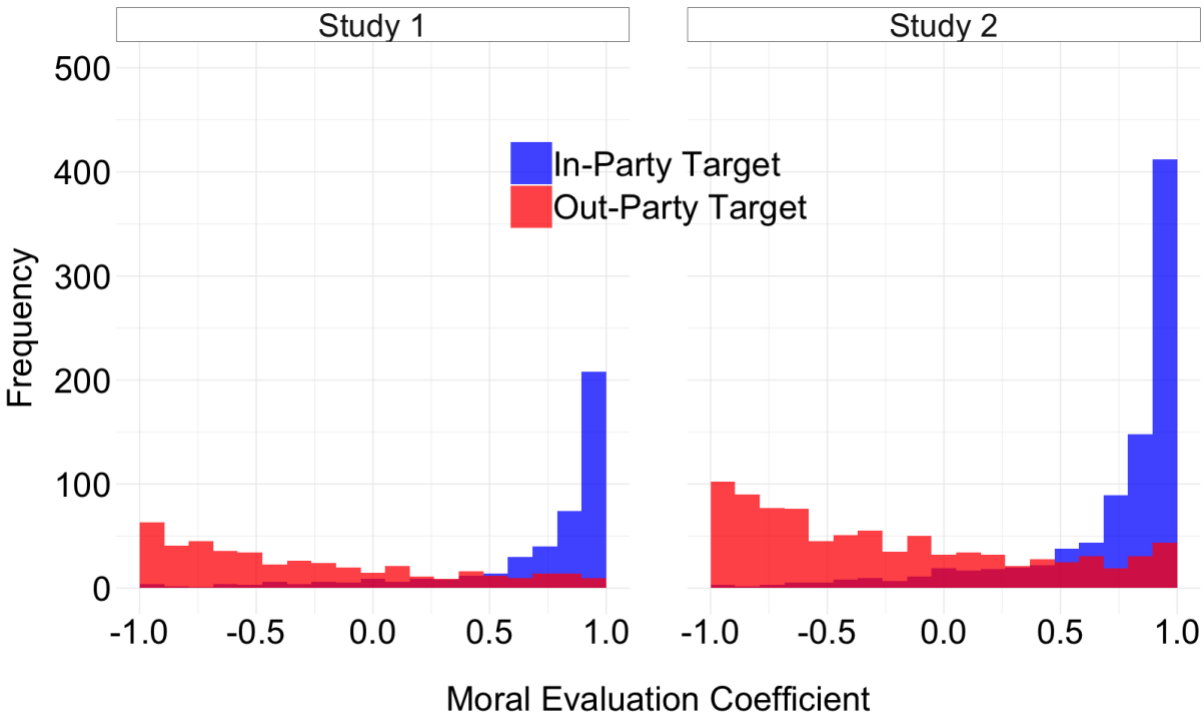


Figure 2. Distribution of the Coefficients of Moral Evaluation for In-Party and Out-Party Targets in Studies 1 and 2. *Each subject has one value corresponding to the in-party target and one value corresponding to the out-party target. Study 1 N = 442; Study 2 N = 874.*

3.2. Data Exclusions

As specified in the preregistered protocols, for the primary analysis we excluded subjects who fulfilled one or more of several criteria. First, we excluded those who failed one or more of three attention checks that were embedded in the trait judgment task (N_{S1} = 23 (5.07%); N_{S2} = 28 (3.11%)). Second, those who provided incomplete data in the trait judgment task (N_{S1} = 1 (0.22%); N_{S2} = 7 (0.78%)) or IPD-MD game (N_{S1} = 2 (0.44%); N_{S2} = 10 (1.11%)). Third, those who clicked through the IPD-MD game instructions too quickly to read them; defined as a recorded page submission time of less than 10 seconds on one or more of three instructions pages (N_{S1} = 74 (16.30%); N_{S2} = 175 (19.44%)). Fourth, those who reported that they did not understand the IPD-MD game instructions (N_{S2} = 8 (1.76%); N_{S2} = 17 (1.89%)). Fifth, and

finally, those subjects who responded *uniformly* on the trait judgment task—that is, recorded zero variance for any type of moral trait judgment (i.e., ratings for the Democratic target, Republican target, and/or social desirability) (N_{S1} = 12 (2.64%); N_{S2} = 22 (2.44%)); this was necessary because a lack of variance prevents correlation coefficients—required for the key measure of moral polarization (see Methods)—from being computed.

In addition to these preregistered exclusion criteria, we identified and excluded duplicate responses (i.e., multiple responses from the same subject) via subjects' unique MTurk IDs (N_{S1} = 1 (0.22%); N_{S2} = 26 (2.89%)). After all data exclusions, we thus retained N_{S1} = 354 and N_{S2} = 671 for the primary preregistered analyses.

3.3. Primary Preregistered Analyses

We fitted binomial logistic regression models to the data. Recall that the outcome of interest is choosing Option 2—expression of out-party hate—in the IPD-MD game (coded 1; the other two choice options were coded 0). The predictor variable is moral polarization, computed as the difference between subjects' coefficient of moral evaluation for the in-party and that for the out-party ($r_{inParty}$ - $r_{outParty}$); higher values therefore correspond to relatively greater moral polarization (M_{S1} = 1.08, SD_{S1} = 0.69; M_{S2} = 0.96, SD_{S2} = 0.73, possible range = [-2, 2]). Consistent with our hypothesis, moral polarization was positively associated with out-party hate in the IPD-MD game, in both studies: Odds Ratio$_{S1}$ (OR_{S1}) = 1.73, p = .027, 95% CI [1.06, 2.81]; OR_{S2} = 1.51, p =.025 [1.05, 2.17]. These odds ratios are plotted in Figure 3 (indexed by "Primary Preregistered" on the y-axis). The meta-analytic OR was 1.59, p =.002 [1.19, 2.12].

According to the models, subjects at the upper limit of moral polarization – i.e., a score of 2 (indicating a coefficient value of +1 for the in-party and -1 for the out-party) – had a predicted probability of 0.21 (Study 1) and 0.17 (Study 2) of expressing out-party hate, respectively. In contrast, subjects whose moral evaluation of the in-party and out-party were similar – a score of 0 on the moral polarization variable (indicating no difference in coefficient values for the in-party and out-party) – had a predicted probability of 0.08 of expressing out-party hate (in both Study 1 and Study 2).

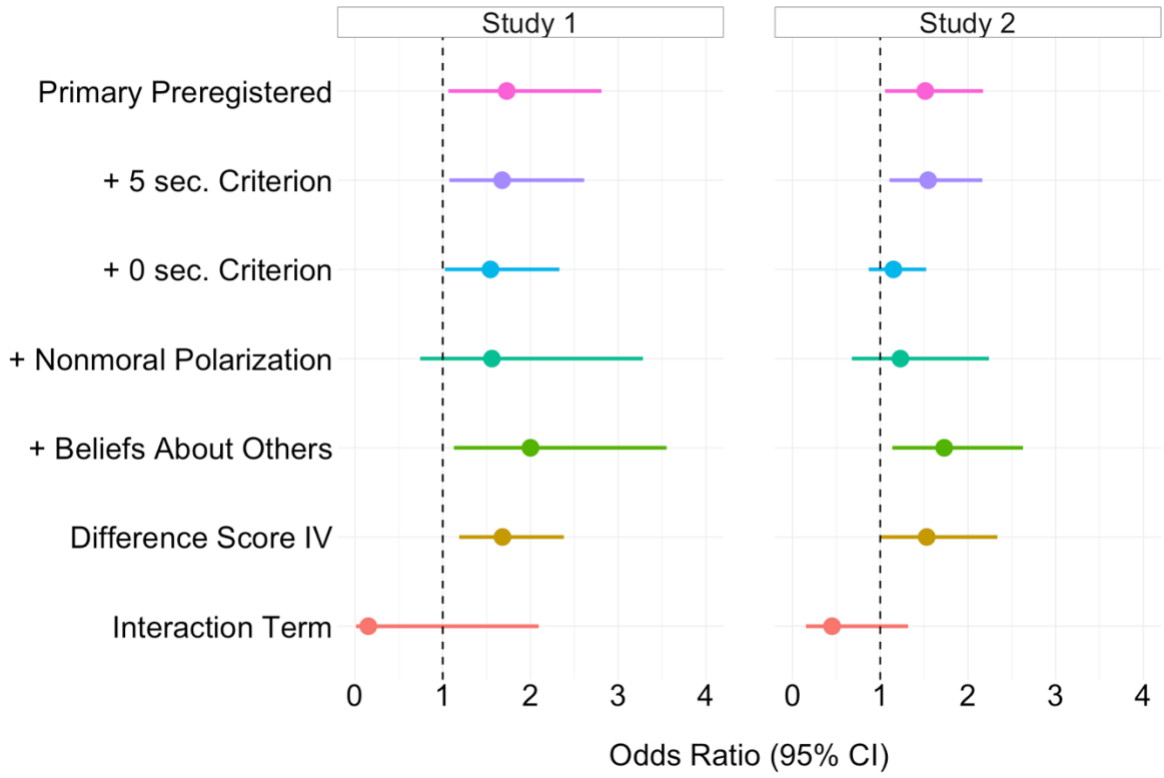


Figure 3. Results of Preregistered and Exploratory Sensitivity Analyses in Studies 1 and 2. *Panels display Odds Ratios (with 95% confidence intervals) from different models with moral polarization predicting out-party hate. Models are indexed on the y-axis.*

3.4. Sensitivity Analyses

We conducted a series of preregistered and exploratory sensitivity analyses as a check on the robustness of the primary result. These are reported below.

3.4.1. Instructions Page Exclusion Criterion

As reported in the data exclusions subsection (3.2), the number of subjects excluded for clicking through one or more of the IPD-MD game instructions too quickly (< 10 seconds) was relatively high in both studies. We thus repeated the primary preregistered analyses after implementing a more conservative exclusion criterion. Specifically, in one exploratory analysis we reduced this exclusion criterion to < 5 seconds (i.e., "5 sec. Criterion" models), and, in another, we removed this particular criterion altogether ("0 sec. Criterion" models).

Consequently, in the 5 sec. Criterion models, the sample sizes increased to N_{S1} = 388 and N_{S2} = 751, respectively. As plotted in Figure 3 ("5 sec. Criterion"), in these models the ORs for the moral polarization variable remained similar in size and statistically significant. The meta-analytic OR was 1.59, p <.001 [1.22, 2.08]. In the "0 sec. Criterion" models, sample sizes increased to N_{S1} = 411 and N_{S2} = 805. The ORs for the moral polarization variable in these models are also plotted in Figure 3 ("0 sec. Criterion"). In contrast to Study 1, in Study 2 the OR decreased noticeably in size and was no longer statistically significant (p >.05). The meta-analytic OR was 1.26, p =.049 [1.00, 1.59]. Overall, we conclude that the primary preregistered result was robust to more conservative specifications of the instructions page exclusion criterion.

3.4.2. Nonmoral Polarization

Recall that subjects also rated *non*moral traits in the trait judgment task, corresponding to the domains of agency and sociability. We preregistered our intention to investigate whether moral polarization was associated with out-party hate independent of polarization in these nonmoral domains of evaluation. We thus computed polarization scores for traits in the agency and sociability domains (in the same fashion as the moral polarization variable was computed), and entered these new variables as additional predictors in the primary preregistered models. As can be seen in Figure 3 ("Nonmoral Polarization"), the confidence intervals on the ORs for moral polarization increased in size after modelling the two nonmoral polarization variables; thus, the ORs were no longer statistically significant (at p <.05) in either study. The meta-analytic OR was 1.35, p =.207 [0.85, 2.15]. This complicates the inference that moral polarization *per se* – distinct from *non*moral polarization – was associated with out-party hate in the IPD-MD. We return to this point in the discussion.

3.4.3. Beliefs about Others

Recall that, after subjects made their own choice in the IPD-MD, they reported their beliefs about what option each other player in the game—their two in-party members, and three out-party members—had chosen. Previous research suggests that patterns of ingroup favouritism are underpinned by beliefs about the differential behaviour of one's ingroup members vs. outgroup members (Brewer, 1999; Yamagishi et al., 1999). We thus preregistered our intention

to investigate whether moral polarization was associated with out-party hate distinct from subjects' beliefs about the behaviour of the other players.

We created two new variables for this analysis. To create these variables, we first dummy-coded whether the subject believed that each in-party and out-party member expressed out-party hate (coded 1) or not (0). We then summed these dummy variables separately for the in-party and out-party members: producing one score between 0-2, indexing the subjects' belief about the number of *in-party* members expressing out-party hate (M_{S1} = 0.72, SD_{S1} = 0.84; M_{S2} = 0.66, SD_{S2} = 0.81); and another score between 0-3, indexing subjects' belief about the number of *out-party* members expressing out-party hate (M_{S1} = 0.84, SD_{S1} = 1.08; M_{S2} = 0.80, SD_{S2} = 1.08). We entered these two new variables as additional predictors in the primary preregistered models. The ORs from these models are plotted in Figure 3 ("Beliefs about Others"), and show that the association between moral polarization and out-party hate in the IPD-MD slightly increased in size (and remained statistically significant) in both Study 1 and 2. The meta-analytic OR was 1.82, p <.001 [1.30, 2.55]. Interestingly, these models revealed that subjects' beliefs about the expressed out-party hate of their two in-party members (but not out-party members) strongly predicted their *own* expression of out-party hate. We return to this result in a later analysis (section 3.5.1).

3.4.4. Difference Score of Moral Polarization

Recall that the preregistered measure of moral polarization was based on correlations between the ascription of moral traits to the Democratic/Republican targets, and the social desirability ratings given to the traits. As outlined in the Methods, this had the advantage of allowing for individual differences in which moral traits subjects considered more vs. less desirable when we computed their coefficient of moral evaluation for each target. Nevertheless, an arguably simpler measure of moral polarization is a difference-in-differences score that first takes the raw difference between the positive and negative moral traits ascribed to each target, and then compares these differences.

We created such a measure in three steps. In step 1, we computed the mean moral trait rating for each target (in-party, out-party), split by the valence of the traits (positive, negative). In step 2, for each subject we computed the difference between their mean positive and mean negative trait rating for each target. Thus, if the difference in means was > 0, subjects ascribed

positive traits more strongly than negative traits (on average) to that target; if the difference in means was < 0, the reverse was true. Finally, in step 3 we subtracted the out-party difference score from the in-party difference score to produce an alternative measure of moral polarization. As with the preregistered measure, therefore, higher values corresponded to relatively greater moral polarization (M_{S1} = 3.84, SD_{S1} = 3.29, $range_{S1}$ = [-3.6, 12]; M_{S2} = 3.81, SD_{S2} = 3.31, $range_{S2}$ = [-11.8, 12]). This new measure was strongly correlated with the preregistered measure of moral polarization, r_{S1} (440) = .82, p <.001, 95% CI [.79, .85]; r_{S2} (872) = .85, p <.001 [.83, .86].

We fitted binomial logistic regression models with this measure as the predictor variable and out-party hate as the outcome variable (as before). Before fitting the model, we rescaled this new measure of moral polarization to lie between [-2, 2], to facilitate comparison with the preregistered measure of moral polarization. In both Study 1 and 2, the ORs were similar in size compared to the primary preregistered analysis, and remained statistically significant at p <.05 (plotted in Figure 3, "Difference Score IV"). The meta-analytic OR was 1.62, p <.001 [1.24, 2.12].

3.4.5. Interaction Term

Recall that the preregistered measure of moral polarization is a single value (per subject) indexing the difference between moral evaluation of the in-party and moral evaluation of the out-party. Use of this measure thus prevents identification of whether moral championing of the in-party, moral demonization of the out-party, or some combination of both is responsible for the association with out-party hate in the IPD-MD. We therefore tested the *interaction* between the two constituent variables of the moral polarization index—that is, the interaction between moral evaluation of the *in*-party and moral evaluation of the *out*-party—in predicting expression of out-party hate. This is arguably a more appropriate test of our key hypothesis. In these models, odds ratios consistent with our hypothesis would be less than 1 (because the upper limit of moral polarization is defined by coefficients of +1 for the in-party and -1 for the out-party). In both Studies 1 and 2, the ORs on the interaction term were < 1—consistent with our hypothesis and the primary preregistered result—but in both cases the confidence intervals overlapped zero; indicating that the ORs were not statistically significant (at p <.05, plotted in Figure 3, "Interaction Term"). The meta-analytic OR was 0.39, p =.060 [0.14, 1.04].

3.5. Additional Exploratory Analyses

Our data afforded a series of exploratory analyses regarding supplemental questions of interest. These are reported below. We note that, for each of these exploratory analyses, we exclude only those respondents with missing values on the relevant variables, as well as duplicate IDs.

3.5.1. Beliefs about In-Party Behaviour

As revealed in the sensitivity analysis of our primary preregistered result, subjects' beliefs about the out-party hate expressed by *in-party* members strongly predicted their *own* out-party hate in the IPD-MD. To examine this relationship distinct from the moral polarization variable, we fitted a binomial logistic regression model where the outcome variable was out-party hate (dummy coded as usual); and the only two predictor variables were belief about the number of (i) in-party members (0-2), and (ii) out-party members (0-3) expressing out-party hate. As in the analysis with moral polarization, the former predictor variable was strongly associated with out-party hate in both studies: $OR_{S1} = 5.44$, $p < .001$, 95% CI [3.67, 8.06]; $OR_{S2} = 4.81$, $p < .001$ [3.64, 6.36]. In other words, the belief that one's in-party members expressed out-party hate shared a strong positive association with expressing out-party hate oneself. Figure 4 displays the data upon which the models are based, and illustrates the starkness of the result. We consider this result further in the discussion. In contrast to beliefs about the in-party, subjects' beliefs about the number of *out*-party members expressing out-party hate did not significantly predict their own expression of out-party hate, in either study: $OR_{S1} = 1.08$, $p = .583$ [0.83, 1.40]; $OR_{S2} = 1.09$, $p = .379$ [0.90, 1.32].

3.5.2. The Ideological "Prejudice Gap"

Our data contribute to debate over the ideological "prejudice gap" (Brandt et al., 2014; Sibley & Duckitt, 2008). In particular, the ideological-conflict hypothesis (Brandt et al., 2014) predicts that people on the ideological left and ideological right exhibit approximately *symmetrical* levels of prejudice toward groups that hold values at odds with their own; as contrasted against the hypothesis of a left-right *a*symmetry in prejudicial behavior (Sibley & Duckitt, 2008). We tested these competing hypotheses by comparing rates of out-party hate between Democratic-identifying and Republican-identifying subjects. To maximize statistical power, we pooled the data from Study 1 and Study 2 before conducting this comparison

(combined N = 1,354). Among Democratic-identifying subjects, N = 111 (14.6%) expressed out-party hate; among Republican-identifying subjects, N = 94 (15.8%) expressed out-party hate. According to a chi-squared test, the difference was not statistically significant: χ^2 (1) = 0.18, p =.673. This result is inconsistent with the prejudice gap (left-right asymmetry) hypothesis.

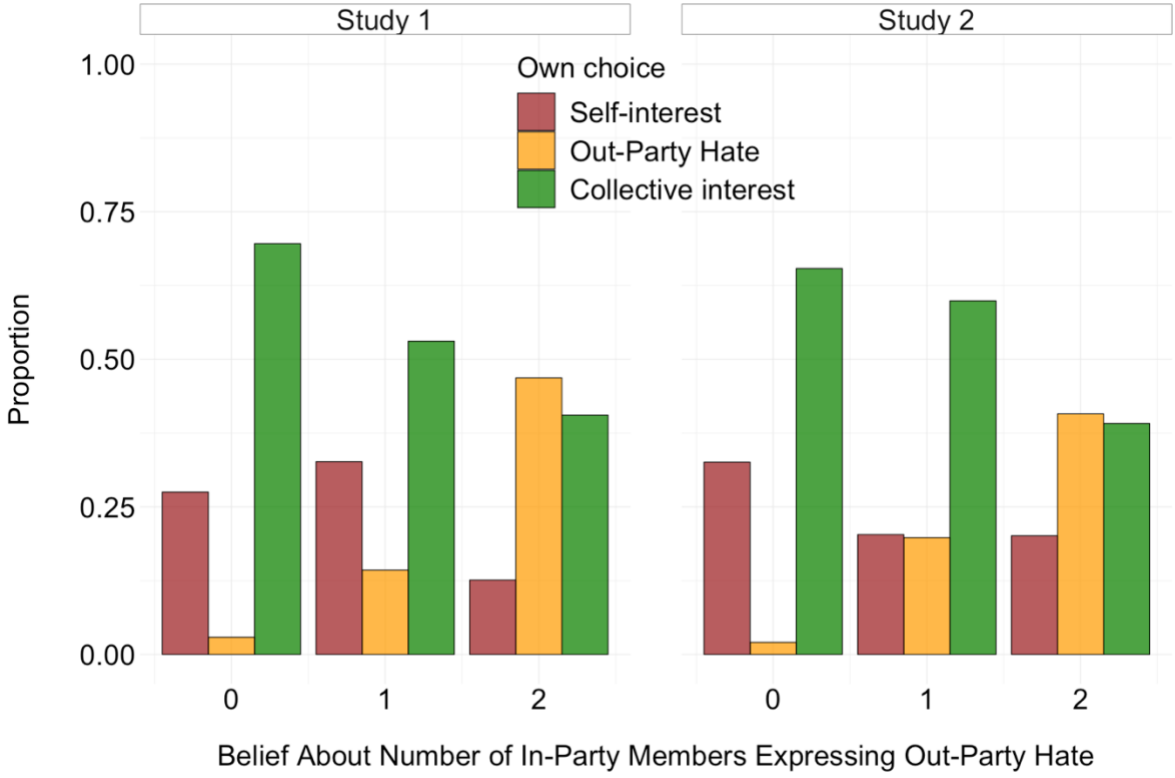


Figure 4. Proportion of Choices in the IPD-MD Game as a Function of Subjects' Beliefs about the Number of In-Party Members Expressing Out-Party Hate. *Study 1 N = 449 (belief 0 group N = 240; belief 1 group N = 98; belief 2 group N = 111); Study 2 N = 864 (belief 0 group N = 488; belief 1 group N = 192; belief 2 group N = 184).*

3.5.3. Perceived Threat Posed by the Out-Party

We examined the association between the perception that the out-party posed a *threat* to the United States and its citizens and moral evaluation of the out-party. Recall that we collected two threat perception variables from subjects (both scored from 1-7); one concerning the "realistic" threat posed by the out-party i.e., threat to the power, safety, and resources of the US, and the other concerning "symbolic" threat; that is, threat to the values and identity of the US. The two variables were strongly correlated: r_{S1} (451) = .78, p <.001, 95% CI [.74, .81]; r_{S2} (872) = .83, p <.001 [.81, .85]. Thus, we combined them into a single threat perception variable by taking their mean (variable: Perceived threat). Perceived threat was strongly negatively correlated with moral evaluation of the out-party (i.e., with the coefficient of moral evaluation for the out-party target), in both studies: r_{S1} (441) = -.58, p <.001 [-.64, -.52]; r_{S2} (850) = -.51, p <.001 [-.56, -.46]. In other words, more negative beliefs about the moral character of the out-party were associated with a stronger belief that they posed a threat to the safety and values of the US and its citizens.

3.5.4. Moral vs. Nonmoral Polarization

Recall that we measured nonmoral traits (as well as moral traits) in the trait judgment task. We sought to compare the magnitude of moral polarization to the magnitude of *non*moral polarization among partisans. We did this in two ways. First, we computed the mean rating given on each individual trait (i.e., for each trait in Table 1) as they were ascribed to each target (in-party, out-party). These mean ratings are plotted in Figure 5 as a function of the trait domain (agency, morality, sociability) and trait valence (negative, positive) (denoted by the faded small data points). We also plot the subsequent mean computed over these individual trait rating means (denoted by the solid large data points). As can be seen in the figure, across all trait domains, when rating the *in*-party target subjects ascribed *positive* traits more strongly than negative traits. However, across studies this valence gap appeared to be slightly larger in the agency and morality domains vs. the sociability domain. When rating the *out*-party target, in contrast, subjects tended to ascribe *negative* traits more strongly than positive traits; but only in the morality and sociability domains (in the agency domain, a similar valence gap was not evident).

To formally compare the domain-specific magnitudes of polarization, we conducted Wilcoxon signed-rank tests between the measures of polarization corresponding to each of the three trait domains. That is, we compared the preregistered measure of moral polarization with the

corresponding measures of polarization computed using the agency and sociability traits (e.g., see Methods section 2.2.2, and Results section 3.4.2). Recall that higher values on all these measures correspond to relatively greater polarization (that is, a greater in-party-favoring difference in trait evaluation). In Study 1, moral polarization (Median = 1.16, IQR = 1.26) was larger than polarization in the domain of agency (Median = 0.58, IQR = 1.03), $p < .001$; but not significantly different from polarization in the domain of sociability (Median = 1.16, IQR = 1.08), $p = .141$. In Study 2, in contrast, moral polarization (Median = 1.12, IQR = 1.24) was larger than polarization in *both* agency (Median = 0.61, IQR = 1.16) *and* sociability (Median = 0.99, IQR = 1.23) domains, $p < .001$ in both Wilcoxon signed-rank tests. These results corroborate the data plotted in Figure 5, and, taken together, show that moral polarization was larger in magnitude than polarization manifested in the nonmoral domains.

3.5.5. Robustness in Multiple Regression

Given the number of inter-correlated variables reported throughout the results section that plausibly share a relationship with out-party hate, in a final exploratory analysis we fitted a binomial logistic regression model where out-party hate was the dummy-coded outcome variable (as usual) and a *multitude* of variables were entered as predictors (including demographics). The odds ratios corresponding to each predictor variable in this joint model are displayed in Figure 6. As shown in the figure, across studies only beliefs about the in-party's expressions of out-party hate were robust in predicting subjects' *own* out-party hate behaviour.

Figure 5. Mean Trait Ratings as a Function of Party Target (In-Party, Out-Party), Trait Valence (Negative, Positive) and Trait Domain (Agency, Morality, Sociability) in Studies 1 and 2. *The faded small data points denote the mean rating for the individual traits in each valence/domain category (see Table 1 for the traits). The data points are horizontally jittered to aid visibility. Each individual trait mean is computed over N = 453 subjects in Study 1 and over N = 874 in Study 2. The solid large data points denote the mean computed over the individual trait rating means. Error bars are 95% CI.*

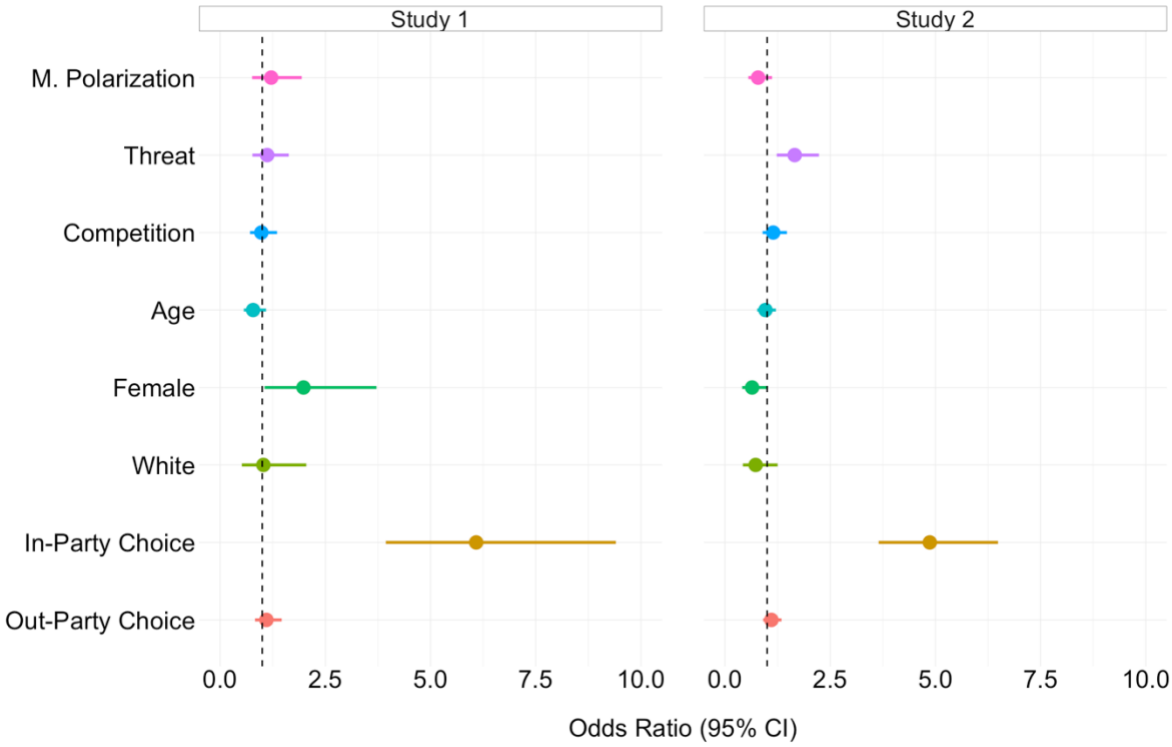


Figure 6. Results of Exploratory Multiple Regression Analyses in Studies 1 and 2. *Panels display Odds Ratios (with 95% confidence intervals) from binomial logistic regression models with multiple variables predicting out-party hate. Variables are indexed on the y-axis. M. Polarization = preregistered moral polarization variable; Threat = perceived threat posed by out-party; Competition = perceived competition between in-party and out-party; Age was measured in years; Female & White are dummy-coded; In-Party/Out-Party Choice = belief about number of in-party/out-party members expressing out-party hate. Threat, Competition, and Age are standardized and mean-centered. Study 1 N = 437; Study 2 N = 842.*

Discussion

We hypothesized that moral polarization would be associated with behavioural expressions of out-party hate in the US political context. In two studies, we tested this hypothesis with large samples of US partisans and a behavioural economic game measure of outgroup hate (Weisel and Böhm, 2015). The primary preregistered analyses were consistent with our hypothesis: Expressions of out-party hate increased in conjunction with moral polarization (meta-analytic

odds ratio = 1.59, 95% CI [1.19, 2.12]). In a series of subsequent preregistered and exploratory sensitivity analyses, we tested the robustness of this result. While these analyses indicated that the primary result was somewhat robust, they also highlighted important constraints on the inference that moral polarization is associated with out-party hate in the US political context. We consider the implications of these and our various other results below.

We observed important exceptions to the general robustness of our primary preregistered result. Most notably, after accounting for (i) polarization observed in *non*moral domains of evaluation (e.g., see Figure 4), and (ii) demographic and other relevant predictor variables (Figure 6), the association between moral polarization and out-party hate was reduced in size and statistically non-significant. In addition, the *interaction* between the constituent variables of the moral polarization index—that is, the interaction between moral evaluation of the *in*-party and moral evaluation of the *out*-party—did not convincingly cohere with the primary preregistered result (meta-analytic odds ratio = 0.39, 95% CI [0.14, 1.04]; recall that, for the interaction term, odds ratios < 1 are consistent with our hypothesis). The test of the interaction term is arguably the more appropriate test of our hypothesis—which was that behavioural expressions of out-party hate would be most common as the perceived moral "gap" between in- and out-party widened. Taking this together with the large sample sizes in our studies, and the rather *mild* form of out-party hate afforded by the IPD-MD, overall our results suggest that the association between moral polarization and out-party hate in the US political context is probably small and somewhat tenuous.

In this respect, our results converge with recent work on partisan prejudice and affective polarization. Lelkes and Westwood (2017) find evidence that even those partisans who are the most affectively polarized are generally unwilling to endorse overtly discriminatory behaviour against the political opposition. Our results extend their findings in two ways. First, by showing that the same pattern holds when using an incentivized, *behavioural* measure of out-party hate, rather than self-report (as those authors used). Second, we measured *moral* polarization, rather than generalized affective polarization. Given that (i) we found moral polarization to be greater in magnitude than polarization in *non*moral domains of evaluation (see section 3.5.4 and Figure 5), and (ii) affective polarization is greatest among partisans who view politics through a moral-psychological lens (Garrett & Bankert, 2018), it is reasonable to assume that ours was an even *easier* test for the affective polarization-partisan prejudice hypothesis to pass—and, yet, still it did not pass this test (at least, not convincingly).

Notwithstanding this convergence in findings, however, there is a particular limitation of our studies that warrants mention and precludes a strong interpretation of our results along the foregoing lines. That is, our sampling population. We recruited subjects from Amazon's Mechanical Turk, a survey platform whose subjects are known to fall short of demographically-representing the wider US population (Chandler & Shapiro, 2016). As highlighted in the introduction, Mason (2016, 2018) finds that US *party* identity is increasingly in alignment with demographic identities (e.g., race, religiosity), and, importantly, that this alignment may serve to weaken barriers to out-party hostility (Mason & Wronski, 2018; Roccas & Brewer, 2002). For this reason, insofar as our subjects did not faithfully represent the demographic identities of the wider US population, it is possible that our analyses mis-estimated the population-level association between moral polarization and out-party hate in IPD-MD. Ultimately, though, we consider this minimally problematic for our overall interpretation of our results, given that (i) there is no evidence that a more faithful demographic representation would have strengthened the hypothesized association—it may just as well have *attenuated* it—and (ii) the results of Lelkes and Westwood (2017), that converge with our own, *are* based on representative samples of US adults.

In contrast to the equivocal association between moral polarization and out-party hate, we observed compelling evidence of moral polarization *per se*. That is, on both the preregistered measure (see Table 3 & Figure 2) and exploratory trait-summary measure (Figure 5), moral polarization among US partisans was large and robust. Moreover, we found evidence that moral polarization was greater than polarization observed in the *non*moral domains of evaluation (see section 3.5.4). We consider two interpretations of these results as they relate to the phenomenon of affective polarization (Iyengar et al., 2012; Iyengar et al., 2018).

One interpretation is that moral polarization is simply a more proximate indicator of whatever underlying construct is manifesting as affective polarization. For example, assuming that domain-general partisan animosity is the underlying construct (e.g., as in Lelkes & Westwood, 2017), one would expect moral polarization to be stronger than nonmoral polarization for the reason that moral traits share a stronger relationship with liking and respecting other people/groups than do nonmoral traits (Hartley et al., 2016). In other words, partisans express their dislike of the out-party and liking of the in-party through whichever route is available; and moral (vs. nonmoral) evaluation just happens to be more "cathartic" in this sense.

On the other hand, it seems clear that moral evaluation also *causes* the (dis-)liking of other people/groups. This is implied by a long programme of research showing that moral content guides—and, in fact, *dominates*—humans' global evaluations of other people and groups, primarily for the reason that the moral character/benevolence of others can have a very direct and consequential impact on one's own wellbeing (reviewed in Wojciszke, 2005). On this view, affective polarization *in general* may be a function of moral polarization *in particular*. That is, partisans "like" the in-party and "dislike" the out-party in part *because* the former are perceived to be fairer, more trustworthy, less prejudiced—in other words, more *benevolent*—than the latter. This perspective accords well with the distinct role of moral psychology in contemporary American politics, as outlined in the introduction (Brady et al., 2017; Koleva et al., 2012; Ryan, 2014, 2017), as well as the apparent moderating effect of moral conviction on affective polarization (Garrett & Bankert, 2018). Unfortunately, which of the foregoing interpretations is ultimately correct cannot be determined on the basis of the current data. Future work might adjudicate by experimentally assigning moral and nonmoral characteristics to in- and out-party targets, and observing affective polarization. At the very least, though, our results *do* illustrate that estimates of the magnitude of affective polarization depend non-trivially on the method by which evaluation of the in- and out-party is measured.

We observed a strong positive association between subjects' beliefs about the number of *in-party* members expressing out-party hate and their *own* expression of out-party hate (e.g., see Figure 4 & section 3.5.1). Though this association was observed in exploratory analyses—and must be interpreted as such—we note that the relevant odds ratios in both studies survive Bonferroni-corrections of 1×10^{14} to the p-values; implying that the association is rather robust. We offer two explanations for this intriguing result. The first—and we think more likely—explanation is that subjects *projected* their own behaviour in the IPD-MD onto their judgment of what the in-party members would do. A long line of research demonstrates that people engage in "social projection" of this kind when asked to make information-deprived judgments of other people (reviewed in Krueger, 2008; Robbins & Krueger, 2005). The logic behind the utility of social projection is that—because most people are in the majority most of the time—projection allows people to make quick and reasonably accurate judgments (on average) about unknown others (Krueger, 2008; Krueger & Chen, 2014). Indeed, in our studies subjects received only sparse information about the other players (i.e., only their political party membership); providing good conditions for social projection.

An alternative explanation for the result is that subjects tailored their own out-party hate behaviour to what they believed the other players in the IPD-MD would do. Specifically, to whether they believed the *in-party* would express out-party hate; akin to a reciprocation- or conformity-type effect. We think this explanation is less likely than social projection. Primarily because a large body of evidence shows that projection to *in*group members is typically greater than projection to *out*group members (for a meta-analysis, see Robbins & Krueger, 2005). This is strongly consistent with our results, where subjects' beliefs about the out-party hate expressed by *out-party members* were only trivially associated with their own out-party hate behaviour (see section 3.5.1 & Figure 6). The alternative explanation—that is, the notion that subjects tailored their behaviour to the expected behaviour of the other players—appears less able to explain this non-association. This is because, assuming this alternative explanation is right, one would expect that beliefs about the expressed out-party hate of the *out-party members* would, to some extent at least, also affect subjects' own choice to express out-party hate. For example, it is reasonable to expect that they would be positively correlated—reflecting a desire for "pre-emptive strike" (Böhm et al., 2016; Simunovic et al., 2013). That we did not observe such an association provides some evidence that subjects were not tailoring their own out-party hate behaviour to what they believed the other players in the IPD-MD would do.

Regardless of which explanation is actually right, the result itself highlights a potentially fruitful avenue by which to predict—ahead of time and with reasonable accuracy—the out-party hate behaviour of partisans. That is, query whether they believe that the typical in-party member would express out-party hate. This may be a particularly useful strategy to identify those most likely to express out-party hate where there exist disincentives to answering in the affirmative oneself. We leave it to future research to explore this idea.

In this paper, we investigated the association between moral polarization—that is, the tendency for people to view opposing partisans' moral character negatively, and co-partisans' moral character positively—and behavioural expressions of out-party hate in the US political context. Our results strike an optimistic chord: Taken together, they suggest that the hypothesized association is probably small and somewhat tenuous. Though moral polarization *per se* was large—and may exceed prior estimates of generalized affective polarization—in our sample even the *most* morally polarized partisans appeared reluctant to engage in a rather mild form

of out-party hate behaviour. These findings converge with recent evidence that polarization—moral or otherwise—has yet (at the time of data collection) to translate into the average US partisan wanting to actively harm their out-party counterparts.

1.4. Doing Good vs. Avoiding Bad in Prosocial Choice: A Refined Test and Extension of the Morality Preference Hypothesis[16]

Abstract

Prosociality is fundamental to human social life, and, accordingly, much research has attempted to explain human prosocial behavior. Capraro and Rand (*Judgment and Decision Making*, *13*, 99-111, 2018) recently provided experimental evidence that prosociality in anonymous, one-shot interactions (such as Prisoner's Dilemma and Dictator Game experiments) is not driven by outcome-based social preferences—as classically assumed—but by a generalized morality preference for "doing the right thing". Here we argue that the key experiments reported in Capraro and Rand (2018) comprise prominent methodological confounds and open questions that bear on influential psychological theory. Specifically, their design confounds: (i) preferences for efficiency with self-interest; and (ii) preferences for action with preferences for morality. Furthermore, their design fails to dissociate the preference to do "good" from the preference to avoid doing "bad". We thus designed and conducted a preregistered, refined and extended test of the morality preference hypothesis (N=801). Consistent with this hypothesis, our findings indicate that prosociality in the anonymous, one-shot Dictator Game is driven by preferences for doing the morally right thing. Inconsistent with influential psychological theory, however, our results suggest the preference to do "good" was *as* potent as the preference to avoid doing "bad" in this case.

[16] The work presented in this section was conducted in collaboration with Valerio Capraro and is published in *Journal of Experimental Social Psychology*: https://www.sciencedirect.com/science/article/pii/S0022103118302841

Introduction

People often pay costs to benefit others; they behave *prosocially*. Fundamental to human social life (Fehr & Gächter, 2002; Gintis et al, 2003; Nowak, 2006; Tomasello, 2014), prosocial behavior is often explained by appeal to reciprocity. If I pay a cost to help *you* today, you – or others who learn about my behavior – are more likely to help *me* tomorrow (Nowak & Sigmund, 2005; Rand & Nowak, 2013; Trivers, 1971). Defying explanations of this kind, however, prosocial behavior is frequently observed in contexts where opportunities for reciprocity are absent. For example, in anonymous, one-shot interactions, individuals often forego some amount of self-interest to the benefit of strangers (Camerer, 2003).

Behavioral economists have classically sought to explain such behavior by assuming that individuals have preferences for minimizing inequity or maximizing efficiency (i.e., social welfare) (Bolton & Ockenfels, 2000; Capraro, 2013; Charness & Rabin, 2002; Engelmann & Strobel, 2004; Fehr & Schmidt, 1999). According to these influential frameworks, prosocial individuals derive utility – psychological benefit – from particular social *outcomes*; thus, realizing those outcomes offsets the cost of behaving prosocially.

A recent alternative perspective is that individuals derive utility from performing *actions* they perceive to be morally right (Bicchieri, 2005; DellaVigna et al., 2012; Huck et al., 2012; Krupka & Weber, 2013). This perspective accords with evidence from social psychology that individuals derive utility from seeing themselves in a positive moral light (Aquino & Reed, 2002; Dunning, 2007) and, in addition, that prosocial individuals in particular view opportunities for prosocial action in moral terms; for example, by considering what the morally "right" action is (Liebrand et al., 1986; Weber et al., 2004).

Building on these converging lines of evidence, recent experimental work advanced the hypothesis that a *generalized morality preference* – rather than preferences for minimizing inequity or maximizing efficiency per se – drives prosocial behavior in anonymous, one-shot interactions (Capraro & Rand, 2018). In other words, that a simple preference for doing (what is perceived to be) the morally "right" thing underpins individuals' prosociality in these contexts.

In their key experiments, Capraro and Rand (2018) used a "Trade-Off Game" (TOG) to empirically dissociate the hypothesized morality preference from outcome-based social preferences for equity and efficiency. In the TOG, participants made a unilateral choice about how to allocate money between themselves and two other (passive) people. While one choice minimized inequity – all participants earned the same amount – the other choice maximized efficiency – participants earned different amounts, but, together, the group earned more. This design effectively pitted preferences for equity and efficiency against one another; creating a decision context where the morally "right" choice was ambiguous. The researchers found that, framing *either* choice as the morally appropriate one dramatically affected participants' choices, such that the majority chose the option framed as morally appropriate; be that the equitable *or* efficient choice.

To support the inference that these moral considerations drive prosociality, however, required additional evidence. To that end, participants also completed, in addition to the TOG, a canonical prosocial choice task; either the Dictator Game (DG), or the Prisoner's Dilemma (PD). In the latter tasks, participants made a unilateral choice about how much money to donate to a new (passive) person (DG), or a simultaneous bilateral choice whether to cooperate with a new person (PD), respectively.

The key finding in Capraro and Rand (2018) was that participants who made the choice framed as morally appropriate in the TOG – be that the equitable choice *or* the efficient choice – were consistently more prosocial in the DG and PD; donating and cooperating (respectively) more than participants who chose otherwise in the TOG. Crucially, this result is *inconsistent* with stable outcome-based preferences for equity or efficiency as explanations for prosociality, which do not predict an association between moral framing in the TOG and prosociality in a different task, such as the DG/PD. The result is instead consistent with the morality preference hypothesis, which predicts that individuals sensitive to which choice is morally right in the TOG – as revealed by the moral framing of those choices – are also revealed to be more prosocial in the DG/PD; where, in contrast to the TOG, the morally right choice is *unambiguous* (Krueger & Acevedo, 2007; Krueger & DiDonato, 2010).

The implication of Capraro and Rand's (2018) findings is important: They suggest their data renders the classic approach to understanding prosocial choice through social preferences insufficient and, in particular, that an account based on a fluid preference for "doing the morally

right thing" is superior. However, their key evidence derives from an experimental design that contains several prominent methodological confounds, and leaves open important questions regarding the mechanism of the hypothesized morality preference. Below we expand on these issues.

Self-Interest

Consider the choice outcomes in the TOG. The *equitable* choice always provided the participants – the chooser, and two passive recipients – the same allocation; 13 Monetary Units (MU) each. The *efficient* choice, in contrast, always provided the chooser with 15 MU, and the passive recipients 23 MU and 13 MU, respectively. Thus, while the efficient choice clearly results in greater overall gains for the group – at the cost of equity, as intended – it also results in *greater gains for the chooser themselves*. In other words, the choice option meant to reveal a preference for efficiency is confounded with self-interest. A plausible consequence of this confound is an *overestimate* of the proportion of individuals with a preference for efficiency. An overestimation of this kind may have affected the key result – an association between TOG choice and prosociality in the DG/PD – in two ways.

First, it may have *inflated* the association between TOG choice under the *equitable*-is-moral frame, and prosociality in the DG/PD. Specifically, this association may not have been driven by participants with a genuine morality preference – who choose the equitable option under this TOG frame, and the prosocial option in the DG/PD – but, rather, by self-interested participants – who choose the *efficient* option under this TOG frame, and the *self-interested* option in the DG/PD. Indeed, in the worst case, the behavior of self-interested participants could *fully account* for the observed association between TOG choice under the equitable-is-moral frame, and prosociality in the DG/PD.

Second, by the opposite logic, the overestimation of individuals with a preference for efficiency may have *deflated* the association between TOG choice under the *efficient*-is-moral frame, and prosociality in the DG/PD. This is because some participants making the efficient choice under that TOG frame did so *not* because of a general morality preference nudged by the framing, but, rather, for their own self-interest. Crucially, these participants would *not* have chosen prosocially in the DG/PD, thereby deflating the observed association between the two choices.

These issues directly affect the key evidence – an association between TOG choice and prosociality in the DG/PD – supporting the morality preference hypothesis. A remedy to these issues is to remove self-interest from the equation by design.

Action-Inaction Asymmetry

Not only do the *efficient*-is-moral and *equitable*-is-moral frames differ in the labels used to describe the two choice options, but, in addition, they differ in which is the *active* choice and which is the *passive* choice. Specifically, in the *efficient*-is-moral frame, participants start with an equitable allocation (13 MU each), while in the *equitable*-is-moral frame they start with an efficient allocation (15, 23, and 13 MU, respectively). In other words, the moral choice is always framed as an *active* choice to change these initial allocations. Choice frame is thus confounded with active/passive frame.

A substantial body of work in social, moral, and decision-making psychology indicates that humans perceive *inaction* differently than *action* (Baron & Ritov, 2004; Spranca et al., 1991). For example, regret is greater for actions that lead to negative outcomes than for *inactions* that lead to the same negative outcomes (Feldman & Albarracín, 2017; Zeelenberg et al., 2002); individuals are biased towards maintaining the status quo in decision-making (Samuelson & Zeckhauser, 1988); and, in moral judgment, harms caused by *action* are considered worse than the same harms caused by *inaction* (Cushman et al., 2006). Finally, most relevant here, action framing influences engagement in prosocial behavior (Teper & Inzlicht, 2011), and there is considerable variation in *who* exhibits action-inaction asymmetries (Baron & Ritov, 2004).

Given this evidence, it is probable that the confounding of choice frame with active/passive frame over- or under-estimated the proportion (and types) of individuals choosing the morally-framed option in the TOG; with unknown consequences for the key association between TOG choice and prosociality in the DG/PD. Decoupling these frames is necessary to make clear inferences about the effect of choice frame in the TOG.

Doing Good vs. Avoiding Bad

An influential hypothesis in social psychology is that immoral, negative, or otherwise "bad" stimuli weigh more heavily than their "good" counterparts in human cognition and behavior

(Baumeister et al., 2001; Rozin & Royzman, 2001; Vaish et al., 2008; see Corns, 2018 for a recent critique).

Consistent with this hypothesis, recent evidence suggests that "self-righteousness" – manifested in, for example, the *average person* rating themselves morally superior to the average person (Tappin & McKay, 2017) – are greater for immoral than moral stimuli (Klein & Epley, 2016, 2017). Relatedly, the correlation between individuals' life satisfaction and their self-perception is reportedly stronger if the latter is computed as the distance between individuals' "real" and "undesired" selves vs. between their "real" and *"desired"* selves (Ogilvie, 1987). In other words, those data suggest the type of person individuals want to *avoid* being weighs more heavily (in their life appraisal) than the type of person they would ideally like to be. A similar asymmetry manifests in the psychology of moral regulation. In particular, in the distinction between *pro*scriptive morality – what we *should* do and be – and *pre*scriptive morality – what we should *avoid* doing and being. Whereas the former is considered discretionary and a matter of personal preference, the latter is considered mandatory and strict (Janoff-Bulman et al., 2009).

To implement the choice framing in the TOG, the two choice options were *jointly* framed as moral and immoral, respectively. Individuals choosing the morally-framed option may thus have been motivated by a preference to do "good" (e.g., a desire to be moral), or motivated by a preference to avoid "bad" (e.g., an aversion to being immoral). These distinct preferences are confounded in the TOG design. Given the preceding evidence, it is plausible that individuals' choices were motivated more by a preference to avoid "bad" than to do "good". Furthermore, assuming this hypothesis, a further plausible hypothesis is that participants who were motivated by a preference to do "good" (vs. avoid "bad") in the TOG were more likely to behave prosocially in the DG/PD. For example, because the preference to avoid bad may reflect a general desire to avoid punishment, whereas a preference to do good may reflect a desire to do good for its own sake. That is, the latter preference is more diagnostic of true prosocial motivation.

The Current Study

Here we address the methodological confounds and open theoretical questions in Capraro and Rand (2018), and, thus, provide a refined and extended test of the morality preference

hypothesis. To do so, we design and implement an improved Trade-Off Game (TOG), and test for an effect of choice frame on TOG choice (Hypothesis 1), and for an association between framing of the TOG and prosociality in a different task, the DG (Hypothesis 3). We also test two novel hypotheses bearing on existing psychological theory. First, that the effect of choice frame on TOG choice is greater under an avoid "bad" than do "good" moral frame (Hypothesis 2) (cf. Baumeister et al., 2001; Rozin & Royzman, 2001; Vaish et al., 2008). Second, that the association between choosing the morally-framed option in the TOG, and prosociality in the DG, is greater under a do "good" than avoid "bad" moral frame (Hypothesis 4).

Methods

The hypotheses, design, sampling and analysis plan were preregistered on the Open Science Framework (protocol: https://osf.io/9nphs/).

Participants

We sought to collect N=200 participants per treatment, giving a total N=800. We determined this sample size by multiplying the N-per-treatment in Capraro and Rand (2018) study 3 by 1.5x; the study most conceptually similar to that which we reproduce here. Sensitivity power analyses (reported in SM) for our key hypothesis tests indicated we had sufficient power (>.80) to detect standardized effect sizes conventionally considered small (r =.10). A total of N=801 participants completed the study. Participants were recruited online via Amazon's Mechanical Turk (AMT) (for the validity of AMT, see e.g., Arechar et al., 2018; Horton et al., 2011; Paolacci et al., 2014; Thomas & Clifford, 2017), and were located in the US at the time of taking part. All participants provided informed consent. This study was reviewed and approved by Middlesex University, and Royal Holloway, University of London ethics procedures.

Procedure

Participants began by playing a Dictator Game (DG). In the DG, they were given $0.10 and they had to decide how much, if any, to give to another anonymous participant who received no starting money allocation. Participants could donate in increments of $0.01; from $0.00 to $0.10. The participant was informed that the other person had no active choice and would only

receive what they decide to give. We asked two comprehension questions to ensure that participants understood the payoff structure of the DG prior to their decision. Specifically, we asked which choice (1) maximized their *own* payoff, and which choice (2) equalized their payoff with that of the other person. Participants who failed either or both comprehension questions were prevented from completing the survey (this condition was made explicit in the consent form). Those who passed the comprehension questions were then asked to make their DG decision.

Following the DG, participants played an improved Trade-Off Game (TOG). In this TOG, participants ("choosers") had to decide between two choice options that affected their own payoff and the payoff of two other people; the latter being passive recipients who did not make any choices. One option was "equitable", in the sense that it minimized payoff differences among the three participants; specifically, they each earned $0.13. The other option was "efficient", in the sense that it maximized the sum of the payoffs of the three participants; specifically, the chooser earned $0.13, while the other two people earned $0.23 and $0.13, respectively. Importantly, in this improved TOG design, because the chooser earns $0.13 by making *either* choice, the confounding of self-interest with preferences for efficiency is eliminated. Furthermore, because participants are not told that one or the other state of money distribution ([13, 13, 13] or [13, 23, 13]) initially holds, *both* choice options are rendered equal in terms of active/passive frame.

Before reading the TOG instructions, participants were randomly assigned to one of four versions of the TOG, each corresponding to a particular framing combination in a 2x2 between-subjects design:

- TOG frame: **Give – Do Good**
- TOG frame: **Give – Avoid Bad**
- TOG frame: **Equalize – Do Good**
- TOG frame: **Equalize – Avoid Bad**

We experimentally manipulate whether the efficient ("Give") or equitable ("Equalize") choice is framed as morally appropriate (choice frame[17]: Give, Equalize). Furthermore, we also manipulate whether the moral framing emphasizes doing "good" or avoiding "bad" (moral frame: Do Good, Avoid Bad):

- Under the **Give – Do Good** frame, the efficient option is labelled "be generous, Option 1", and the equitable option is "Option 2"
- Under the **Give – Avoid Bad** frame, the efficient option is labelled "Option 2", and the equitable option is "be ungenerous, Option 1"
- Under the **Equalize – Do Good** frame, the efficient option is labelled "Option 2", and the equitable option is "be fair, Option 1"
- Under the **Equalize – Avoid Bad** frame, the efficient option is labelled "be unfair, Option 1", and the equitable option is "Option 2"

Importantly, notice that the experimental manipulation of moral frame decouples the preference to do "good" from the preference to avoid "bad". After making their decision in the TOG, participants provided standard demographic information, at the end of which they were given the completion code needed to submit the survey on AMT. After the end of the survey, we downloaded the data file and computed the bonuses, which were paid on top of the base participation fee received by all participants ($0.50). No deception was used. We refer to the SM for verbatim experimental instructions (available here: https://osf.io/m7w2s/). We report all measures, manipulations and exclusions.

Results

Data analysis was conducted in R (v.3.4.0, R Core Team, 2017) using RStudio (v.1.1.423, RStudio Team, 2016). R packages used in analysis and figures: *ggplot2* (v.2.2.1, Wickham, 2009), *plyr* (v.1.8.4, Wickham, 2011), *dplyr* (v.0.7.4, Wickham et al., 2017), *reshape* (v.0.8.7, Wickham, 2007), *gridExtra* (v.2.3, Auguie & Antonov, 2017), *effsize* (v.0.7.1, Torchiano,

[17] Following Capraro & Rand (2018), in our preregistered protocol and analysis script we labelled the efficient-is-moral frame the "Give" frame, and the equitable-is-moral frame the "Equalize" frame. For consistency, we follow that convention here.

2017), *datatable* (v.1.10.4-3, Dowle et al., 2017). The raw data and code to reproduce all results and figures in this paper are available at https://osf.io/x5stj/.

Data Exclusions

N=288 (26.45%) participants answered one or more of the comprehension questions incorrectly, or did not answer these questions, and were thus prevented from completing the study (following our preregistered protocol). Of the remaining N=801 participants who completed the study, there were N=15 (1.87%) duplicate responses according to participants' IP address/unique Mechanical Turk ID. In line with our preregistered protocol, we excluded these duplicates, retaining the earliest responses only—defined by the date/time they began the study. Finally, after these exclusions, we identified N=2 (0.25%) participants that dropped out of the study prior to making their decision in the TOG, and are thus unable to be included in the analysis (leaving N=784 for analysis).

Hypotheses 1 & 2

Preregistered analyses. We first test whether participants were more likely to choose the efficient option in the TOG under the "give" choice frame than under the "equalize" choice frame (Hypothesis 1). We then test whether this framing effect was stronger under the "avoid bad" moral frame than under the "do good" moral frame (Hypothesis 2). To that end, we fit a binomial logistic regression model with two dummy-coded treatment variables as predictors: choice frame [0=equalize frame, 1=give frame] and moral frame [0=avoid bad, 1=do good], and choice in the TOG as the dependent variable [TOG choice: 0=equitable choice, 1=efficient choice]. Consistent with Hypothesis 1, choice frame predicted TOG choice in the expected direction, Odds Ratio (OR) = 3.55, 95% CI [2.34, 5.40], Z = 5.94 p <.001. A majority of participants chose the efficient option under the give choice frame (69.21%), whereas only a minority of participants chose this option under the equalize choice frame (39.90%).

Inconsistent with Hypothesis 2, there is no statistically significant interaction between choice frame and moral frame, OR = 0.91 [0.50, 1.64], Z = -0.32, p =.752. In other words, the effect of choice frame appeared largely independent of whether the choice was framed as "doing good" or "avoiding bad". Both H1 and H2 results remain similar after adjusting for age, gender [not female=0, female=1] and education [0=less than college, 1=college or above] in the model:

Main effect of choice frame OR = 3.74 [2.45, 5.72], Z = 6.11, p <.001; interaction between choice frame and moral frame OR = 0.89 [0.49, 1.61], Z = -0.39, p =.697. The proportion of choices in each of the four treatments is displayed in Figure 1.

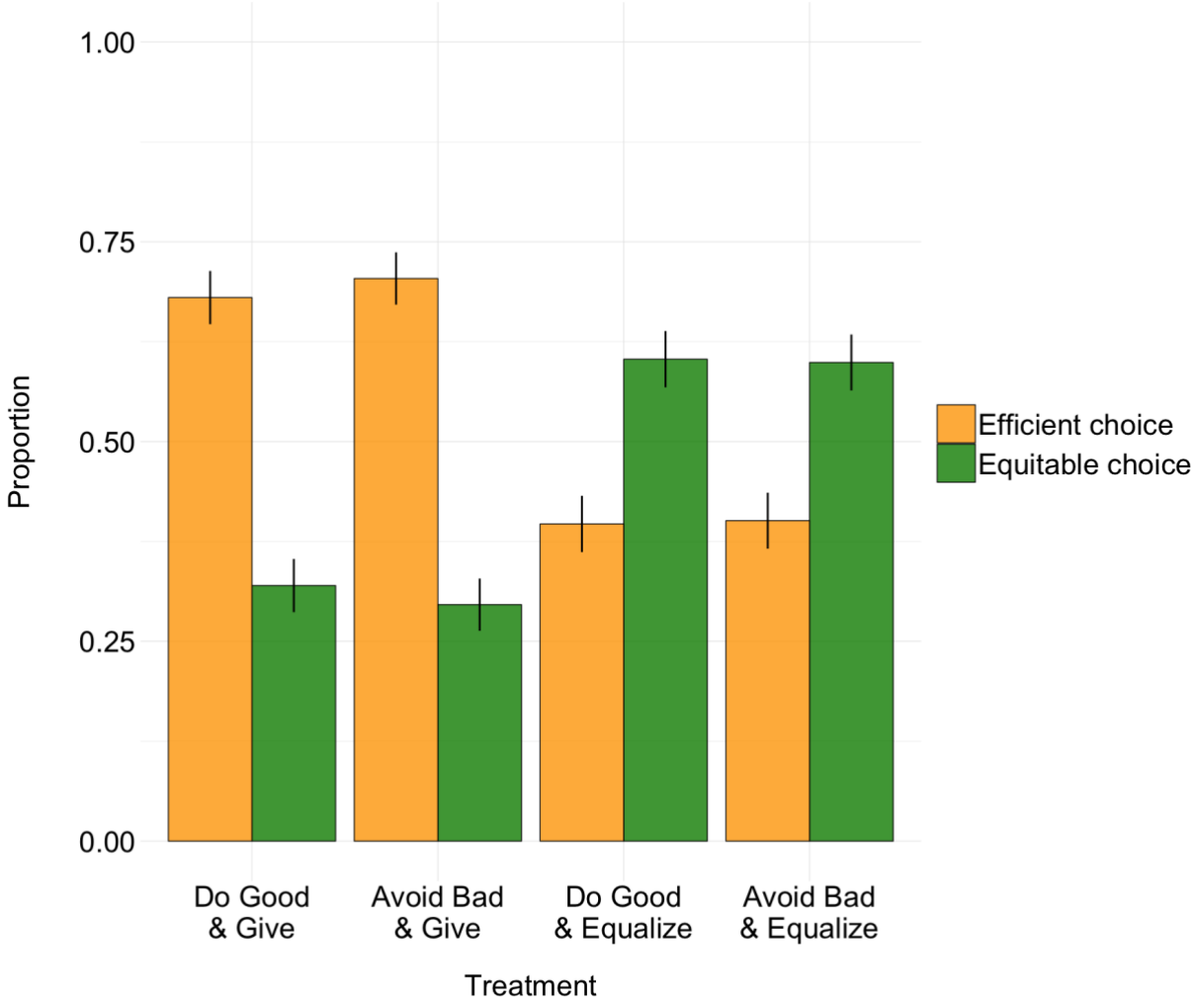


Figure 1. Proportion of efficient and equitable choices as a function of treatment. *Error bars are ± 1 SEM.*

Hypothesis 3

Preregistered analyses. Next, we test the hypothesis that participants who make the 'moral' choice in the TOG – that is, choose the efficient option under the "give" frame, or choose the equitable option under the "equalize" frame – donate more to their partner in the DG (Hypothesis 3). We fit a linear regression model with two dummy-coded variables as predictors: TOG choice frame [0=equalize frame, 1=give frame] and the choice the participant made in the TOG [0=equitable choice, 1=efficient choice], respectively. The DV is amount donated in the DG [from 0 to 10].

Consistent with Hypothesis 3, there is an interaction in the predicted direction, b = 1.68, SE = 0.40, t = 4.17, p <.001. Under the equalize frame, participants who made the equitable choice donated more in the DG (M = 3.40, SD = 2.60) than participants who made the efficient choice (M = 2.38, SD = 2.57), t (389) = 3.81, p <.001, hedges' g = 0.39, 95% CI [0.19, 0.60]. This pattern was reversed under the give frame. There, participants who made the equitable choice donated *less* in the DG (M = 2.53, SD = 2.75) than participants who made the efficient choice (M = 3.19, SD = 2.80), t (391) = -2.19, p =.029, hedges' g = -0.24 [-0.45, -0.02]. (Note: All t-tests are post-hoc.) This interaction between choice frame and TOG choice remains similar after adjusting for age, gender, and education, b = 1.67, SE = 0.40, t = 4.14, p <.001. The interaction pattern is displayed in Figure 2. As seen in Figure 2 (panel A), DG donations follow an approximately bimodal distribution peaking over donations of 0 and 5. We thus conducted exploratory analyses to test the robustness of the preceding linear regression results (reported in SM). These were consistent with the linear regression models.

Hypothesis 4

Preregistered analyses. In our final preregistered analysis, we test whether the difference in DG donations between participants who made the 'moral' vs. 'non-moral' choice in the TOG is larger under the "do good" frame than under the "avoid bad" frame (Hypothesis 4). In line with our preregistered protocol, to simplify this analysis, we collapse across two variables: TOG choice frame [equalize, give], and the TOG choice the participant made [efficient, equitable]. This provides a new binary variable denoting whether the participant made the moral choice in the TOG [0=no, 1=yes], where the moral choice is simply defined as *either* the efficient option under the "give" frame, *or* the equitable option under the "equalize" frame. We then fit a linear regression model with two variables as predictors: moral choice [0=no, 1=yes],

and moral frame [0=avoid bad, 1=do good]. As before, the DV is amount donated in the DG [0-10].

Inconsistent with Hypothesis 4, there is no statistically significant interaction between moral choice and moral frame on DG donations, b = -0.32, SE = 0.40, t = -0.79, p =.429. In other words, while participants who made the moral choice in the TOG tended to donate more in the DG than participants who made the non-moral choice (i.e., Hypothesis 3), this effect appeared largely independent of whether the participants made their TOG choice under the moral frame of "doing good" or "avoiding bad". Adjusting for age, gender, and education in the model did not meaningfully change this result, b = -0.33, SE = 0.40, t = -0.81, p =.418 (see SM for robustness checks).

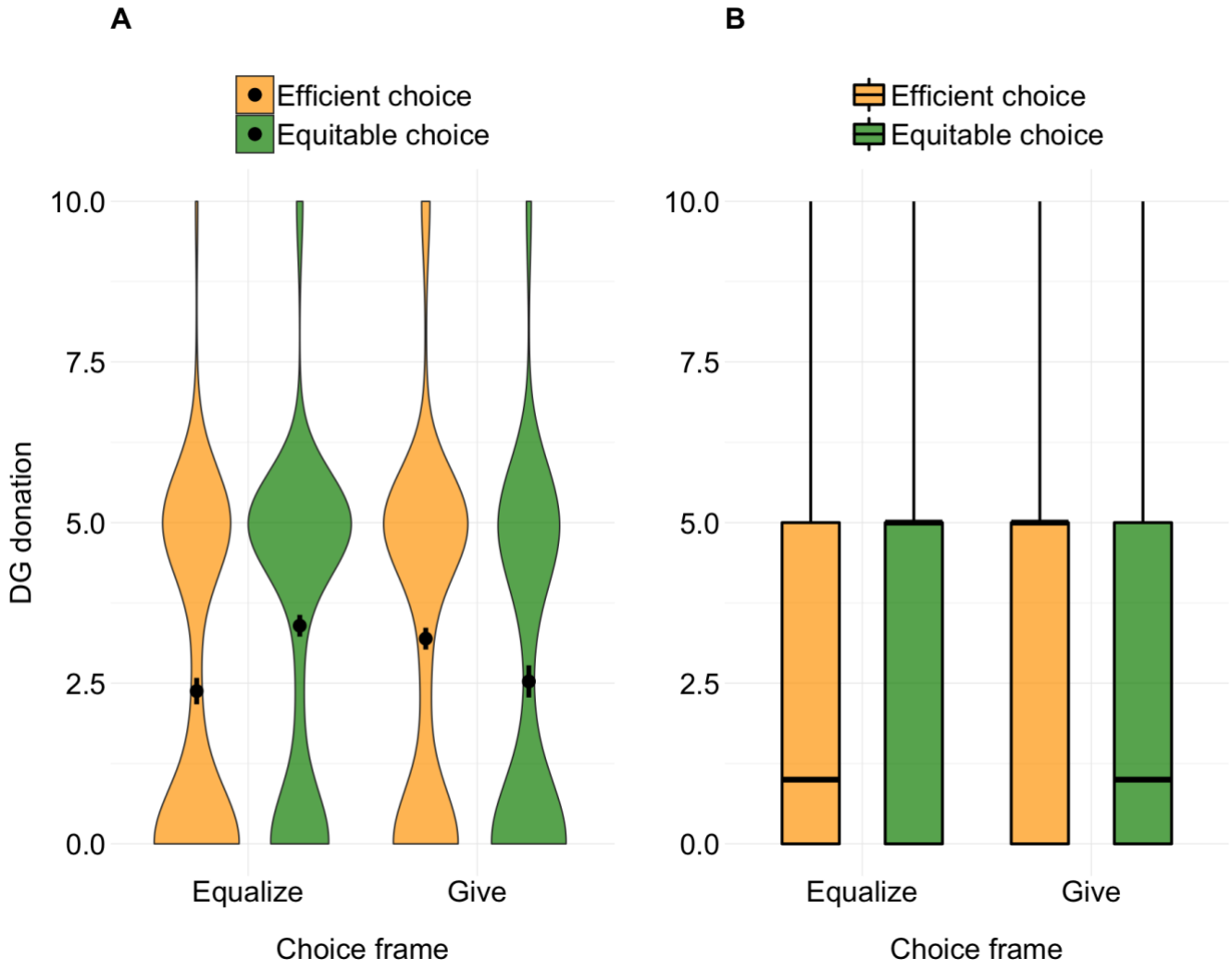


Figure 2. Violin plots (A) and boxplots (B) of DG donations as a function of choice frame, and the choice the participants made, in the TOG. A*, points denote the mean values and error bars are ± 1 SEM.* ***B****, bolded centre lines denote the median values.*

Discussion

Converging evidence suggests that prosociality in anonymous, one-shot interactions is not solely motivated by outcome-based social preferences, but that it is also motivated by what individuals perceive to be the morally right action (Bicchieri, 2005; DellaVigna et al., 2012;

Eriksson et al, 2017; Kimbrough & Vostroktunov, 2016; Krupka & Weber, 2013); perhaps serving to maintain a positive moral self-image (Aquino & Reed, 2002; Dunning, 2007). Building on this work, recent experimental evidence advanced the hypothesis that a generalized morality preference drives prosocial behavior in anonymous, one-shot interactions like that in Dictator and Prisoner's Dilemma games (Capraro & Rand, 2018). This hypothesis rejects the classic view in behavioral economics that prosociality in these situations is driven by social preferences for equity and efficiency.

Here we identified prominent methodological confounds and open theoretical questions in the key experiments reported in Capraro and Rand (2018). In particular, their Trade-Off Game (TOG) design (i) confounds preferences for efficiency with self-interest, and (ii) preferences for action with preferences for morality. Moreover, the design fails to dissociate preferences to do "good" from preferences to avoid doing "bad". It is highly likely these issues affected the observed association between choice in the TOG, and prosociality in the DG/PD; the key evidence for the morality preference hypothesis. Likewise, the failure to decouple the preference to do "good" from that to avoid doing "bad" leaves the mechanism of the proposed morality preference unclear, and misses a key prediction from influential psychological theory; that "bad is stronger than good" (Baumeister et al., 2001; Rozin & Royzman, 2001; Vaish et al., 2008).

To address these issues, we designed and implemented an improved TOG/DG experiment – eliminating the confounds identified in the original experiments reported in Capraro and Rand (2018). In doing so, we replicated the key results in support of the morality preference hypothesis. We found that framing one or the other TOG choice as morally appropriate – by labelling the focal choice "fair" or "generous", or the counterpart choice "unfair" or "ungenerous" – strongly affected individuals' choices. Specifically, approximately 70% of individuals chose the efficient option when that choice was framed as morally appropriate, dropping to 40% when the equitable choice was framed as morally appropriate; a swing of 30%, and a reverse in the majority decision. More importantly, we found that individuals who chose the morally appropriate option in the TOG – be that the efficient option *or* the equitable option – were more prosocial in the preceding DG; donating more money to a stranger. This result was robust to various analytic specifications, and provides evidence that prosociality (in the DG) is driven by a preference for doing what is perceived to be morally right.

Our results lend experimental support to various alternatives to outcome-based preference models (e.g., Alger & Weibull, 2013; Brekke et al., 2003; Kimbrough & Vostroktunov, 2016; Krupka & Weber, 2013; Lazear et al., 2012; Levitt & List, 2017). In particular, these models assume that individuals have moral preferences that guide their prosocial decision-making in unilateral interactions; an assumption consistent with the current findings. A parallel class of models has sought to explain prosocial decision-making using an intention-based framework, according to which people are sensitive to others' intentions (Falk et al., 2008; McCabe et al., 2003; Rabin, 1993). These models have been useful in explaining prosociality observed in interactions with more than one active player. However, they are of limited use in the case of unilateral interactions – which are the focus of the current study – where beliefs about the intentions of others do not apply. Reputation-based models (e.g., Heck & Krueger, 2017; Jordan et al., 2016; Nowak & Sigmund, 2005) also appear of limited use for our results because choices were anonymous and one-shot.

Our findings are consistent with work in social psychology that suggests individuals are motivated to maintain a positive moral self-image (Dunning, 2007). Indeed, an open question at the intersection of this research is whether individuals who assign greater value to a moral self-image (Aquino & Reed, 2002) show stronger choice framing effects in the TOG, and/or a stronger association between TOG choice and prosociality in the DG/PD. That said, we note that it is unlikely individuals are *solely* motivated by what they perceive to be the right thing, especially across different choice contexts. In line with Capraro and Rand (2018), we do expect that these results extend to the PD. There is some recent evidence, however, that behavior in other economic games is driven by outcome-based preferences, and *not* by a general morality preference (Capraro, 2018). An interesting avenue for future work is thus to further explore the boundary conditions of the morality preference account. One avenue might be situations in which people have to trade-off conflicting moral principles, like in the case of "altruistic" lying; that is, lying to benefit others (Biziou-van-Pol et al., 2015; Erat & Gneezy, 2012). An open empirical question is whether and how preferences for morality – as revealed by choice in the TOG – predict moral trade-offs of this kind.

We found that a moral frame emphasizing "good" affected TOG choice as strongly as a moral frame emphasizing "bad". For example, the proportion of individuals switching from the efficient choice to the equitable choice was essentially identical (approx. 30%) whether the latter choice was labelled "fair" or the former choice "unfair". We thus found little evidence

that "bad" (framing) was stronger than "good" (framing) in prosocial choice (cf. Baumeister et al., 2001).

An explanation for this discrepancy is found in recent work that has criticized the dominance of "bad" over "good" (termed the *negativity bias*) on theoretical and empirical grounds (Corns, 2018). A particular criticism concerns the credible alternative explanations for much of the evidence base. For example, asymmetries in perception that are taken to support the stronger effect of negatively-valenced stimuli may instead be explained by differences in the informativeness of negative vs. positive stimuli (Corns, 2018). In research on impression formation, as a case in point, the greater weight assigned to immoral (vs. moral) traits may be explained by the fact that immoral traits tend to be more *informative* of others' character (Kellermann, 1984). In the case of *self*-perception, in contrast, it is likely this informativeness bias does not hold; for example, because individuals can introspect on their full "moral history", unlike when they are judging other people. Absent a difference in informativeness, immoral and moral traits may yield similar impact on judgment and behavior.

This provides a plausible explanation for our results, and is consistent with a recent critique of the negativity bias hypothesis (Corns, 2018). This explanation also clarifies why the observed association between choosing the morally-framed option in the TOG, and prosociality in the DG, was similar under the "good" vs. "bad" moral frame (i.e., rejection of H4). Specifically, because the frames were *equally* motivating, individuals choosing the morally-framed option in either frame were *equally* liable to donate more in the DG. It is important to highlight that our "bad" frame label was *symmetric* vis-à-vis our "good" frame label (e.g., "unfair" vs. "fair", respectively). Had we used frames with stronger, but asymmetric labels – such as "steal" in place of "unfair" – we may have elicited an asymmetric moral frame effect in line with the negativity bias hypothesis. As noted above, however, in that case it may have been unclear whether the moral frame asymmetry was due to the valence of the frames per se (i.e., bad vs. good), or *other* differences such as asymmetric strength or salience of the labels (cf. Corns, 2018).

Finally, our results add to a growing body of research on framing effects in prosocial choice. Prior work finds that situational labels sometimes impact prosocial decision-making (e.g., Capraro & Vanzo, 2018; Kay & Ross, 2003; Krupka & Weber, 2013; Larrick & Blount, 1997; Liberman et al., 2004), but not always (Dreber et al., 2013; Goerg et al., 2017). This suggests

the effect of labels may depend on the specific interaction or game type. Our findings replicate recent work showing that labels are highly effective in the Trade-Off Game (Capraro, 2018; Capraro & Rand, 2018).

In summary, recent experimental work advanced the hypothesis that prosociality in anonymous, one-shot interactions is not driven by outcome-based social preferences for equity or efficiency per se, but by a generalized morality preference for "doing the right thing" (Capraro & Rand, 2018). We identified prominent methodological confounds and open theoretical questions in this work, and, consequently, conducted a refined and extended test of the morality preference hypothesis. Consistent with this hypothesis, our findings indicate that prosociality in the anonymous, one-shot Dictator Game is driven by preferences for doing the morally right thing. Furthermore, consistent with a recent critique of the negativity bias hypothesis, our results suggest the preference to do "good" was *as* potent as the preference to avoid doing "bad" in this case.

Supplemental Material

Sensitivity Power Analyses

Hypotheses 1 and 2 were tested via binomial logistic regression. H1 test is for a main effect (in the presence of another main effect and an interaction), while H2 test is for an interaction (in the presence of two main effects). In our sensitivity analysis, we assume that the two *non-focal* effects in each of the H1 and H2 tests jointly account for an R^2 of .10. We note that increasing the R^2 to .50 (i.e., implausibly high) does not dramatically change the sensitivity of our H1 and H2 tests (we are still able to detect r <.20). Moving on, we further assume 80% power, an alpha threshold of .05, a sample of N=784 (i.e., our post-data exclusion N) and a binomial distribution for the independent variables – because these were dummy-coded [0, 1]. Given these parameters, we were able to detect an Odds Ratio of 1.64. This is equivalent to an r coefficient of .13; conventionally considered small. The GPower (Faul et al., 2017) input/output of this sensitivity analysis is displayed in Figure S1.

Hypotheses 3 and 4 were tested via linear regression. Both H3 and H4 tests were for an interaction (in the presence of two main effects). In our sensitivity analysis, we thus specify three predictor variables in a linear multiple regression model. Assuming 80% power, alpha of .05, and an N=784, we were able to detect an f^2 of .01; equivalent to an r coefficient of .10, conventionally considered small. The GPower output is displayed in Figure S2.

z tests - Logistic regression

Options: Large sample z-Test, Demidenko (2007) with var corr

Analysis: Sensitivity: Compute required effect size

Input:	Tail(s)	=	Two		
	Effect direction		=	p2 >= p1	
	α err prob	=	0.05		
	Pr(Y=1|X=1) H0		=	0.2	
	Power (1-β err prob)		=	0.8	
	Total sample size		=	784	
	R^2 other X	=	0.1		
	X distribution		=	Binomial	
	X parm π		=	0.5	
Output:	Critical z	=	1.9599640		
	Odds ratio	=	1.6407844		
	Actual power		=	0.8000000	

Figure S1. *GPower output of sensitivity analysis for H1 and H2.*

t tests - Linear multiple regression: Fixed model, single regression coefficient

Analysis: Sensitivity: Compute required effect size

Input:	Tail(s)	=	Two	
	α err prob	=	0.05	
	Power (1-β err prob)	=	0.8	
	Total sample size	=	784	
	Number of predictors	=	3	
Output:	Noncentrality parameter δ	=	2.8050371	
	Critical t	=	1.9630100	
	Df	=	780	
	Effect size f^2	=	0.0100360	

Figure S2. *GPower output of sensitivity analysis for H3 and H4.*

Robustness Checks

As shown in Figure 2 in the main text, DG donations are clearly not normally distributed; rather, they follow an approximately bimodal distribution peaking over donations of 0 and 5. We thus conducted several exploratory analyses to test the robustness of the linear regression results reported in the main text (Hypotheses 3 and 4).

Hypothesis 3

Following the bimodal distribution of the DG donations, we computed a binary variable identifying participants who donated less than 5 and participants who donated 5 or above [coded 0 and 1, respectively]. We fitted an exploratory binomial logistic regression model with this new variable as the DV, and TOG frame and TOG choice as the predictor variables. There was an interaction between TOG frame and TOG choice, OR = 3.09 [1.70, 5.61], Z = 3.71, p <.001. A larger proportion of participants donated 5 or above if they made the equitable choice (59.57%) vs. the efficient choice (42.31%) under the equalize frame. This pattern was reversed under the give frame. There, a *smaller* proportion of participants donated 5 or above if they made the equitable choice (42.98%) vs. the efficient choice (53.68%).

We also conducted an exploratory Kruskal-Wallis (K-W) rank sum test with a 4-level factor specifying each combination of TOG choice frame [give, equalize] and TOG choice [efficient, equitable] as a separate group. The DV was DG donation [0-10]. This test showed that the groups differed in DG donations, X^2 (3) = 18.16, p <.001. Follow-up K-W tests showed that, under the equalize frame, participants who made the equitable choice in the TOG donated more in the DG (Median = 5, IQR = 5) than participants who made the efficient choice in the TOG (Median = 1, IQR = 5), X^2 (1) = 13.10, p <.001. In contrast, under the give frame, the reverse was true (equitable choice Median = 1, IQR = 5; efficient choice Median = 5, IQR = 5), X^2 (1) = 5.07, p =.024 (Figure 2, panel B). Both the logistic regression and nonparametric test results are consistent with those of the preregistered linear regression model.

Hypothesis 4

As before, given the approximately bimodal distribution of DG donations (Figure 2, main text), we fitted an exploratory binomial logistic regression as a robustness check on the linear regression model. We specified the binary DG variable computed previously as the DV [0=donated less than 5, 1=donated 5 or above], and moral choice and moral frame as the predictor variables. Consistent with the results of the preregistered linear regression model, there was little evidence of an interaction between moral choice and moral frame, OR = 0.86 [0.48, 1.56], Z = -0.49, p =.627.

Experimental Instructions

Dictator Game

For this task, you will be paired with another person taking this survey.

The amount of money you can earn depends only on your choice. You are given 10c and the other person is given nothing. You have to decide how much, if any, to donate to the other person. The other person has no choice and will accept your donation.

The other person is REAL and will really get your donation. After the survey has ended, your choice will be matched to them to determine each of your bonus earnings.

Here are some questions to make sure that you understand the rules.

Remember that you have to answer all of these questions correctly in order to get the completion code. If you fail any of them, the survey will automatically end and you will not get any payment.

What donation by you means that you and the other person earn the same amount?
(Available answers: 0c/1c/…/10c)

How much should YOU donate to maximize YOUR earnings?
(Available answers: 0c/1c/…/10c)

(Here there was a skip logic which redirected to the end of the survey all subjects who fail either or both the comprehension questions)

Congratulations, you successfully answered all the questions. It is now time to make your decision.

What is your donation?
(Available answers: 0c/1c/…/10c)

Trade-Off game (Give – Do Good frame)

This is the second part of the HIT. Here, you will complete another task.

You are Person A. You are completing this task with two other people taking the survey, Person B and Person C. They are different from the person you were paired with in the previous task.

You get to make a choice. Person B and Person C do not make any choices.

You can either be generous by choosing Option 1, or you can choose Option 2.

If you decide to be generous by choosing Option 1, then you earn 13 cents, Person B earns 23 cents, and Person C earns 13 cents as a bonus.

If you choose Option 2, then you earn 13 cents, Person B earns 13 cents, and Person C earns 13 cents as a bonus.

This is the only interaction you have with Person B and Person C. They will not have the opportunity to influence your earnings in later parts of the HIT.

As with the previous task, the other people are REAL and will really get your donation. After the survey has ended, your choice will be matched to them to determine each of your bonus earnings.

What do you want to do?
(Available answers: Be generous, Option 1/Option 2)

Trade-Off game (Give – Avoid Bad frame)

This is the second part of the HIT. Here, you will complete another task.

You are Person A. You are completing this task with two other people taking the survey, Person B and Person C. They are different from the person you were paired with in the previous task.

You get to make a choice. Person B and Person C do not make any choices.

You can either be ungenerous by choosing Option 1, or you can choose Option 2.

If you decide to be ungenerous by choosing Option 1, then you earn 13 cents, Person B earns 13 cents, and Person C earns 13 cents as a bonus.

If you choose Option 2, then you earn 13 cents, Person B earns 23 cents, and Person C earns 13 cents as a bonus.

This is the only interaction you have with Person B and Person C. They will not have the opportunity to influence your earnings in later parts of the HIT.

As with the previous task, the other people are REAL and will really get your donation. After the survey has ended, your choice will be matched to them to determine each of your bonus earnings.

What do you want to do?
(Available answers: Be ungenerous, Option 1/Option 2)

Trade-Off game (Equalize – Do Good frame)

This is the second part of the HIT. Here, you will complete another task.

You are Person A. You are completing this task with two other people taking the survey, Person B and Person C. They are different from the person you were paired with in the previous task.

You get to make a choice. Person B and Person C do not make any choices.

You can either be fair by choosing Option 1, or you can choose Option 2.

If you decide to be fair by choosing Option 1, then you earn 13 cents, Person B earns 13 cents, and Person C earns 13 cents as a bonus.

If you choose Option 2, then you earn 13 cents, Person B earns 23 cents, and Person C earns 13 cents as a bonus.

This is the only interaction you have with Person B and Person C. They will not have the opportunity to influence your earnings in later parts of the HIT.

As with the previous task, the other people are REAL and will really get your donation. After the survey has ended, your choice will be matched to them to determine each of your bonus earnings.

What do you want to do?
(Available answers: Be fair, Option 1/Option 2)

Trade-Off game (Equalize – Avoid Bad frame)

This is the second part of the HIT. Here, you will complete another task.

You are Person A. You are completing this task with two other people taking the survey, Person B and Person C. They are different from the person you were paired with in the previous task.

You get to make a choice. Person B and Person C do not make any choices.

You can either be unfair by choosing Option 1, or you can choose Option 2.

If you decide to be unfair by choosing Option 1, then you earn 13 cents, Person B earns 23 cents, and Person C earns 13 cents as a bonus.

If you choose Option 2, then you earn 13 cents, Person B earns 13 cents, and Person C earns 13 cents as a bonus.

This is the only interaction you have with Person B and Person C. They will not have the opportunity to influence your earnings in later parts of the HIT.

As with the previous task, the other people are REAL and will really get your donation. After the survey has ended, your choice will be matched to them to determine each of your bonus earnings.

What do you want to do?
(Available answers: Be unfair, Option 1/Option 2)

Part II: Desires, Identities and Bias in Political Belief Formation

Preface

In Part II, I investigate several factors that are purported to influence belief formation in the morally-charged context of contemporary US politics. In particular, I study (i) whether belief updating is biased by the beliefs people already hold or by their desired political outcomes (or both), and (ii) the extent to which cognitive sophistication facilitates biased information processing, such that people who are more sophisticated are more likely to form factual beliefs that are favourable to their political identities.

Specifically, in Part 2.1 I report a study designed to pit two putative "biases" in belief updating against one another. The first bias, which I refer to here as *confirmation bias*, predicts that people incorporate belief-confirming evidence to a greater extent than belief-disconfirming evidence when updating their beliefs (all else being equal). In contrast, the second—which I refer to here as *desirability bias*—predicts that people incorporate desirable evidence to a greater extent than undesirable evidence when updating their beliefs (all else being equal). These biases are often conflated in past work and thus it is unclear which accounts for variance (and how much) in belief updating. To tease the biases apart, I conducted a study capitalizing on the context of the 2016 US presidential election. Individuals were asked who they desired to win and who they believed would win the election. I recruited an equal number of people whose desire and belief were *congruent* (e.g., a desire for Trump and the belief he would win) and whose desire and belief were *in*congruent (e.g., a desire for Trump but the belief Clinton would win). I then randomly assigned information about who was more likely to win—*decoupling* desires and prior beliefs in the aggregate sample of the study—and re-measured people's beliefs. I found evidence to suggest that people updated their beliefs by a greater magnitude if the information was desirable (vs. undesirable). In contrast, I found little evidence for the corresponding asymmetry in updating for information that confirmed (vs. disconfirmed) prior beliefs.

In Part 2.2, I report five studies that investigated the hypothesis that cognitive sophistication facilitates identity-protective bias in political belief formation. The logic of this hypothesis is that people with distinctive cognitive resources are expected to bring those resources to bear

on information that threatens their political identities; specifically, to resist and disregard it. I test this hypothesis in the context of two processes involved in belief formation: (a) *belief updating*—that is, how beliefs change after receipt of evidence—and (b) *reasoning/evidence evaluation*—that is, beliefs about the validity or quality of the evidence itself. I infer cognitive sophistication from performance on the Cognitive Reflection Test (CRT), a behavioural measure of the propensity to think analytically. Regarding (a), I find that—contrary to the target hypothesis—people who score higher on the CRT deviate less from the posterior beliefs of a Bayesian agent; implying *less* bias in belief updating. Regarding (b), I find evidence to suggest that people who score higher on the CRT condition more strongly on their *prior beliefs*—rather than on their political identities *per se*—when evaluating new information. I discuss these findings with respect to existing models of identity-protective belief formation, and I conclude that the evidence taken in favour of the target hypothesis is rather undiagnostic.

2.1. The Heart Trumps the Head: Desirability Bias in Political Belief Revision[18]

Abstract

Understanding how individuals revise their political beliefs has important implications for society. In a preregistered study (N=900) we experimentally separated the predictions of two leading theories of human belief revision—desirability bias and confirmation bias—in the context of the 2016 US presidential election. Participants indicated who they desired to win, and who they believed would win, the election. Following confrontation with evidence that was either consistent or inconsistent with their desires or beliefs, they again indicated who they believed would win. We observed a robust desirability bias—individuals updated their beliefs more if the evidence was consistent (versus inconsistent) with their desired outcome. This bias was independent of whether the evidence was consistent or inconsistent with their prior beliefs. In contrast, we find limited evidence of an independent confirmation bias in belief updating. These results have implications for the relevant psychological theories and for political belief revision in practice.

[18] The work presented in this section was conducted in collaboration with Leslie van der Leer and Ryan McKay (supervisor) and is published in the *Journal of Experimental Psychology: General*: http://psycnet.apa.org/fulltext/2017-23363-001.pdf

Introduction

People are routinely exposed to a bewildering array of information relevant to their political beliefs. Whether and how they incorporate this information has profound consequences for society. The belief that vaccines have harmful side effects (Moritz, 2011), or that climate change is a hoax (Lewandowsky et al., 2013), can reduce people's intentions to vaccinate (Gangarosa et al., 1998; Jolley & Douglas, 2014a; Horne et al., 2015), or to minimize their carbon footprint (Douglas & Sutton, 2015; Jolley & Douglas, 2014b). Even simple infographics displayed during live televised election debates can meaningfully shape beliefs about debate outcome, potentially influencing the voting intentions of millions of viewers (Davis et al., 2011). A clear understanding of how people incorporate information into their political beliefs is thus of considerable practical importance.

Two prominent theories offer similar yet distinct predictions regarding when and how people incorporate new information into their beliefs. One theory contends that individuals assign greater weight to information that is desirable versus undesirable—i.e., a *desirability bias*. This bias is reported to underlie an asymmetry whereby people update their prior beliefs to incorporate new and desirable information more than new but *un*desirable information (Sharot & Garrett, 2016). The other theory, *confirmation bias*, contends that people preferentially search for, evaluate and incorporate new information that confirms their prior beliefs (Nickerson, 1998). This bias is reported to underlie an asymmetry whereby people update their prior beliefs to incorporate new and confirming information more than new but disconfirming information—even if they receive a balanced set of both types of information (Lord et al., 1979; Taber & Lodge, 2006; Taber et al., 2009).

Unfortunately, the predictions of desirability bias and confirmation bias are often conflated. In the domain of self-belief, the tendency for people to believe desirable things about themselves and their futures (Sedikides & Strube, 1997; Weinstein, 1980) means that new information is typically either confirming *and* desirable, or disconfirming *and* undesirable (Eil & Rao, 2011). In the domain of political belief, rigorous separation of desirable and confirming information is similarly difficult. Of the few experiments that *are* appropriately designed to disentangle them, group identity is taken as a proxy for the desirability of information—that is, whether the information is consistent (i.e., desirable) or inconsistent (undesirable) with the position of

an individual's cultural group—and belief updating is not the target outcome measure (e.g., see Kahan, 2016a; 2016b).

Here we experimentally separated desirability bias and confirmation bias in political belief updating. To do so, we capitalized on the political context prior to the 2016 US presidential election. To illustrate the advantage of this context, consider that many supporters of candidate Donald Trump may have believed Hillary Clinton would win the election—owing to her establishment support (Green & Kapur, 2016) or, more conspiratorially, a rigged ballot (Graham, 2016). In such circumstances, new information may have been simultaneously confirming but undesirable—for instance, polls indicating a Clinton win—or disconfirming but desirable—polls indicating a Trump win: causing desirability bias and confirmation bias to yield divergent predictions for belief updating.

We exploited the profusion of close polling results[19] to credibly suggest to individuals that either Donald Trump or Hillary Clinton would become the next president, and measured how individuals with congruent (i.e., same candidate) desire-belief profiles, and incongruent (different candidate) desire-belief profiles updated their beliefs following receipt of this information. We thus independently manipulated whether information was consistent or inconsistent with (a) who individuals *desired* to win the election, or (b) who they *believed* would win the election.

Methods

Participants

We collected data from 900 participants online via Amazon's Mechanical Turk (59% female; M_{age} = 37.89 SD = 12.91). Participants were US residents as determined by IP address (IP addresses located outside of the US were blocked prior to the start of the experiment). We required 779 participants to attain greater than 80% power ($\alpha = .05$) to detect a small effect of partial eta squared (η_p^2) = .01 in our primary analyses of covariance. We added approximately 15% to this number to guard against power loss due to planned data exclusions. Following

[19] At the time of study (data collection commenced 26th September 2016); see http://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton-5491.html

these data exclusions, we retained 811 participants for analyses. The study hypotheses, design, data collection, and analysis plan were pre-registered (see https://aspredicted.org/idxgj.pdf).

Procedure & Design

At the beginning of the survey, participants completed a brief screening questionnaire designed to determine who they (a) *desired* to win, and (b) *believed* would win the 2016 US presidential election. Responses to (a) were provided in a nominal choice format: Participants selected "Donald Trump", "Hillary Clinton", or "neither". Responses to (b) were provided on a bipolar sliding scale from 0-100 with "Hillary Clinton" (0) at one end, and "Donald Trump" (100) at the other (the numerical values were hidden from participants). Participants were instructed that the more confident they were that a candidate would win, the closer they should slide the pointer to that candidate's name. Those who responded with scores greater than 50 were categorized as *believing* Trump would win, and scores less than 50 as *believing* Clinton would win. Participants selecting "neither" for (a), or exactly 50 for (b), were directed to an end-of-survey message and were unable to continue. This yielded two quasi-experimental groups; those whose *desire-believe* candidates were congruent, and those whose *desire-believe* candidates were incongruent. We balanced these condition assignments to obtain approximately 450 in each quasi-experimental condition (final condition samples after data exclusions: congruent desire-belief: n=406 [$\text{desire}_{\text{Trump}}$ / $\text{believe}_{\text{Trump}}$: n=127, $\text{desire}_{\text{Clinton}}$ / $\text{believe}_{\text{Clinton}}$: n=279]; incongruent desire-belief: n=405 [$\text{desire}_{\text{Clinton}}$ / $\text{believe}_{\text{Trump}}$: n=91, $\text{desire}_{\text{Trump}}$ / $\text{believe}_{\text{Clinton}}$: n=314])[20].

Participants in both conditions then completed a filler task (the 16-item Balanced Inventory of Desirable Responding; Hart et al., 2015) before being randomly presented with evidence either consistent, or inconsistent, with who they believed would win the election. Specifically, participants read a short passage about nationwide polling results, which emphasized either that Hillary Clinton or Donald Trump was likely to win the upcoming election. Participants were also presented with a bar graph figure illustrating such an outcome (study materials are available in the Materials Supplement). Evidence presentation was balanced within each specific candidate that participants initially believed would win the election. For example, of

[20] The substantial variance in condition sizes per candidate reflects the fact that approximately three-quarters of our sample initially believed Clinton would win (see Figure 1 in the results section)—explaining the smaller number of individuals in the "$\text{believe}_{\text{Trump}}$" condition(s).

those participants who initially *believed* Trump would win, half received the polling manipulation suggesting Clinton would win, and half received the polling manipulation suggesting Trump would win (likewise for those who initially *believed* Clinton would win). Thus, collapsing over specific candidates, this yielded four between-subjects conditions in a 2 x 2 design: Evidence consistent or inconsistent with who the participant initially *believed* would win (Confirmation: Confirmatory or Disconfirmatory) and consistent or inconsistent with who they *desired* to win (Desirability: Desirable or Undesirable). Following the evidence presentation participants responded to several filler questions about polling data—e.g., "*To what extent have you been following the polling data for the upcoming US presidential election?*"—before again indicating who they believed would win the election, on the same bipolar scale used initially.

Belief Updating

We calculated how much participants updated their confidence in who they believed would win the election in the following steps. First, we converted both the participants' initial confidence (T1), and their subsequent confidence (T2), onto a comparable scale indicating the *absolute* confidence they had in the candidate they initially believed was most likely to win. Thus, for those who initially believed Trump would win we subtracted 50 from T1 and T2 scores, whereas for those who initially believed Clinton would win we subtracted T1 and T2 scores from 50. Next, we computed the absolute difference between these newly converted T1 and T2 scores for each participant. Finally, we multiplied this difference by either 1 (if the participant updated towards the presented evidence) or -1 (if the participant updated away from the presented evidence); meaning that higher numbers represented greater belief updating towards the presented evidence.

Results

Data Exclusions

Participants were excluded from all analyses for fulfilling one or more of the pre-registered criteria: Failing an attention check embedded in the filler task (n=22, 2.44% of sample), answering "yes" to a question asking them if they responded dishonestly or mistakenly during

the survey (n=48, 5.33%), or recording a belief update score of greater than the mean ± 3SD in their respective condition (n=26, 2.89%). We excluded 1 (0.11%) further participant for taking the survey more than once (identified via their unique Amazon Mechanical Turk ID). Following these exclusions, 811 participants were retained for analyses.

Descriptives

Figure 1 displays the proportion of participants reporting who they (a) desired to win and (b) initially believed would win the election (for these results split by gender, age group and ethnicity, see figures S1-S3 in the Analysis Supplement).

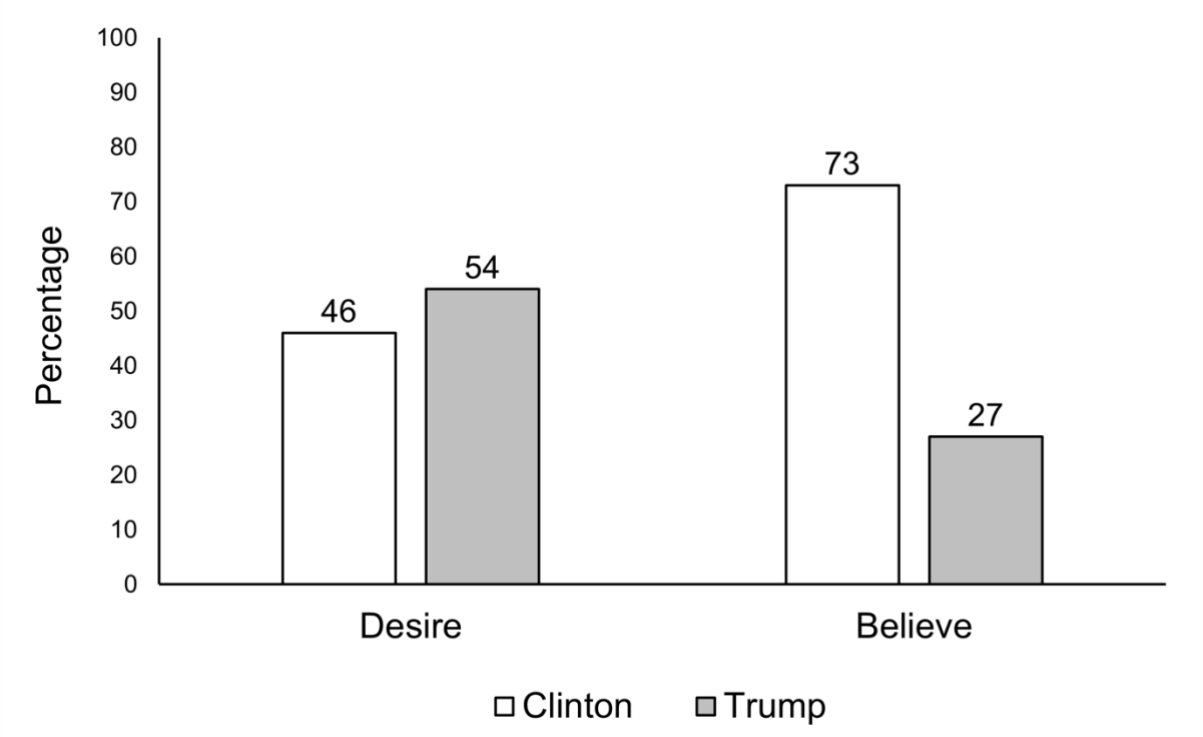
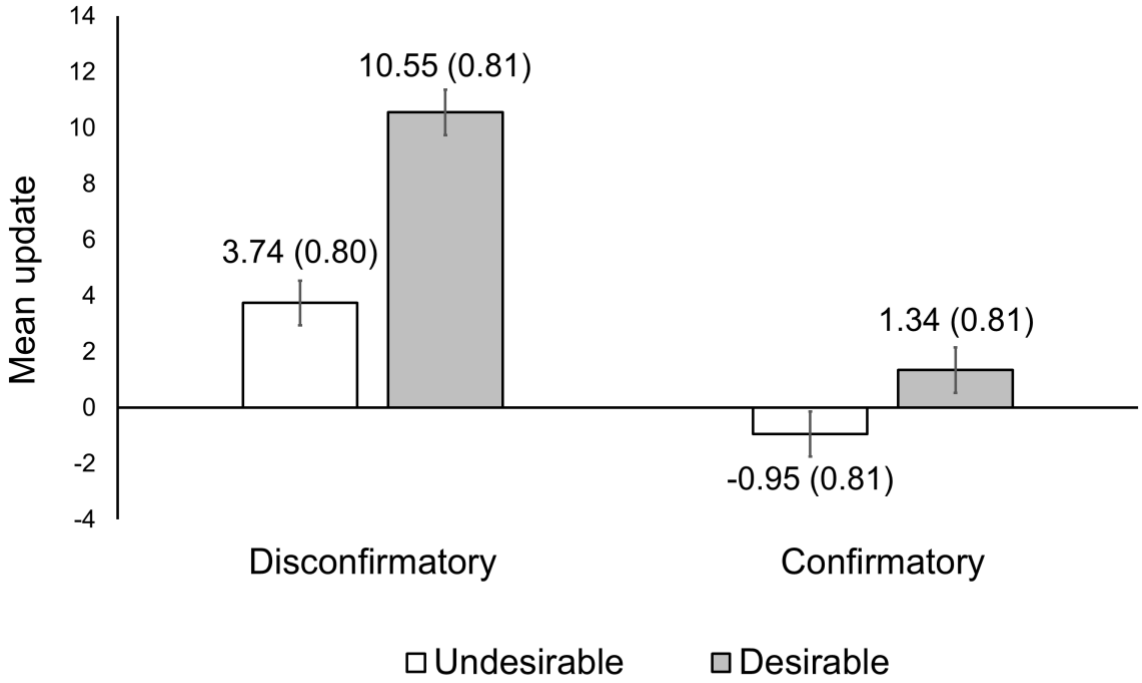


Figure 1. *Percentage of participants reporting which candidate they (a) desired to win and (b) initially believed would win the 2016 US presidential election. N=811.*

Preregistered Analyses

We conducted an analysis of covariance (ANCOVA) to investigate the effect of Desirability and Confirmation factors on belief updating, adjusting for absolute T1 confidence scores[21] (Figure 2 displays the adjusted mean update in each condition[22]). There was a main effect of Desirability: $F(1, 806) = 32.81$, $p < .001$, $\eta_p^2 = .04$ 90% CI [0.02, 0.06], such that participants updated more towards the evidence when it was consistent (versus inconsistent) with the candidate they desired to win. There was also a main effect of Confirmation: $F(1, 806) = 76.63$, $p < .001$, $\eta_p^2 = .09$ [0.06, 0.12], but in this case participants updated more towards the evidence when it was *inconsistent* (versus consistent) with the candidate they initially believed would win. In other words, we observed a *dis*confirmation bias. Finally, we observed a small interaction between Desirability and Confirmation: $F(1, 806) = 7.15$, $p = .008$, $\eta_p^2 = .01$ [0.00, 0.02]. To decompose this interaction, we conducted planned ANCOVAs comparing updating in each condition—while adjusting for absolute T1 confidence scores.



[21] This prevents regression to the mean spuriously affecting belief updating.
[22] The raw means and distributions of update scores are reported in the Analysis Supplement.

Figure 2. *Mean update by condition. Error bars and parentheses denote standard error of the mean. Note: Means are adjusted for absolute T1 confidence and based on the 2x2 ANCOVA model. One unit of update corresponds to a 1% adjustment on the bipolar scale used to measure belief. N=811.*

For those participants receiving disconfirming information, updating was greater if that information was desirable (versus undesirable): $F (1, 407) = 36.58$, $p < .001$, $\eta_p^2 = .08$ 90% CI [0.04, 0.13]. This pattern was the same for those receiving confirming information, albeit less pronounced: $F (1, 398) = 20.62$, $p < .001$, $\eta_p^2 = .05$ [0.02, 0.09]. Next, we examined those participants who received undesirable information—updating was greater for *disconfirming* (versus confirming) information: $F (1, 406) = 23.76$, $p < .001$, $\eta_p^2 = .06$ [0.02, 0.09]. This disconfirmation pattern was the same, yet more pronounced, for those receiving desirable information: $F (1, 399) = 47.72$, $p < .001$, $\eta_p^2 = .11$ [0.06, 0.16]. Finally, directly comparing the unique effect of desirable information (disconfirming-desirable condition) against the unique effect of confirming information (confirming-undesirable condition) revealed that updating was greater for the former: $F (1, 402) = 75.26$, $p < .001$, $\eta_p^2 = .16$ [0.11, 0.21].

In the following sections, we report a series of exploratory analyses to examine (a) the robustness of our results, and (b) extant debates in the field of politically motivated cognition.

Robustness Checks

Prior exposure. It is likely that participants had different amounts of prior exposure to the election polls. Examination of the distribution of one of our filler questions— "*To what extent have you been following the polling data for the upcoming US presidential election?*"—suggested this was the case (see Figure S4 in the Analysis Supplement). It is possible this affected our manipulation and subsequent results. We thus repeated our preregistered ANCOVA with the addition of this variable as a covariate. However, the pattern of results remained the same.

Initial confidence. Participants' initial (T1) confidence scores were negatively skewed—in particular, a substantial number reported complete (or strong) confidence in their initial belief regarding which candidate would win (see Figure 3). This constrains belief updating for those

receiving confirming information because they are unable to update towards the new information (i.e., increase their confidence). In contrast, those receiving disconfirming information *can* update towards the new information (i.e., decrease their confidence). This may account for the disconfirmation bias we observed[23].

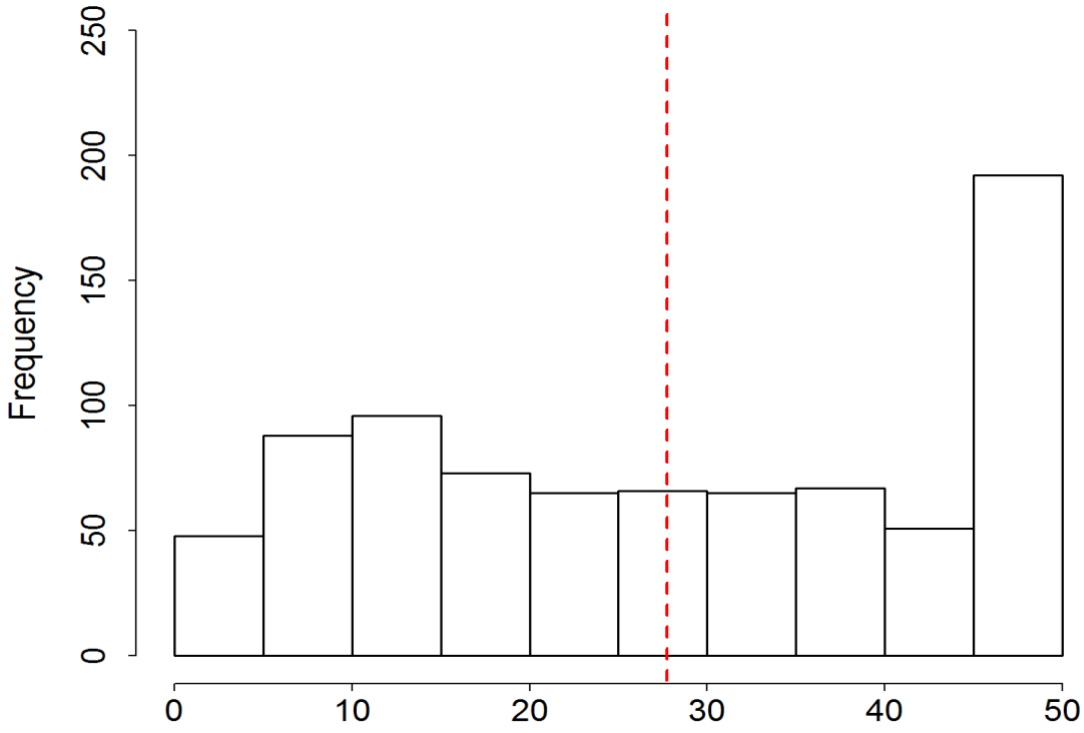


Figure 3. *Distribution of absolute T1 confidence in belief about which candidate would win the election. Note: The dashed line denotes the median. N=811.*

To explore this possibility, we selected a subset of participants (N = 370)—excluding those with high initial confidence (absolute T1 confidence scores > 25, $N_{excluded}$ = 441)—and recomputed the mean update in each condition (Figure 4 displays the results). The pattern of means in this truncated sample indicated a diminished disconfirmation bias, but an enduring

[23] We are grateful to two anonymous reviewers for emphasizing this point.

desirability bias. To confirm this statistically we conducted separate Kruskal-Wallis tests on the distribution of belief updating in the Confirmation and Desirability conditions, respectively[24]. As suspected, there was now only a trivial difference in updating for participants who received disconfirmatory (Median = 2.01, IQR = 11.69) versus confirmatory (Median = 1.78, IQR = 6.58) information: χ^2 (1, N = 370) = 2.01, p =.156. In contrast, participants receiving desirable information updated more (Median = 3.08, IQR = 11.61) than those receiving undesirable information (Median = 0.71, IQR = 6.16): χ^2 (1, N = 370) = 25.84, p <.001.

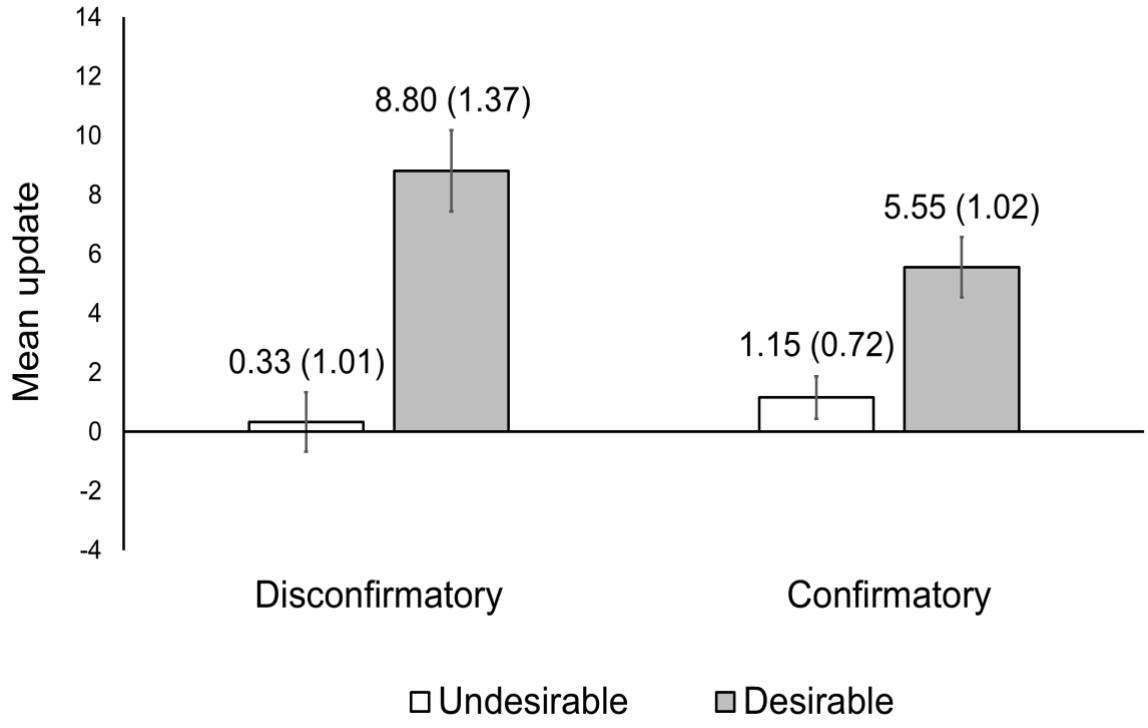


Figure 4. *Mean update by condition following sample truncation. Error bars and parentheses denote standard error of the mean. Note: Means are unadjusted. One unit of update corresponds to a 1% adjustment on the bipolar scale used to measure belief. N=370.*

[24] Parametric analyses were inappropriate as the cell N's across conditions were unequal following sample truncation.

To supplement this analysis, we also specifically examined updating among those with *weak* confidence in their initial belief. This is worthwhile because participants with particularly low confidence may have been (a) less constrained by the upper limit of the confidence scale, or (b) simply more receptive to confirming information, compared to their higher confidence counterparts. Thus, we selected those participants with low confidence (absolute T1 confidence scores $\leq$ 12.5, $N_{excluded}$ = 622) and again recomputed the mean update in each condition. Because the resultant N was small ($N_{low\text{-}confidence}$ = 189) and unevenly distributed across conditions, we simulated belief updating scores using the parameters from the low confidence sample. Specifically, for each of the four conditions, we drew 500 scores from a random normal distribution centered on the respective condition mean, and the pooled SD (i.e., computed across the four conditions) (the simulation script and data are available in the Simulation Supplement).

This simulated sample conferred greater than 99% power to detect small effects (η_p^2 = 0.01, α = 0.05). Conducting an ANOVA on this data revealed a main effect of Desirability, F (1, 1996) = 53.73, p <.001, η_p^2 = .03 90% CI [0.02, 0.04], similar in size and equivalent in direction to that observed in the preceding empirical analyses. The main effect of Confirmation was trivial in size, F (1, 1996) = 2.06, p =.151, η_p^2 = .001 [0.000, 0.005], as was the interaction between the two factors, F (1, 1996) = 1.54, p =.215, η_p^2 = .001 [0.000, 0.004].

Ideological Asymmetry Hypothesis

There is ongoing debate over whether motivated cognition is more pronounced among individuals on the political right than the political left (Jost et al., 2003; Kahan, 2016b). We thus explored whether supporters of Donald Trump demonstrated greater desirability bias than supporters of Hillary Clinton. We conducted an ANCOVA (adjusting for absolute T1 confidence as before) with two factors: Desirability, and a dummy coded variable denoting which candidate the participant desired to win ("Supporter"). There was a small Desirability by Supporter interaction, F (1, 806) = 8.58, p =.004, η_p^2 = .01 90% CI [0.00, 0.03]. Separate ANCOVA models revealed a stronger desirability bias among supporters of Donald Trump, F (1, 438) = 34.07, p <.001, η_p^2 = .07 [0.04, 0.11], than supporters of Hillary Clinton, F (1, 367) = 2.54, p =.112, η_p^2 = .01 [0.00, 0.03].

Further exploration, however, revealed this asymmetry was due to the previously identified ceiling effect in initial (T1) confidence. First, a large number of participants supported Clinton and also believed she would win (n=279)—whereas fewer than half this number supported Trump while also believing he would win (n=127). Second, these participants (i.e., those with *congruent* desire and prior belief) had strong negative skew in their initial confidence, with many believing that their desired candidate was certain to win (see Figure S5 in the Analysis Supplement). Taking these facts together implies that supporters of Clinton were more numerous among those who received desirable information but were constrained (by virtue of their extreme initial confidence) in updating their belief towards this information.

This was confirmed by examining participants who (i) had a congruent desire-belief profile, (ii) received desirable information, and (iii) reported extreme initial confidence (absolute T1 confidence > 45). Of these participants (n=69), 67% supported Clinton (n=46) and 33% supported Trump (n=23). This discrepancy may have disproportionately suppressed desirability bias among Clinton supporters. Indeed, truncating the sample to exclude those with extreme initial confidence (absolute T1 confidence > 45, $N_{excluded}$ = 192, $N_{included}$ = 619), and repeating the ANCOVA analysis, diminished the size of the Desirability by Supporter interaction, $F(1, 614) = 1.08$, $p = .298$, $\eta_p^2 = .002$ 90% CI [0.000, 0.012]. Supporters of Donald Trump and supporters of Hillary Clinton demonstrated similar desirability bias in this sample: $F(1, 348) = 27.19$, $p < .001$, $\eta_p^2 = .07$ [0.03, 0.12] and $F(1, 265) = 10.55$, $p = .001$, $\eta_p^2 = .04$ [0.01, 0.08], respectively.

Discussion

Understanding how people revise their political beliefs has important implications for society. In the context of the 2016 US presidential election, we observed a robust desirability bias: individuals incorporated information more if it was consistent (versus inconsistent) with their desired outcome. This bias was independent of whether the information was consistent or inconsistent with individuals' prior beliefs. In contrast, we found limited evidence of an independent confirmation bias in belief updating. These results have implications for the underlying psychological theories and for political belief revision in practice.

A substantial body of work spanning neuroscience, economics, and clinical psychology reports an asymmetry in the updating of self-beliefs whereby desirable information is incorporated more than undesirable information. This asymmetry has been observed when individuals receive information about their personality traits (Korn et al., 2016; Korn et al., 2012), abilities and attractiveness (Eil & Rao, 2011; Mobius et al., 2011), or risk of experiencing future negative life events (Moutsiana et al., 2013; Sharot et al., 2011; but see Shah et al., 2016; Garrett & Sharot, 2017). A similar yet distinct asymmetry has been reported in the updating of political beliefs whereby individuals become more confident in their prior beliefs despite receiving a balanced set of confirming and disconfirming information. When two individuals with conflicting prior beliefs are thus exposed to the same stream of information, polarization of political beliefs is an oft-observed outcome (e.g., Lord et al., 1979; Taber & Lodge, 2006; Taber et al., 2009).

The present study advances this work twofold. First, we find a robust asymmetry in political belief updating that is consistent with desirability bias, *independent* of individuals' prior beliefs. In contrast, we find little independent effect of prior beliefs on belief updating. This suggests that the belief polarization reported in previous studies may be due to individuals' conflicting desires, not their prior beliefs per se. Second, whereas past investigations of political belief updating have mainly focused on political *attitudes* (e.g., support for or against a policy), here we examined belief updating about political *reality*—specifically, individuals' belief about which presidential candidate was going to be elected. Though one might expect biased belief updating in the former case—after all, attitudes are guided by preferences and desires—it is somewhat more surprising to find that individuals' desires biased their belief updating over a question of fact (Kahan, 2016a).

A recent study reported that individuals updated their beliefs about the facts of global warming asymmetrically, but that the specific pattern depended upon whether they were weak or strong believers in anthropogenic climate change (Sunstein et al., 2016). Particularly, when confronted with new information regarding global temperature increase, strong believers updated their beliefs more upon receipt of ostensibly *undesirable* information (i.e., a faster temperature increase than expected), whereas weak believers updated their beliefs more upon receipt of ostensibly *desirable* information (a slower increase than expected). Though this pattern appears consistent with an independent confirmation bias, such an outcome may emerge when individuals are personally invested in "being right"—indeed, for many climate

change activists a belief that the world is warming constitutes a core part of their identity (Stern et al., 1999). For such people, objectively undesirable (but confirming) information about the rate of global warming may be subjectively *desirable*: vindicating their commitment to combatting climate change (Sunstein et al., 2016) and affirming their cultural group identity (Kahan et al., 2012).

It is unlikely that our design inadvertently conflated confirmation with desirability in this way. Ahead of an election, it is difficult to imagine an individual being personally invested in the belief that their desired candidate would *not* get into office. Indeed, in the domain of self-belief updating, rigorous separation of confirming and desirable information yields identical results to those reported in the present study—namely, a robust desirability bias but limited evidence of confirmation bias (Eil & Rao, 2011). We note the important distinction, however, between (a lack of) confirmation bias observed in *belief updating* as measured here, and confirmation bias observed in measures of *information search* and *evaluation* (e.g., Ditto & Lopez, 1992). We did not directly examine the latter, which may yet manifest independent of information desirability. Additional exploration of our own data lent support to this distinction (see "informational value of polls" available in the Analysis Supplement).

Finally, our results offer a mechanistic explanation for why impassioned political disagreements in the US, such as those over gun control or immigration, appear increasingly polarized and intractable (Pew Research Center, 2016). Insofar as individuals have strong preferences concerning these issues (Koleva et al., 2012), our findings suggest they selectively incorporate new evidence into what they believe to be true regarding the relevant facts—provided it is consistent with what they *desire* to be true. Polarization over factual beliefs is inimical to the effective functioning of democratic society (Kahan et al., 2012); it is thus a priority to continue exploring which interventions ameliorate the motivated integration of evidence (Lewandowsky & Oberauer, 2016).

Supplemental Material

SM: Materials

Screening Questionnaire

Participants completed the following questions as part of the screening procedure (in a fixed order):

1. *Please enter your age (in years):*
2. *Please select your gender:*
    a. *Female*
    b. *Male*
    c. *Other*
3. *Please select your ethnicity:*
    a. *Asian*
    b. *Black*
    c. *Hispanic*
    d. *White*
    e. *Other (please enter):*
4. *Please indicate your religious affiliation:*
    a. *Agnostic*
    b. *Atheist*
    c. *Christian*
    d. *Jewish*
    e. *Muslim*
    f. *Other (please enter):*
5. *Which political candidate do you **want** to win the upcoming US presidential election?*
    a. *Donald Trump*
    b. *Hillary Clinton*
    c. *Neither*
6. *Which political candidate do you **think** will win the upcoming US Presidential election?*

*Please provide your response using the sliding scale below (**note:** dragging the slider closer towards the name of the candidate indicates how confident you are).*

Hillary Clinton |----------◊----------| Donald Trump

Filler Questions

After the screening questionnaire, those who were eligible to continue with the survey completed the *16-item balanced inventory of desirable responding* (BIDR, Hart et al., 2015). Completed on a scale from 1 (Strongly disagree) to 7 (Strongly agree). The question order was fixed.

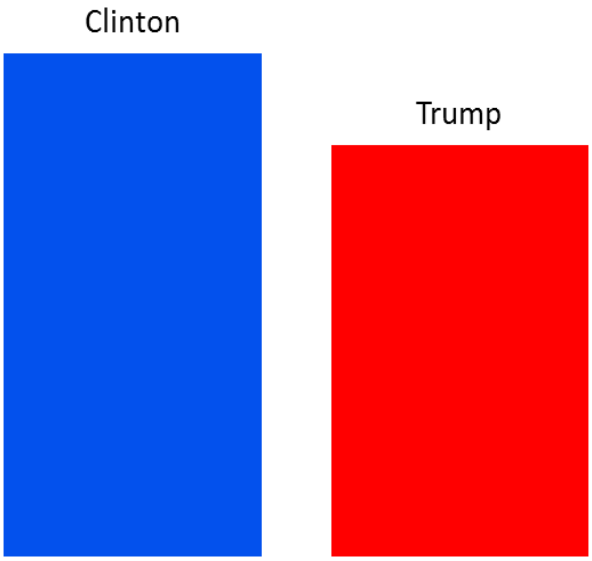
1. *I have not always been honest with myself.*
2. *I always know why I like things.*
3. *It's hard for me to shut off a disturbing thought.*
4. *I never regret my decisions.*
5. *I sometimes lose out on things because I can't make up my mind soon enough.*
6. *I am a completely rational person.*
7. *I am very confident of my judgments.*
8. *I have sometimes doubted my ability as a lover.*
9. *I sometimes tell lies if I have to.*
10. *I never cover up my mistakes.*
11. *There have been occasions when I have taken advantage of someone.*
12. *I sometimes try to get even rather than forgive and forget.*
13. *I have said something bad about a friend behind his/her back.*
14. *Please select response number three.* [**Attention check**]
15. *When I hear people talking privately, I avoid listening.*
16. *I never take things that don't belong to me.*
17. *I don't gossip about other people's business.*

Polling Manipulation

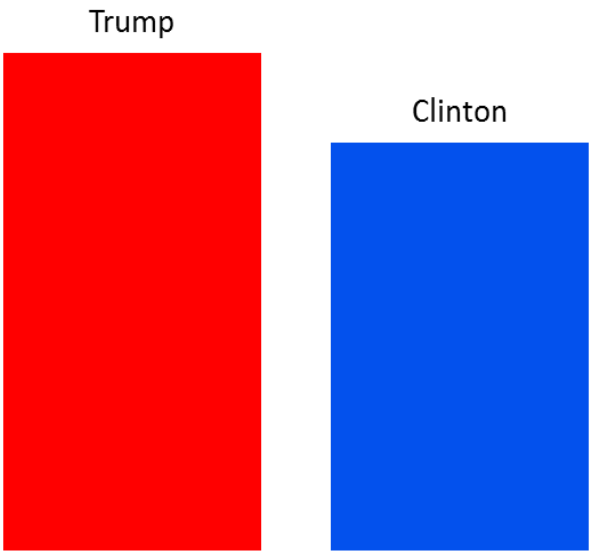
Following the 16-item BIDR, participants (according to condition) read the following passage:

> *Over the past several months there have been many polls conducted to try and predict the outcome of the upcoming US Presidential election: specifically, whether it will be Hillary Clinton or Donald Trump who assumes the mantle of commander-in-chief.*
>
> *These polls are conducted at both the local (i.e., state-wide) and nationwide level. As Election Day draws closer, several of the large nationwide polls become increasingly indicative of who will go on to win the election and become the next President of the United States. Examples of such nationwide polls are the USC/Los Angeles Times and NBC/Survey Monkey tracking polls, the Google Consumer Surveys poll, and the Ipsos/Reuters Core Political Data survey.*
>
> *As you may be aware, data from several of these nationwide polls have recently suggested that, on Election Day, Hillary Clinton* [Donald Trump] *is likely to obtain the largest proportion of votes, and thus be elected as President of the United States. The results of nationwide polling data are not definitive, but results this close to Election Day have proven accurate in correctly identifying the election of Presidential candidates in the past.*

Participants in the Clinton-win condition saw the following graphic:



Participants in the Trump-win condition saw the following graphic:



Polling Data (Filler) Questions

Following the manipulation, participants responded to the following three filler questions (fixed order):

1. *In general, do you think polling data is informative?* [Scored from 1 (Not at all) to 7 (Very much so)]
2. *Do you think there should be more or less polling data made available to the public prior to US presidential elections?* [Scored from 1 (Definitely less) to 7 (Definitely more)]
3. *To what extent have you been following the polling data for the upcoming US presidential election?* [Scored from 1 (Not at all) to 7 (Very much so)]

Time 2 Belief

Participants then once again indicated who they believed would win the election (on the same bipolar scale used previously):

> *Finally, given these nationwide polling data and your own opinion, please indicate which political candidate you **think** will win the upcoming US Presidential election.*
>
> *Please provide your response using the sliding scale below (**note:** dragging the slider closer towards the name of the candidate indicates how confident you are).*

Hillary Clinton |----------◊----------| Donald Trump

Final Questions

On the last page of the survey participants were asked:

> *Accurate data are very important for our research so please answer the following question honestly.*

Your answer to this question is completely anonymous, and we can assure you that there will be no negative consequences whatsoever for answering "yes".

You will get your HIT code on the next screen regardless of how you respond, and your response will have no influence on what Mechanical Turk HITs you can choose to do in the future.

During this survey, did you answer dishonestly or mistakenly at any point?

1. Yes *2. No*

Lastly, we would like to hear any feedback you have about the survey. Please leave any feedback in the box below. If you have no feedback, please enter "none" into the box.

SM: Analyses

Descriptives

Figures S1-S3 (below) show whom participants (a) desired to win and (b) initially believed would win the 2016 US presidential election, by gender (Figure S1), age group (Figure S2) and ethnicity (Figure S3).

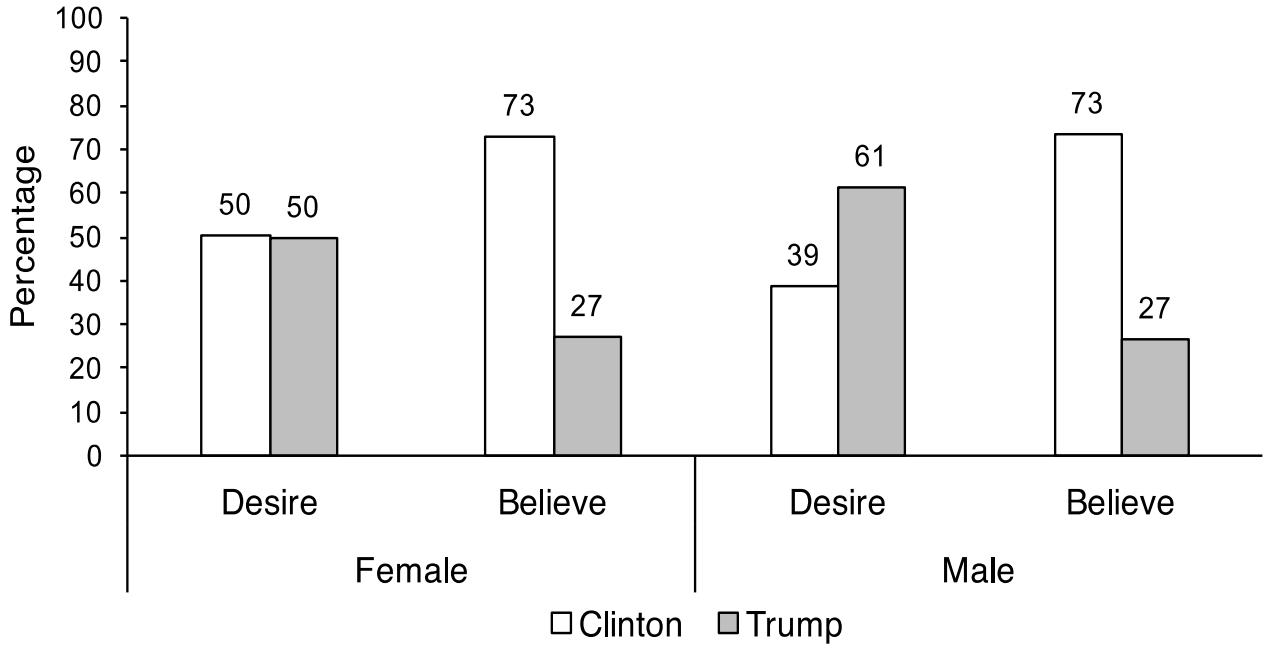


Figure S1. *Percentage of males and females reporting which candidate they (a) desired to win and (b) initially believed would win the 2016 US presidential election. Please note: N=810. One participant did not identify as male or female; they reported desiring Trump but believing Clinton would win.*

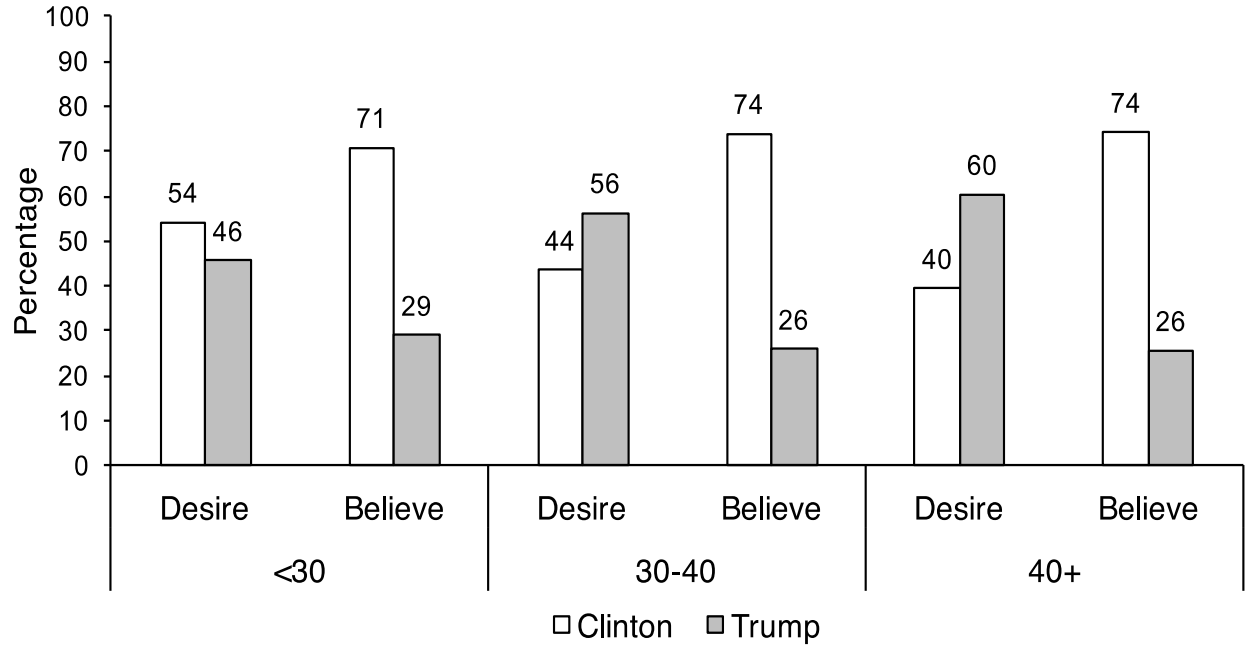


Figure S2. *Percentage of participants in each age group (in years) reporting which candidate they (a) desired to win and (b) initially believed would win the 2016 US presidential election. Note: <30 n=265; 30-40 n=250; 40+ n=296. Total N=811.*

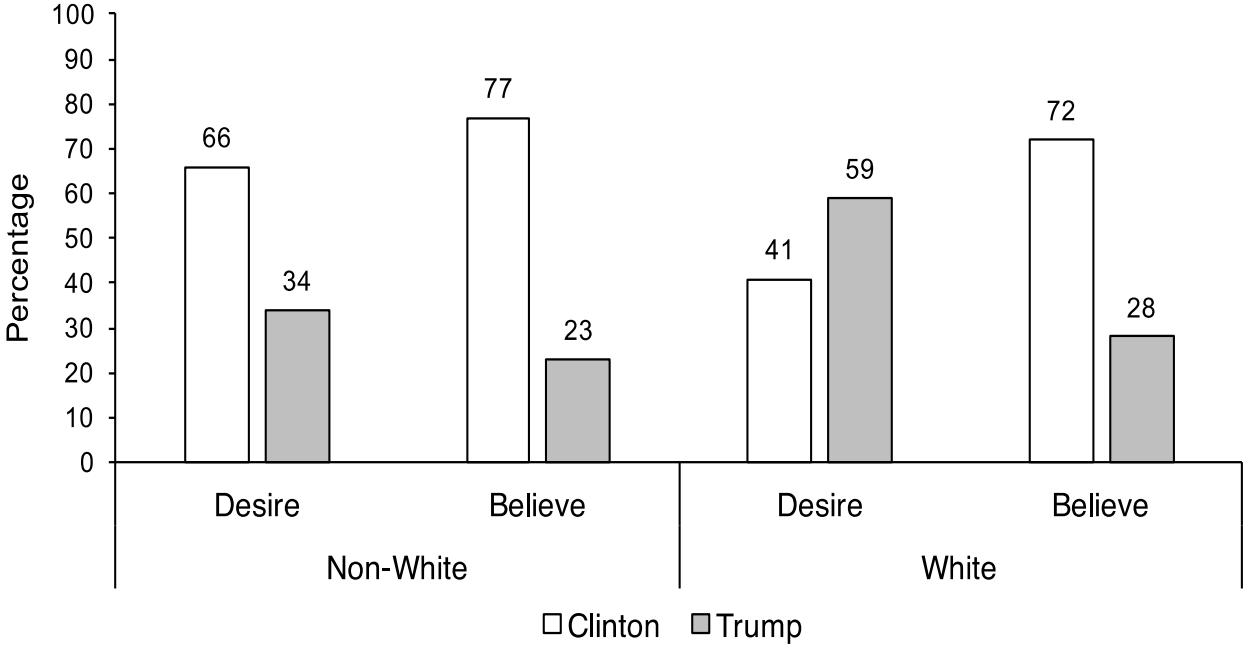


Figure S3. *Percentage of Non-White and White participants reporting which candidate they (a) desired to win and (b) initially believed would win the 2016 US presidential election. Note: Non-White n=152; White n=659. Due to the low number of Non-White participants we pooled them into a single category to provide an interpretable summary.*

Robustness Checks

Prior exposure. Figure S4 (below) shows the distribution of scores on the filler question "*To what extent have you been following the polling data for the upcoming US presidential election?* [Scored from 1 (Not at all) to 7 (Very much so)].

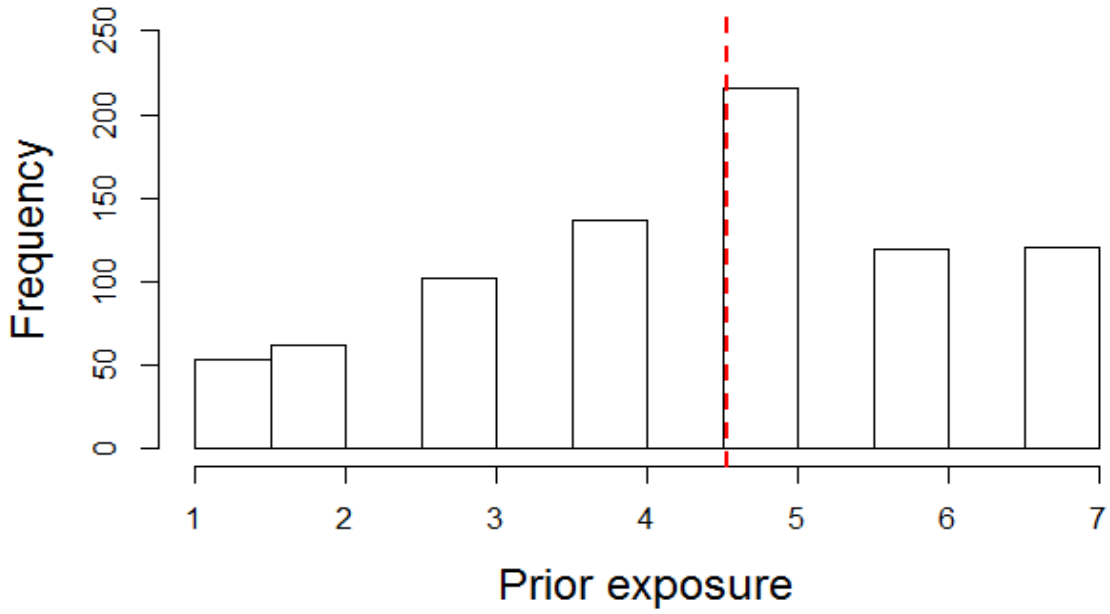


Figure S4. *Distribution of scores indicative of prior exposure to US election polls. Note: Scored from 1 (Not at all) to 7 (Very much so); the dashed line denotes the mean. N=811.*

Ideological Asymmetry Hypothesis

Figure S5 (below) shows the distribution of initial (T1) confidence in the candidate participants believed would win the election, crossed with whom they desired to win the election.

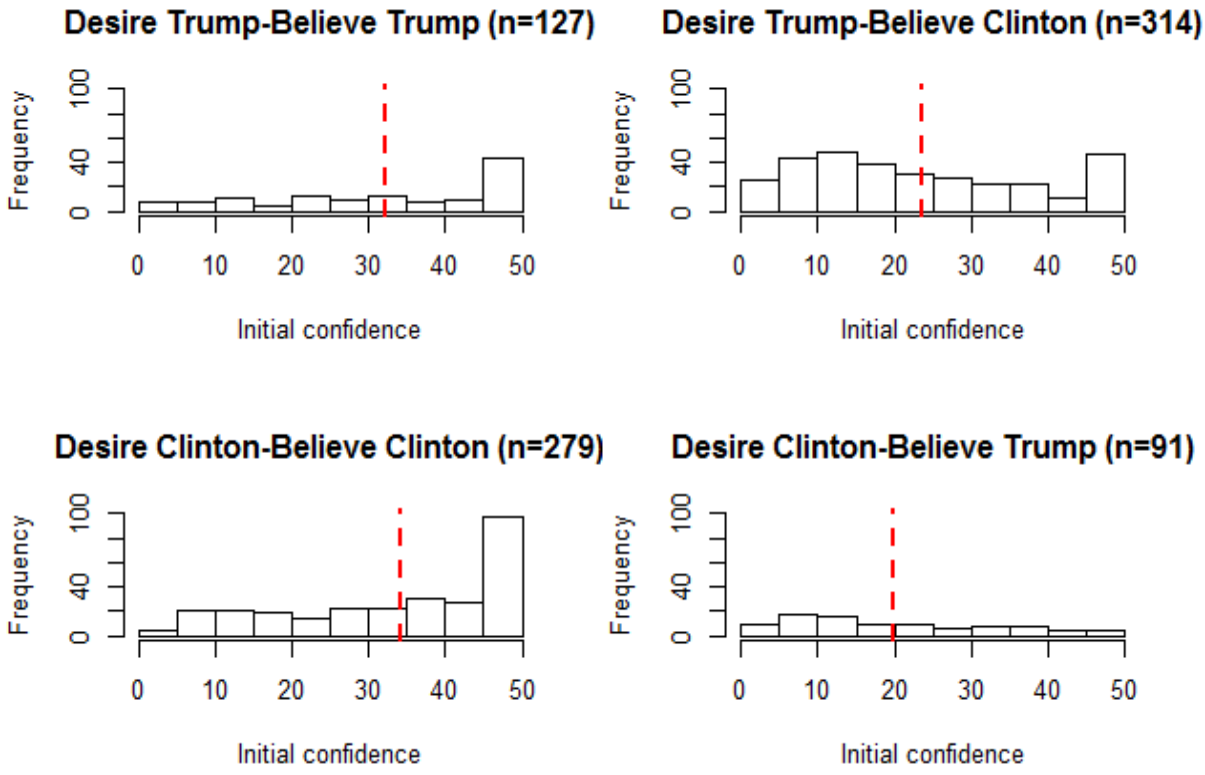


Figure S5. *Distribution of absolute T1 confidence according to which candidate participants (a) desired to win, and (b) initially believed would win. Note: The dashed line denotes the mean. N=811.*

Additional Exploratory Analyses

Informational Value of Polls

We examined the effect of Confirmation and Desirability on participant responses to one of our filler questions— "*In general, do you think polling data is informative?*" [Scored from 1 (Not at all) to 7 (Very much so)]. Conducting a two factor ANOVA revealed a main effect of Desirability, $F(1, 807) = 28.86$, $p < .001$, $\eta_p^2 = .04$ 90% CI [0.02, 0.06], such that participants thought polling data was more informative when it was consistent ($M = 4.95$, $SD = 1.40$) versus inconsistent ($M = 4.41$, $SD = 1.56$) with whom they desired to win the election. We also observed a main effect of Confirmation, $F(1, 807) = 65.86$, $p < .001$, $\eta_p^2 = .08$ [0.05, 0.11], such that participants thought polling data was more informative when it was consistent ($M = 5.08$, $SD = 1.42$) versus inconsistent ($M = 4.27$, $SD = 1.48$) with whom they initially believed would win the election. There was only a trivial interaction between the factors, $F(1, 807) = 0.01$, $p = .933$, $\eta_p^2 < .001$.

We also investigated to what extent supporters of Donald Trump and supporters of Hillary Clinton differed in their judgments of the informational value of polling data overall. An independent samples t-test revealed a trivial difference between supporters of Trump ($M = 4.61$, $SD = 1.58$) and supporters of Clinton ($M = 4.75$, $SD = 1.42$): $t(809) = 1.31$, $p = .192$, $d = 0.09$ 95% CI [-0.05, 0.23].

Participants Who Changed Their Belief

Of the full sample (n=811), 62 (7.64%) changed their (qualitative) belief about which candidate was going to win the election. This was determined by examining the sign of absolute T2 confidence: if participants crossed the midpoint of the bipolar scale when giving their belief for a second time (i.e., indicating they now believed that a different candidate was most likely to win the election) the sign would be negative. As to be expected, participants who changed their belief had lower confidence in their initial belief ($M_{T1\ confidence} = 15.66$, $SD_{T1\ confidence} = 12.42$) than those who did not change their belief ($M_{T1\ confidence} = 29.15$, $SD_{T1\ confidence} = 15.81$). Participants who changed their belief also had much larger belief updating scores ($M_{updating} = 25.10$, $SD_{updating} = 23.02$) than those who did not ($M_{updating} = 1.92$, $SD_{updating} = 8.61$).

Of those who changed their mind, half were supporters of Donald Trump (n=32, 7.26% of all Trump supporters), and half were supporters of Hillary Clinton (n=30, 8.11% of all Clinton supporters). Of those who initially believed Clinton was most likely to win (n=593), 32 (5.40%) subsequently believed Trump was most likely to win; whereas, of those who initially believed

Trump (n=218), 30 (13.76%) changed their belief to Clinton. Finally, receiving the polling manipulation which suggested that Trump was going to win (n=414) prompted 40 (9.66%) individuals to change their belief about which candidate was most likely to win; whereas only 22 (5.54%) of those who received the Clinton manipulation (n=397) did the same. This was probably a function of the fact that approximately three-quarters of the sample initially believed Clinton was more likely to win.

2.2. Rethinking the Link between Cognitive Sophistication and Identity-Protective Bias in Political Belief Formation[25]

Abstract

Popular accounts of political belief formation argue that citizens' cognition is biased toward the formation of factual beliefs that are favorable to their political identities. An alarming hypothesis derived from these accounts is that cognitive sophistication *facilitates* identity-biased processing of political information. In this paper, we re-examine this hypothesis, presenting five studies (total N>5000) that investigated the role of analytic thinking (as measured by the Cognitive Reflection Test, CRT) in political belief formation. Our key results are twofold. In Studies 1 and 2, we investigated *belief updating*, and found no evidence that analytic thinking is associated with biased updating. On the contrary, individuals who scored higher on the CRT tended to deviate *less* from the posterior beliefs of a Bayesian agent, whether new information was concordant or discordant with their political identity. In Studies 2 through 5, we investigated *reasoning* about the validity of politically concordant or discordant information. Our results suggested that highly analytic individuals deferred more to their prior beliefs when reasoning, rather than to their political identities *per se*. An important implication of these results is to highlight the possibility that, rather than being deployed to disregard or resist identity-threatening evidence, cognitive sophistication may instead be deployed to assess and integrate new evidence in light of what the person currently believes to be true. This perspective has implications for theory, provides a reinterpretation of past findings, and may offer greater cause for optimism regarding the role of cognitive sophistication in political belief formation than previously suggested.

[25] The work presented in this section was conducted in collaboration with Gordon Pennycook and David Rand.

Introduction

The ideal of an informed electorate is central to democratic governance. Relevant to promoting this ideal is an understanding of how citizens acquire their beliefs about political issues; and, in particular, their beliefs about politically-relevant *facts*—what is true "out there" in the world—distinct from values, attitudes or opinions. Although this is a longstanding research topic, it has taken on new impetus in recent years as the potential influence of misinformation and disinformation on democratic functioning has piqued scientific, public, and governmental interest (Allcott & Gentzkow, 2017; Digital, Culture, Media, and Sports Committee, 2018; Lazer et al., 2018; Nyhan, 2016), and public consensus over politically-relevant science continues to elude even the more advanced democratic nations (Drummond & Fischhoff, 2017; Kahan, 2015).

According to an influential theoretical perspective, one barrier to public convergence on true beliefs in politics is citizens' *identity commitments*. That is, who they are, and the groups—most commonly, political parties—with which they identify. More specifically, the proposal is that citizens' cognition is *biased* toward the formation of factual beliefs favorable to their political identities. As this bias arises from the application of cognition, an alarming hypothesis that has been advanced is that the most cognitively *sophisticated* partisans will exhibit the *most* identity-bias when processing political information. This hypothesis presents a considerable challenge to the ideal of an informed electorate, and casts an ominous shadow over the prospect of achieving convergence on true beliefs in politics. Chiefly, because it implies that the distinct proficiencies of cognitively sophisticated partisans will be deployed to disregard and resist evidence that threatens their political identities.

In the current paper, we test this hypothesis and offer an alternative hypothesis to explain past evidence. Specifically, we report on a multi-study investigation of how political belief formation relates to the interaction between analytic thinking and political identity. Our main results are twofold. First, benchmarked against a normative Bayesian agent, we find evidence that more analytic individuals are overall *less* biased in their belief updating after receipt of new political information, whether or not it aligns with their political identity. Second, we find evidence that highly analytic individuals may defer more to their prior beliefs – rather than to their political identities *per se* (i.e., independent of their prior beliefs) – when reasoning about

the validity of this information. An important implication of these results is to highlight the possibility that, rather than being deployed to disregard or resist identity-threatening evidence *per se*, cognitive sophistication may instead be deployed to assess and integrate new evidence in light of what the person currently believes to be true. We argue that these results offer a somewhat more optimistic perspective on the role of cognitive sophistication in political belief formation. However, we also highlight that, from a practical perspective, reliance on prior beliefs may be similarly problematic for the prospect of achieving convergence on true beliefs in politics.

1.1. Identity-Protective Cognition in Political Belief Formation

According to several overlapping theoretical accounts, the cognitive processes involved in political belief formation are biased by people's group identities (Bermúdez, 2018; Kahan, 2016a; Van Bavel & Pereira, 2018). While these identities may be diverse in scope, they are often construed along (US) political lines: "Democrat" and "Republican", "liberal" and "conservative" (Kahan, 2016a; Pereira & Van Bavel, 2018; Van Bavel & Pereira, 2018).

The central proposition common to these accounts is that a person's political identity biases their cognition towards formation of factual beliefs that are *concordant* with that identity, and away from factual beliefs that are *discordant* with the identity[26]. For brevity, we will call this biasing influence "identity-protective cognition" (IPC) (cf. Kahan, 2017).

A further similarity between these accounts is their appeal to the logic of utility maximization in belief formation to explain IPC (Bénabou & Tirole, 2016; Loewenstein & Molnar, 2018; Van Bavel & Pereira, 2018). To illustrate this logic, consider that people often depend upon their identity-defining personal and social commitments for psychological and material wellbeing, commitments that may be put at risk by adoption of politically-discordant beliefs (e.g., Crawford & Pilanski, 2014; Huber & Malhotra, 2017). The accuracy of those beliefs—particularly beliefs about political issues which lack a perceptible impact on day-to-day life—is of limited value by comparison (Kahan, 2016a). Thus, the logic underpinning IPC is that

[26] As indicated in the opening paragraphs, the focus of the current paper is *factual* beliefs—that is, people's beliefs about what is true "out there" in the world—rather than values, attitudes or opinions. We use "belief" throughout the paper to refer to factual beliefs.

cognition is sensitive to—and adjusts for—this value asymmetry; manifesting as bias in favor of political identity in belief formation.

A final and related similarity between these theoretical accounts is their construal of *bias* as deviation from accuracy. That is, the bias caused by IPC is considered detrimental to epistemically normative processes of belief formation; because of the prioritization of the identity concordance of political beliefs over their truth value (Bermúdez, 2018; Kahan, 2016a; Van Bavel & Pereira, 2018).

Despite these similarities, there are some differences between the theoretical accounts described. Most notably, they diverge in the specific cognitive processes through which IPC is assumed to exert bias on belief formation. For example, whereas Kahan's (2016a) account assumes the biasing influence of IPC operates when individuals reason about the diagnosticity of new evidence, Van Bavel and Pereira's (2018) account assumes a broader biasing influence of IPC; including on reasoning, memory, perception, and other cognitive processes. While these differences between the accounts are explicit, less explicit is why IPC should be limited to some cognitive processes but not others. After all, both accounts invoke the same logic—utility maximization in belief formation—to explain the biasing influence of IPC.

To illustrate the potential implications of IPC more concretely, Figure 1 displays a triad of distinct cognitive-behavioral processes involved in belief formation; processes through which IPC could conceivably exert bias on the formation of beliefs in the political domain.

The top of the triad—*sampling*—refers to the process of sampling information from the environment. Following the logic of IPC (utility maximization), individuals may be expected to deploy information sampling strategies such that they are more likely to sample information that is concordant (vs. discordant) with their political identity. For example, selecting into politically like-minded media outlets (Bolin & Hamilton, 2018; Newman et al., 2018; Rodriguez et al., 2017), and choosing to hear from politically like-minded others (Marks et al., 2018).

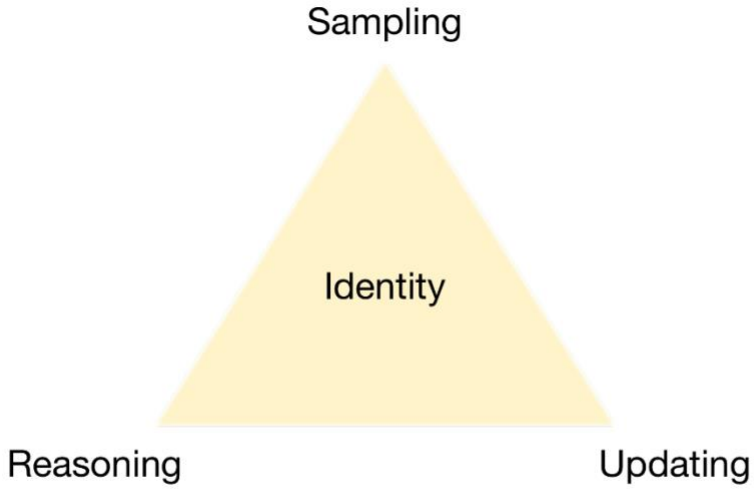


Figure 1. *Three distinct cognitive-behavioral processes involved in political belief formation through which IPC conceivably exerts bias. Sampling refers to information sampling from the environment. Reasoning refers to reasoning (deliberating) about the validity of information. Updating refers to integrating new information with existing beliefs to arrive at a revised (updated) set of beliefs. We do not assume these processes are exhaustive of those involved in belief formation.*

The lower left of the triad—*reasoning*—refers to the process of reasoning or deliberating about the validity of information. Following the logic of IPC, individuals may be expected to reason about information in such a way that politically-concordant information is judged to be more valid or accurate than politically-discordant information. For example, "liberal Democrats" and "conservative Republicans" (in the US) will judge the same information as less or more valid, conditional on its implications for their political identity (Kahan, 2013). Identity-protective reasoning is the process most commonly studied in the context of IPC and political belief formation. A meta-analytic review of such evidence was recently conducted by Ditto and colleagues (2018a; for a critique see Baron & Jost, 2018; for a reply see Ditto et al., 2018b).

The lower right of the triad—*updating*—refers to the process of integrating new information with existing beliefs; providing a revised (updated) set of beliefs. Following the logic of IPC, individuals may be expected to integrate new information with existing beliefs in a biased

fashion; updating to a greater extent after receipt of politically-concordant than politically-discordant information (Sunstein et al., 2017).

1.2. Cognitive Sophistication and Identity-Protective Belief Formation

An alarming hypothesis derived from the logic of IPC is that cognitive *sophistication* will facilitate identity-protective processing in political belief formation (Kahan, 2013, Kahan et al., 2017). By extension, cognitive sophistication may be expected to *increase* bias in this domain; polarizing, rather than unifying, the beliefs of people who identify with opposing political groups. This hypothesis, which we will call the "IPC facilitation hypothesis", casts an ominous shadow over the prospect of achieving convergence on true beliefs in politics; chiefly, because it implies that the distinct proficiencies of cognitively sophisticated partisans will be deployed to disregard or resist new evidence that threatens their political identities.

The IPC facilitation hypothesis is consistent with numerous sets of observational data reported in the last decade. Surveys of US adults repeatedly find that educational attainment is associated with *greater* belief polarization among individuals of opposing political identities, on issues as diverse as climate change (Bolin & Hamilton, 2018; Drummond & Fischhoff, 2017; Buttel & Flinn, 1978; Ehret et al., 2017; Hamilton et al., 2015; Hamilton et al., 2018; Hamilton & Keim, 2009; Hamilton & Stampone, 2013; McCright & Dunlap, 2011; Shao et al., 2014), the safety of vaccination (Hamilton et al., 2015; Joslyn & Sylvester, 2017), concern about the environment (Ehret et al., 2017; Hamilton, 2008; Hamilton & Safford, 2015; Hamilton et al., 2010), the trustworthiness of the scientific community (Gauchat, 2012), and various others (Drummond & Fischhoff, 2017; Hamilton & Saito, 2015; Joslyn & Haider-Markel, 2014; Scheitle, 2018). While educational attainment is a rather crude measure of cognitive sophistication, more targeted measures reveal a similar pattern: High scores on tests of science literacy, numeracy, topic knowledge, and open-minded and reflective thinking are likewise associated with greater belief polarization across various political issues (Bolsen et al., 2015; Drummond & Fischhoff, 2017; Hamilton, 2011; Hamilton et al., 2012; Hart et al., 2015; Kahan et al., 2012; Kahan & Corbin, 2016; Kahan & Stanovich, 2016; Malka et al., 2009; Sarathchandra et al., 2018).

Although these observational data are consistent with the IPC facilitation hypothesis, their purely correlational nature does not provide particularly compelling evidence of IPC. Neither

do they provide insight into the specific processes—for example, sampling, reasoning or updating—through which IPC might exert bias on political belief formation to produce such polarization (Figure 1). To our knowledge, there has been no investigation of whether cognitive sophistication correlates with more biased belief updating, and there is only mixed evidence regarding a link with selective information sampling (e.g., Bolin & Hamilton, 2018; Cragun, 2018; Knobloch-Westerwick et al., 2017).

Stronger evidence for the IPC facilitation hypothesis instead comes from several experimental studies linking cognitive sophistication with identity-protective *reasoning* in political belief formation. In a representative study, Kahan (2013) found that individuals who scored highest on the Cognitive Reflection Test (CRT) – a behavioral measure of the propensity to engage in reflective/analytic thinking (Frederick, 2005; Pennycook et al., 2016) – were *most* polarized in their evaluation of information that was manipulated to be concordant vs. discordant with their political identities (see also Kahan et al., 2017)[27]. These data thus draw a clearer link between cognitive sophistication and identity-protective processing than provided by the aforementioned observational studies[28]; and, at the same time, they offer a mechanism to explain those observational findings. Specifically, that cognitive sophistication—indexed by educational attainment, CRT performance, and so on—confers individuals the *aptitude* to more effectively reason about information such that it "fits" with their political identity (Kahan, 2017).

1.3. Rethinking Cognitive Sophistication and Identity-Protective Processing

Here we re-examine the IPC facilitation hypothesis by further investigating how political belief formation relates to the interaction between cognitive sophistication and political identity. We focus on the processes of (a) *reasoning* about political information, and (b) *belief updating* after receipt of political information. Regarding (a), we propose and test an alternative hypothesis for why cognitively sophisticated individuals putatively defer more to their political identities when reasoning about information in the political domain. Regarding (b), we provide

[27] A similar pattern is reported by Bakker et al. (2018). However, in those experiments, the outcome measure was policy support, not beliefs about "facts" out in the world—as is the focus of the current paper.

[28] But, still, not a *causal* link: Political identity and cognitive sophistication are not randomly assigned in these experiments. This prohibits the inference that political identity (or cognitive sophistication) *causes* biased reasoning, because of the potential confounding influence of unobserved variables (e.g., Gerber & Green, pp. 301-303). Indeed, we posit one such confounding variable—*prior beliefs*—in the following section.

the first test (to our knowledge) of whether cognitively sophisticated individuals are more biased in favor of their political identities when updating their beliefs. In the latter case, we also test the alternative hypothesis that cognitively sophisticated individuals are *less* biased in their belief updating.

Our alternative hypotheses are motivated by multiple and distinct lines of recent evidence. Regarding (a), evidence suggests that cognitively sophisticated individuals defer more to their *prior beliefs* when reasoning about information—in both political *and a*political domains. Because prior beliefs about political issues are often correlated with political identity—perhaps in part due to selective exposure to news media[29] (Bolin & Hamilton, 2018; Rodriguez et al., 2017)—stronger deference to political identity among the cognitively sophisticated may instead reflect a more general and direct deference to prior beliefs. Regarding (b), recent evidence suggests that cognitively sophisticated people have more accurate beliefs about the truth and falsity of political news headlines—independent of the identity concordance of those headlines—a pattern that suggests overall *less* biased belief formation among this group. Below we briefly elaborate on this evidence.

First, a substantial body of research using the "belief bias" paradigm suggests that individuals are influenced by their prior beliefs when reasoning about the deductive validity of arguments (Evans et al., 1983; Klauer et al., 2000; Markovits & Nantel, 1989). In the typical paradigm, individuals are tasked with endorsing (or not) a conclusion that follows from two assumed-to-be-true premises, where the conclusion either does (valid) or does not (invalid) logically follow, and either contradicts or aligns with individuals' prior beliefs (e.g., that whales can walk). Recent work with this paradigm indicates that, while individuals with greater cognitive ability in general—as well as those who score higher on the CRT in particular—perform better overall, they are also more influenced by their prior beliefs (Trippas et al., 2015; Trippas et al., 2018).

To illustrate, consider a recent study by Trippas and colleagues (2015). Among low CRT scorers, the difference in endorsement rates between valid and invalid conclusions was similar irrespective of whether the conclusions contradicted or aligned with subjects' prior beliefs. Specifically, low CRT scorers endorsed valid over invalid conclusions at roughly the same rate

[29] But possibly also due to other factors unrelated to political identity.

across prior belief manipulations. Among high CRT scorers, in contrast, the difference in valid/invalid endorsement rates was *greater* for conclusions that contradicted prior beliefs. In other words, whether a given conclusion contradicted or aligned with subjects' prior beliefs affected task performance in this way *only* for those who scored high on the CRT. Importantly, the content of the task was unrelated to identity or politics, implying that cognitively sophisticated individuals are more sensitive to how information squares with their prior beliefs *per se*; a phenomenon that may also manifest in reasoning about information in the political domain.

Second, several recent studies indicate that people who score higher on the CRT are better able to recognize political *dis*information (specifically, "fake news")—rating it as less accurate—than those who score lower on the CRT (Bronstein et al., 2018; Pennycook & Rand, 2018a, 2018b). Notably, Pennycook and Rand (2018a) found that this association was *independent* of the identity-concordance of the disinformation; that is, better CRT performance was associated with more accurate beliefs about both politically-discordant *and* politically-concordant fake news. This pattern provides some evidence that cognitive sophistication is associated with less—not more—biased belief formation (in the context of fake news headlines, at least; but see Pennycook, Fugelsang, & Koehler, 2015a for a review of similar results in other domains).

Third, and finally, when judging the accuracy of real and fake news headlines, people who scored higher on the CRT were more sensitive to the *plausibility* of the headlines than to their concordance with political identity. In particular, high CRT scorers were more skeptical of headlines that were rated (out of sample) as implausible, compared to those that were more ambiguous. At the same time, however, they were *less* skeptical of headlines that were rated as plausible (Pennycook & Rand, 2018a). This pattern of results suggests that cognitively sophisticated individuals may be more sensitive to how information in the political domain squares with their prior knowledge—to infer plausibility (Mercier, 2017)—and their accuracy ratings defer accordingly[30].

[30] Of course, cognitively sophisticated individuals may also have a richer or more accurate web of prior beliefs which they bring to bear on assessment of plausibility—rather than being more *sensitive* to their prior beliefs per se. Thus, while the results of Pennycook and Rand (2018a) suggest *some* differential influence of prior beliefs among the more (vs. less) cognitively sophisticated, because cognitive sophistication and prior beliefs are not randomly assigned this association is susceptible to unobserved confounding (a point to which we return in the discussion).

1.4. Current Investigation

In four studies (Study 1-4) and an augmented reanalysis of previously published data collected during the 2016 US presidential election (Study 5), we investigated how political belief formation relates to the interaction between cognitive sophistication—as indexed by CRT performance—and political identity. Across these studies, we distinguish between two processes in belief formation: belief *updating* – the integration of new evidence with prior beliefs (Studies 1 and 2) – and *reasoning* – explicit evaluations of the validity or accuracy of new evidence (Studies 2-5).

In particular, in Studies 1 and 2 we adapted a recent learning paradigm to investigate whether individuals who score higher (vs. lower) on the CRT are more biased in favor of their political identities when updating their beliefs. Recall that, according to the IPC perspective, the bias caused by IPC implies deviation from accuracy. In their review of psychological research on bias, Hahn and Harris (2014) argue that charges of bias are rarely backed up by rigorous and systematic evaluation of subjects' behavior against an accuracy criterion. Where use of an accuracy criterion is difficult, Hahn and Harris (2014) recommend that putative bias is evaluated against other normative or optimality criteria; for example, in the case of belief updating, Bayesian rationality. For this reason, in Study 1 and 2 we assess bias in individuals' belief updating against the normative benchmark of a Bayesian agent.

In Study 2, we also examined differences in reasoning about evidence among higher vs. lower CRT scorers. Study 3 and 4 did the same using a different paradigm: We conducted two replications of Kahan's (2013) influential experiment that examined differences in reasoning among higher vs. lower CRT scorers. Critically, however, we also measured and modelled *prior beliefs* alongside political identity. Finally, in the reanalysis of data collected during the 2016 US presidential election (Study 5), we retroactively matched subjects' patterns of reasoning in the data to their CRT performance measured in separate, unrelated research. Furthermore, as a function of the design of Study 5, political identity and prior beliefs were only weakly correlated; allowing us to test the interaction between CRT performance and prior beliefs on reasoning about evidence somewhat independent of political identity.

Across these five studies—which comprised a range of designs, stimuli, analytic approaches, and dependent variables—we found consistent evidence to suggest that individuals who scored higher on the CRT (i) deferred more to their prior beliefs when reasoning about information in the political domain, and (ii) were *less* biased in their belief updating after receipt of such information. In contrast, we found no evidence that these individuals (i) deferred more to their political identities *per se* (i.e., independent of their prior beliefs) when reasoning about such information, or (ii) were more *biased* in favor of their political identities when updating their beliefs.

1.5. Preregistration Protocols and Study Data

We preregistered the hypotheses, sample sizes, designs, and primary analysis plans for Studies 1-4 (Study 1: https://osf.io/e39kq/; Study 2: https://osf.io/9yj57/; Study 3: https://osf.io/j7hrb/; Study 4: https://osf.io/2byaq/). All non-preregistered analyses are designated their own sections—entitled *exploratory analyses*—or, when reported alongside the preregistered analyses, are clearly labelled *post-hoc*. All analyses were conducted in R (v.3.4.0, R Core Team, 2017), using R Studio (v.1.1.423, RStudio Team, 2016). The complete list of R packages (with versions) used in data analysis is reported in this footnote[31]. The data and analysis code to reproduce the results and figures in Studies 1-4 are available via the project hub on the OSF: https://osf.io/yt3kd/. Unfortunately, data from Study 5 is not publicly available due to the wording of the consent form signed by subjects.

Study 1

In Study 1, we adapted a recent learning paradigm (Hill, 2017) to compare subjects' belief updating in the political domain to that of a normative Bayesian agent. We tested two hypotheses. The first is that individuals who score higher on the CRT deviate *less* from Bayesian updating overall; combining their prior beliefs with new political information in a

[31] Complete list of R packages used in data analysis: *Data.table* (v.1.10.4-3, Dowle & Srinivasan, 2017); *reshape* (v.0.8.7, Wickham, 2007); *plyr* (v.1.8.4, Wickham, 2011); *dplyr* (v.0.7.6, Wickham et al., 2018); *lme4* (v.1.1-15, Bates et al., 2015); *ggplot2* (v.3.0.0, Wickham, 2016); *gridExtra* (v.2.3, Auguie, 2017); *effects* (v.4.0-0, Fox, 2003); *ppcor* (v.1.1, Kim, 2015); *psych* (v.1.8.4, Revelle, 2018); *sjPlot* (v.2.5.0, Lüdecke, 2018); *MASS* (v.7.3-47, Venables & Ripley, 2002); *lmerTest* (v.3.0-0, Kuznetsova et al., 2017). We are grateful to the authors of these packages.

less biased (more Bayesian) manner (H1). The second, consistent with the logic of the IPC facilitation hypothesis, is that individuals who score higher on the CRT deviate from Bayesian updating in a pattern consistent with identity-protective processing (H2). Specifically, their deviation from Bayesian updating is *conditional* on the political concordance of new information, and they are more biased in favor of their political identities relative to low CRT scorers.

2.1. Methods

In this and subsequent studies, for brevity, we report methodological detail sufficient to orient readers but refer to the supplementary materials (SM) for minutiae (also available on the OSF: https://osf.io/yt3kd/).

2.1.1. Sample

We sought to collect N = 500 subjects, on the assumption of a small association between CRT performance and deviation from Bayesian updating (*a priori* power analysis reported in the preregistered protocol). A total of N = 501 subjects completed the study. In Studies 1-3, and Study 5, subjects were from the US, recruited online via Amazon's Mechanical Turk (MTurk). Subjects in Study 4 were recruited from Lucid, a marketplace for online survey research whose samples are more representative of the US general population than MTurk (Coppock & McClellan, 2017).

While MTurk is an intensely studied convenience sample—presenting valid concerns about the non-naiveté of subjects and generalizability of results—recent work suggests that cognitive psychological phenomena are generally reproducible on MTurk, *despite* non-naiveté (Bialek & Pennycook, 2017; Stagnaro et al., 2018; Zwaan et al., 2017). Furthermore, MTurk is valid for research on political identity in particular—insofar as political partisans recruited via MTurk have similar psychological profiles to partisans in nationally representative samples (Clifford et al., 2015). Finally, past work administering the CRT to representative samples finds over a third of individuals score *zero*; limiting the discriminability of the test at low-end performance (Kahan, 2013; also see Bialek & Pennycook, 2017). CRT scores in our MTurk samples, by contrast, are considerably less skewed (in Study 1, 2 and 3, see SM for the distributions); and, consequently, better discriminate among low-end performers. While

MTurk thus possesses certain limitations as a sampling population, existing evidence suggests that it can provide for valid inferences in the case of the current investigation.

2.1.2. Belief Update Task

The task consisted of two phases. In phase one (P1), subjects made sequential likelihood judgments about the truth of various political statements; their *prior beliefs*. Subjects were informed before P1 began that the statements were either true or false. Likelihood judgments were provided in percentages on a sliding scale from 0 ("certainly false") to 100 ("certainly true"), in whole integers. There were 16 political statements, corresponding to 16 trials in P1. Of these 16 statements, 8 favored the Democratic Party if they were true (pro-Democratic); 4 of which were (in fact) true, and 4 of which were false. The remaining 8 statements favored the Republican Party if true (pro-Republican); again, 4 were in fact true, while 4 were false. The political statements were selected via a three-step pre-testing procedure (reported in full in the SM). Four example statements are displayed in Table 1.

Table 1. Four Example Political Statements Used in Studies 1 and 2.

	Pro-Democratic	**Pro-Republican**
True	During former President Barack Obama's final 4 years in office, wages of the average American worker went up.	Under President Donald Trump's administration, unemployment has fallen to a 17-year low.
False	Within two years of "Obamacare" being signed into law, health care premiums were going up more slowly than at any time in the previous 50 years.	Only 10 cents on every dollar from the Clinton Foundation goes to charitable causes.

Immediately after rating a given political statement in P1, subjects received a signal about whether that statement was in fact true (i.e., "TRUE") or false (i.e., "FALSE"). Signals were correct with probability 2/3. Signals thus provided noisy but, on average, *accurate* evidence about the truth (or falsity) of the political statements. Importantly, subjects were informed of the probability of receiving an accurate signal and answered comprehension questions prior to

P1 to ensure their understanding. Verbatim task instructions are available on the OSF: https://osf.io/yt3kd/. Subjects were reminded of the probability of receiving an accurate signal after each signal they received in P1.

After rating and receiving signals for each of the 16 statements in P1, subjects moved onto phase two (P2). In P2, subjects saw each political statement presented in P1 again (sequentially, over 16 trials), and made another likelihood judgment about the truth of each statement; their *posterior beliefs*. Subjects were not reminded of their P1 likelihood judgment or the signal they received. Upon completion of P2, the task was over. The presentation order of political statement stimuli in P1 and P2 was randomized. Prior to starting P1, all subjects completed a practice P1 and P2 trial with an apolitical statement: "Henry VII was King of England between the years 1485 and 1509" (not included in analysis). The structure of the task is displayed in Figure 2.

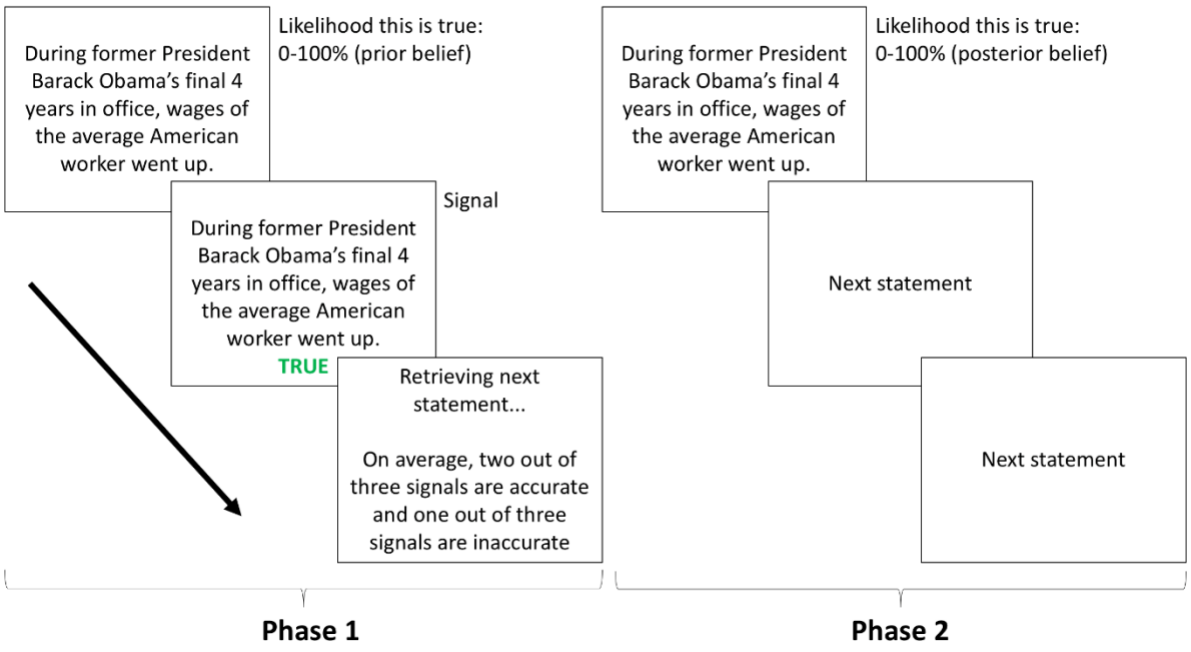


Figure 2. *Task structure in Study 1.*

2.1.3. Comparison with Bayesian Agent

To compare the posterior beliefs of our subjects—obtained in P2—to those of a (simple) normative Bayesian agent, we first consulted the prior beliefs provided by subjects in P1 and the signal they received for each statement during the belief update task. With this information, we computed the posterior belief of a Bayesian agent for each statement using Bayes' rule,

$$P(T|S) = \frac{P(T)P(S|T)}{P(T)P(S|T) + P(\neg T)P(S|\neg T)}$$

where $P(T|S)$ is the posterior probability the statement is true, given the signal received on that trial (i.e., the "Bayesian posterior belief"); $P(T)$ is the prior probability that the statement is true (i.e., the subject's prior belief); and, $P(S|T)$ is the probability of receiving the signal assuming the statement is true (i.e., equal to 2/3 for TRUE signals, and 1/3 for FALSE signals). The terms $P(\neg T)$ and $P(S|\neg T)$ refer to the prior probability that the statement is false ($1 - P(T)$), and to the probability of receiving the signal assuming the statement is false ($1 - P(S|T)$)[32], respectively. Using this equation, we computed Bayesian posterior beliefs for each statement rated by each subject. Note that prior and posterior beliefs are on the 0-1 scale for these computations, but are converted back to 0-100 scale for analysis.

We computed two dependent variables (DVs) to quantify subjects' deviation from Bayesian posterior beliefs: (i) the *difference* between subjects' posterior beliefs and Bayesian posterior beliefs (difference index), and (ii) the *ratio* of subjects' posterior beliefs to Bayesian posterior beliefs (ratio index). To illustrate the DVs, assume that a subject receives a signal of "TRUE" regarding statement X. They subsequently report a posterior belief of 72% that statement X is true. The Bayesian posterior belief on this trial, however, is calculated to be 65%. Thus, the raw difference between the subject and Bayesian posterior belief is 72 – 65 = 7 (difference index). The ratio of the two posterior beliefs is 72/65 ≈ 1.11 (ratio index). For all analyses, we log-transform the ratio index due to positive skew (this transformation was preregistered). These two indices react in different ways to the bounded nature (0-100%) of the likelihood judgment scale; thus, analyzing both provides for more valid inferences (Shah et al., 2016). The two DVs are computed such that, relative to Bayesian updating, values *smaller* than zero

[32] We note that is it not a mathematical *necessity* that P(S|-T) is equal to 1 – P(S|T), but this is a reasonable assumption in the this context given the symmetric wording of "accuracy" in the instructions provided to subjects.

imply *under*-updating, and values *greater* than zero imply *over*-updating. Values of exactly zero imply the (normatively correct) updating of a Bayesian agent.

2.1.4. Post-Task Measures

Following the belief update task, subjects completed a memory test (for exploratory purposes). Specifically, they were shown each political statement again and were asked to indicate whether they saw a TRUE or FALSE signal for that statement. Subjects then completed a 7-item CRT; comprised of a reworded version of the original 3-item Frederick (2005) CRT (from Shenhav, Rand, & Greene, 2012), and the 4-item non-numeric CRT from Thomson and Oppenheimer (2016). The CRT is a behavioral task that measures the propensity to engage in reflective/analytic thinking, and to override intuitive but incorrect responses (Frederick, 2005; Pennycook et al., 2016). Test-retest reliability estimates for the original 3-item measure range from .75 to .81 (Stagnaro et al., 2018), and CRT performance shares a substantial positive correlation with cognitive ability and aptitude over a range of rational thinking measures (Toplak et al., 2011) and predicts a number of everyday beliefs and behaviors (Pennycook, Fugelsang, & Koehler, 2015a). Correct responses were summed to create a 0-7 score for each subject ($M = 3.85$, $SD = 2.15$, $\alpha = .79$, 95% CI [.77, .82], average inter-item $r = .35$) (items and distributions of sum scores are reported in the SM). Finally, subjects completed a self-report scale measuring their responsiveness to evidence (for exploratory purposes), and provided demographic information including which US political party they preferred; Democratic or Republican (forced choice).

2.1.5. Political Concordance of Evidence

For analyses, we computed a variable indexing whether the signal (i.e., "evidence") subjects received on each trial was concordant or discordant with their political identity. This variable was a joint function of subjects' political party preference (Democratic or Republican), the partisanship of the political statement stimulus (i.e., the statement favors Democrats if it is true vs. favors Republicans if it is true), and the signal received (signal: "TRUE" or "FALSE"). The classifications are displayed in Table 2.

Table 2. Computing the Political Concordance of Evidence.

Subject party preference	Partisanship of political statement	Signal received	Political concordance
Democratic	Democrat-favor	True	Concordant
Democratic	Democrat-favor	False	Discordant
Democratic	Republican-favor	True	Discordant
Democratic	Republican-favor	False	Concordant
Republican	Democrat-favor	True	Discordant
Republican	Democrat-favor	False	Concordant
Republican	Republican-favor	True	Concordant
Republican	Republican-favor	False	Discordant

2.2. Results

2.2.1. Data Exclusions

For all hypothesis tests, we exclude subjects with duplicate IP addresses (N = 9, 1.80%); retaining the earliest responses only. We also exclude trials on which subjects' prior belief was provided as 0 or 100, and the signal they received was FALSE or TRUE, respectively; because updating is not possible on these trials (N trials = 254, 3.23%). Lastly, missing trial data (prior and posterior beliefs) are excluded from both H1 and H2 tests (N = 0), as are subjects who did not report a US political party preference (preregistered test of H2 only, N = 2, 0.41%). These exclusion criteria were preregistered prior to data collection.

2.2.2. H1: High CRT Scorers Deviate Less from Bayesian Updating

After data exclusions, we retained N = 492 for the preregistered test of H1. To test H1—that individuals who score higher on the CRT deviate less from Bayesian updating overall—we computed *absolute* values of both DVs (|difference index|, |ratio index|). The absolute values thus represent deviation from Bayesian posterior beliefs collapsing over the particular *direction* of that deviation (since direction is the focus of H2). We then computed the mean absolute deviation from Bayesian posterior beliefs over the 16 trials for each subject.

2.2.2.1. Preregistered Tests

Due to anticipated skew in the absolute DVs, we preregistered nonparametric Kendall's tau (τ) correlations between the DVs and CRT sum scores. Consistent with H1, CRT performance was negatively correlated with absolute deviation from Bayesian posterior beliefs, though the association was larger for the difference index: $\tau = -.20$, $Z = -6.27$, $p < .001$, than for the ratio index: $\tau = -.06$, $Z = -1.72$, $p = .086$. The raw data are displayed in Figure 3 (A).

The magnitude of the negative correlations was similar across politically concordant and politically discordant evidence: Politically discordant: $\tau = -.18$, $Z = -5.66$, $p < .001$ (difference index), $\tau = -.04$, $Z = -1.24$, $p = .215$ (ratio index), politically concordant: $\tau = -.17$, $Z = -5.42$, $p < .001$ (difference index), $\tau = -.06$, $Z = -1.99$, $p = .046$ (ratio index). (Note: These correlation tests—split by the political concordance of the evidence—were conducted *post-hoc*).

Broadly consistent with H1, subjects who scored higher on the CRT deviated less from Bayesian posterior beliefs overall, combining their prior beliefs with new politically-relevant evidence in a more normative (less biased) manner.

2.2.2.2. Exploratory Tests

As a check on the robustness of H1 results, we conducted a range of exploratory analyses. These analyses are reported in full in the SM, and, on the aggregate, show that the preregistered result is robust to a variety of analytic specifications and potential confounds. Here we report one particularly noteworthy such analysis regarding memory errors.

The memory test following the belief update task indicated that subjects scoring higher on the CRT had better memory for the signals they received (i.e., they made fewer errors in recall, see SM for analysis). It is possible this difference—rather than differences in updating *per se*—accounted for the relationship between CRT performance and deviation from Bayesian posterior beliefs. To examine this possibility, we conducted nonparametric partial correlations between (a) CRT scores and (b) mean absolute deviation from Bayesian posterior beliefs, adjusting for (c) subjects' proportion of memory errors (see SM for further details). The results closely reproduced the preregistered H1 results: $\tau = -.20$, $Z = -6.58$, $p < .001$ (difference index),

τ = -.07, Z = -2.42, p =.015 (ratio index). In fact, the coefficient on the ratio index slightly increased, implying that differences in memory were slightly suppressing the relationship between CRT and Bayesian updating in the preregistered test of H1.

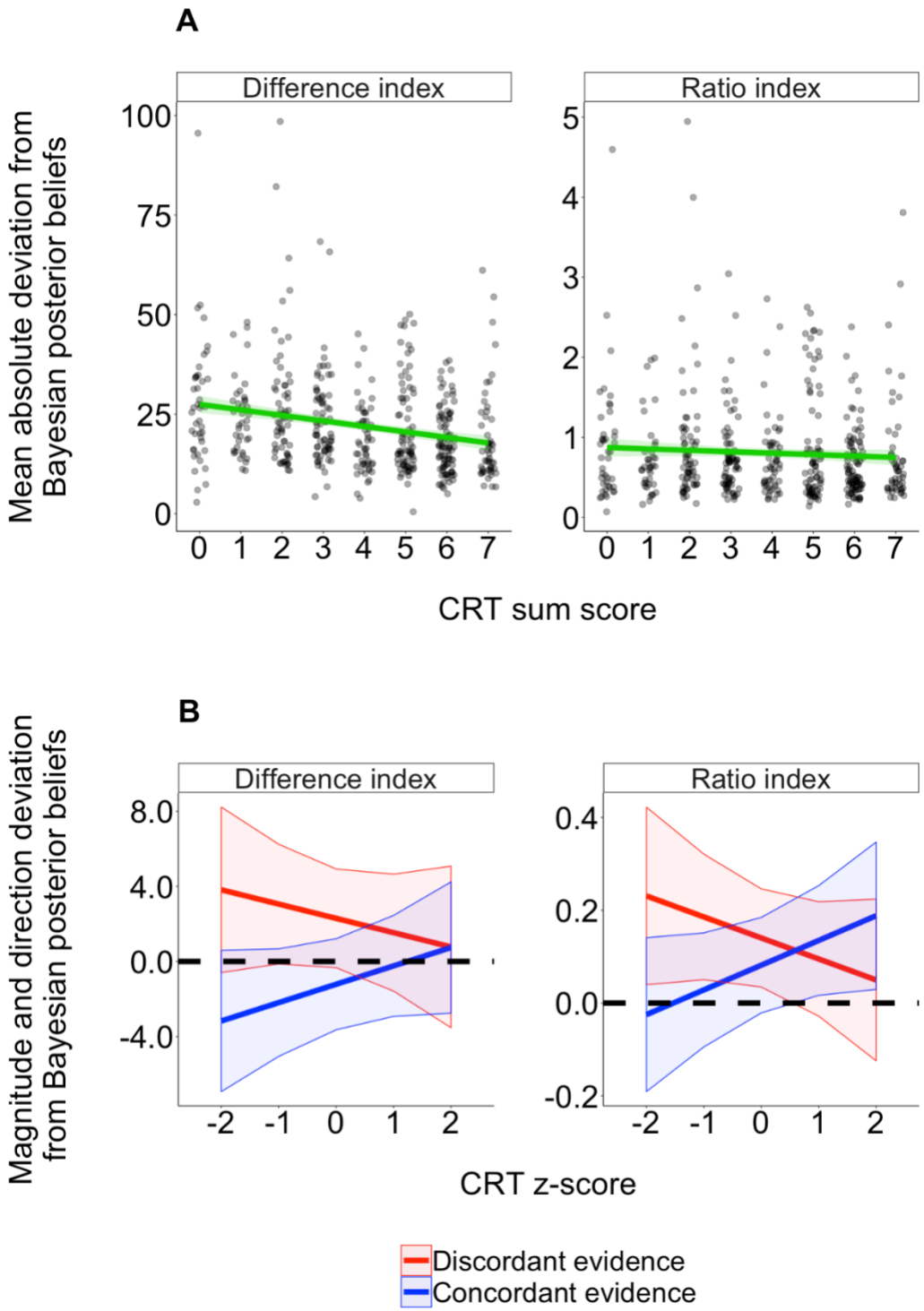


Figure 3. Deviation from Bayesian posterior beliefs as a function of CRT performance and the political concordance of evidence in Study 1. A*, mean absolute deviation from Bayesian posterior beliefs (represented at y = 0). Each point corresponds to the mean absolute deviation of one subject. Data points have slight jitter for visibility.* ***B****, predicted magnitude and direction deviation from Bayesian posterior beliefs (represented at y = 0). Predicted values are obtained from the models reported in Table 3. Shaded regions are 95% CI. CRT = Cognitive Reflection Test.*

2.2.3. H2: High CRT Scorers Deviate from Bayesian Updating Conditional on their Political Identities

After data exclusions, we retained N = 490 for the preregistered test of H2. H2 is that individuals who score higher on the CRT deviate from Bayesian posterior beliefs conditional on the political identity concordance of the evidence (to a greater extent than individuals who score lower on the CRT). Furthermore, H2 concerns the direction of deviation from Bayesian posterior beliefs; that is, whether individuals who score high on the CRT (i) deviate from Bayesian posterior beliefs conditional on the political concordance of the evidence (to a greater extent than low CRT scorers), and, specifically, (ii) deviate in the direction that is more favorable to their political identities.

To assess both magnitude and direction of deviation conditional on political concordance, we use the two DVs as initially computed; that is, before their absolute transformations for the test of H1. Relative to the Bayesian benchmark, therefore, values *greater* than zero imply *over*-updating; values *smaller* than zero imply *under*-updating; and values of exactly zero imply the normatively correct updating of a Bayesian agent. The critical test of H2 is on the interaction between CRT performance and the political concordance of the evidence in predicting these values.

2.2.3.1. Preregistered Tests

We fitted two linear mixed effects models at the trial-level; one for each DV (difference index, ratio index). Linear mixed effects models possess several advantages over classic ANOVA (see

e.g., Barr et al., 2013; Brauer & Curtin, 2017; Judd et al., 2012). Most notably, they can model non-independent data while maintaining the nominal (5%) type I error rate, and allow researchers to generalize beyond the stimuli used in their particular experiment. Our model fitting procedure is detailed in the SM. In the text below, we report *Likelihood Ratio Tests* (LRT) statistically evaluating the contribution of the key fixed-effect interaction term: CRT performance x political concordance (of the evidence). We first attempt to fit a "maximal" random effects structure[33] before conducting the LRT, because this was our preregistered critical test. Where the maximal model does not converge, we simplify the model by iteratively dropping random effects terms and re-estimating the model at each iteration. This protocol was preregistered. In Table 3, we report the parameter estimates from the final models described in text below.

The maximal model fitted on the difference index DV failed to converge. To achieve convergence, we re-estimated the model after dropping the random slope of [CRT x political concordance] on stimuli. The critical LRT on the fixed-effect interaction term in this model was (just) statistically significant, χ^2 (1) = 3.94, p =.047. That is, the interaction between CRT performance and the political concordance of evidence improved model fit. Predicted values from this model are displayed in Figure 3 (left panel B). We thus proceeded to estimate the simple slopes of CRT performance over trials where the evidence was (a) politically concordant, and (b) politically discordant, by fitting separate models for each. Higher CRT performance was associated with relatively greater over-updating on trials where the evidence was politically concordant, b = 1.03 (SE = 0.67), but this missed the significance threshold, χ^2 (1) = 2.34, p =.126. This pattern was reversed on trials where the evidence was politically *dis*cordant. There, higher CRT performance was associated with relatively greater *under*-updating, b = -0.74 (SE = 0.92), but again this slope was not statistically significant itself, χ^2 (1) = 0.65, p =.421.

In the ratio index maximal model, the critical LRT was statistically significant, χ^2 (1) = 6.04, p =.014. That is, the interaction between CRT performance and the political concordance of evidence improved model fit. Predicted values from the maximal model are displayed in Figure 3 (right panel B). Once again, we proceeded to estimate the simple slopes of CRT performance

[33] This means estimating a full random effects structure; providing a conservative test of the contribution of the fixed effects (for more discussion we refer to Barr et al., 2013).

by fitting separate models for each. Higher CRT performance was again associated with relatively greater over-updating on trials where the evidence was politically concordant, b = 0.05 (SE = 0.03), but this missed the significance threshold, χ^2 (1) = 3.01, p =.083. This pattern was reversed on trials where the evidence was politically discordant. There, higher CRT performance was associated with relatively greater under-updating, b = -0.04 (SE = 0.04), but once again this slope was not statistically significant itself, χ^2 (1) = 1.40, p =.236.[34]

Table 3. Linear Mixed Effects Model Output in Study 1.

	Difference index			Ratio index		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
Fixed Effects						
(Intercept)	2.29	1.34	.087	0.14	0.05	.009
CRT	-0.76	0.89	.391	-0.05	0.04	.234
Political concordance	-3.51	0.93	<.001	-0.06	0.05	.202
CRT x Political concordance	1.74	0.88	.047	0.10	0.04	.011
Observations	7586			7586		
$N_{Subject}$	490			490		
$N_{Stimuli}$	16			16		
R^2 / Ω_0^2	.004 / .217			.002 / .239		
Deviance	73038.56			24978.64		

Note. For brevity, only the fixed effect estimates are displayed. Random effects estimates are reported in the SM. The ratio index model is estimated via ML because REML did not converge (the difference index model is estimated via REML). P-values in the table are estimated via Wald test. CRT = Cognitive Reflection Test (z-score).

[34] In the preregistered protocol for the test of H2, we slightly mis-specified the intended maximal random effects structure of the models. Thus, the foregoing LRT results and interaction estimates in Table 3 are from the models with a *correctly* specified random effects structure (see SM for details). For full transparency, we also report here the LRT results and fixed-effect interaction estimates from the *incorrect* maximal models, difference index: χ^2 (1) = 5.70, p =.017, b = 2.33 (SE = 0.95), ratio index: χ^2 (1) = 6.54, p =.011, b = 0.10 (SE = 0.04). As can be seen from these LRT results, the outcomes are substantively similar.

Despite the presence of an interaction in both models—consistent with H2—the pattern of results in fact provide relatively little support for the IPC facilitation hypothesis; as can be seen in the predicted values in Figure 3 (B), and the simple slopes analyses conducted above. Specifically, while subjects who scored higher on the CRT appeared more "biased" towards their political identities than their low scoring counterparts—updating more on politically *con*cordant evidence but less on politically *dis*cordant evidence—in three of four cases the simple slopes of CRT *converged towards the posterior beliefs of a Bayesian agent*. That is, instead of high CRT subjects exhibiting identity-protective deviation from Bayesian posterior beliefs, we observed that low CRT subjects exhibited the *opposite* of identity-protective bias; relative to the Bayesian prediction, they tended to over-update on politically discordant evidence, and under-update on politically concordant evidence. The updating of high CRT subjects, in contrast, was closer to that of a normative Bayesian agent. In this light, the statistically significant interaction is not particularly consistent with H2; in fact, it seems more consistent H1. More generally, this result highlights the value in assessing putative bias in political belief formation against a normative (e.g., Bayesian) benchmark (Gerber & Green, 1999; Hahn & Harris, 2014).

2.2.3.2. Exploratory Tests

To increase the sensitivity of the H2 test, we also fitted models where political identity was represented continuously (not dichotomized into a preference for Democratic or Republican Party). The details of this exploratory analysis are reported in the SM. The results closely reproduced the pattern of results reported here.

2.3. Discussion

In Study 1, we found that subjects who scored higher on the CRT deviated less from Bayesian updating overall; combining their prior beliefs with new politically-relevant information in a less biased (more normative) manner (H1). In testing H2, we found some evidence consistent with the IPC facilitation hypothesis—wherein more cognitively sophisticated subjects (indexed by CRT performance) appeared relatively more "biased" in favor of their political identities when updating their beliefs. However, when assessed against the Bayesian benchmark, this pattern of updating was revealed to be *more*, not less, normatively appropriate.

We note, however, that the pattern of subjects' belief updating vis-à-vis the Bayesian agent may depend in part upon how they interpreted the task instructions. Specifically, recall that we told them that the signals were accurate on average two out of three times. We assumed that subjects interpreted this instruction as $P(S_{TRUE}|T) = .67$; that is, the likelihood of receiving a signal saying "TRUE" assuming the statement is true is equal to 2/3 (and, analogously, we assumed they interpreted $P(S_{FALSE}|F) = .67$). However, it is possible they interpreted this instruction instead as $P(T|S_{TRUE}) = .67$; that is, the posterior probability that the statement is true upon receipt of a signal saying "TRUE". These conditional probabilities are not equivalent.

To test our assumption that subjects interpreted the task instructions as $P(S_{TRUE}|T) = .67$ not as $P(T|S_{TRUE}) = .67$, we can examine subjects' posterior beliefs after receiving "TRUE" and "FALSE" signals. Specifically, if subjects were interpreting the task instructions about signal accuracy as $P(T|S_{TRUE}) = .67$ (violating our assumption), their posterior beliefs should "stack" on 67% or 33% following receipt of a "TRUE" or "FALSE" signal, respectively. If there is no such stacking on these values in the posterior beliefs, we can be more confident that subjects did not misinterpret the task instructions as we describe above. In Figure S3 in the SM, we plot the distributions of posterior beliefs as a function of political statement stimuli and signal received ("TRUE" or "FALSE"). There is little evidence that the posterior beliefs stack on the values of 67% or 33%, respectively.

Study 2

A particular limitation of Study 1 was that we *imposed* on subjects the diagnosticity of the evidence they received; the "likelihood ratio", in Bayesian terms. Specifically, we informed subjects of the probability that signals were accurate (i.e., 2/3), and we assumed all subjects applied this knowledge uniformly in their updating behavior. This assumption is perhaps unrealistic. Indeed, the prediction of Kahan's (2016a) IPC account is that individuals' reasoning about the likelihood ratio—the diagnosticity of evidence—is conditional on their political identities. Belief updating and reasoning about the likelihood ratio are thus confounded in Study 1, possibly biasing against H2.

To resolve this issue, in Study 2 subjects provided repeated *subjective* judgments regarding the diagnosticity of signals in the task, and we used this in conjunction with their prior beliefs to

compute Bayesian posterior beliefs; isolating the process of belief updating—distinct from reasoning about the likelihood ratio. The hypotheses were the same as in Study 1 (H1 & H2). In addition, we obtained self-reported accuracy ratings on each individual signal to test the IPC facilitation hypothesis that high CRT scorers *reason* about the validity of evidence conditional on their political identities, to a greater extent than low CRT scorers (H3) (cf. Kahan, 2013).

3.1. Methods

3.1.1. Sample

We sought to double the sample size from Study 1 and collect N = 1000 subjects. A total of N = 1004 subjects completed the study. Subjects who completed Study 1 were unable to take part in Study 2.

3.1.2. Belief Update Task

The task was identical to Study 1, except for the following adjustments. First, subjects were not told the probability of receiving an accurate signal; instead, they were simply informed that signals were accurate with some fixed probability, and that signals were, on average, accurate (i.e., that this probability was > 0.5). Second, upon receipt of each signal in P1, subjects reported on a scale from 1 to 5 whether they believed that *particular* signal was accurate (1 = definitely NOT; 2 = probably NOT; 3 = not sure; 4 = probably YES; 5 = definitely YES). We refer to these as *signal accuracy ratings*. Following this rating, subjects were asked to consider *all* the signals they had seen thus far in the task, and to provide a judgment about the overall likelihood of receiving an accurate signal in the task as a whole. We refer to these as *likelihood judgments*. These judgments allowed us to infer subjects' *subjective* likelihood ratios—that is, their perception of the diagnosticity of the evidence—at repeated stages throughout the task. The likelihood judgments were provided on a sliding scale in whole integers, anchored from 51% – "signals are almost random/uninformative" – through 75% – "signals are mostly accurate/quite informative" – to 99% – "signals are almost perfectly accurate/very informative". Subjects answered comprehension questions and completed a P1 and P2 practice trial prior to starting P1. P2 was the same as in Study 1. Verbatim task instructions are available on the OSF: https://osf.io/yt3kd/.

3.1.3. Comparison with Bayesian Agent

Bayesian posterior beliefs were computed as in Study 1. The only difference was that, in Study 2, $P(S|T)$ and $P(S|\neg T)$ were defined by subjects themselves, via their subjective likelihood judgments provided immediately after each statement-signal pairing in P1. In particular, to calculate the Bayesian posterior belief for political statement i, we consulted the signal received for that statement ("TRUE" or "FALSE"), and the subjective likelihood judgment (scaled to 0-1), L, provided by subjects immediately after that statement-signal pairing. Formally, for TRUE signals,

$$P(T_i|S_{TRUE}) = \frac{P(T_i)L_i}{P(T_i)L_i + P(\neg T_i)(1 - L_i)}$$

For FALSE signals,

$$P(T_i|S_{FALSE}) = \frac{P(T_i)(1 - L_i)}{P(T_i)(1 - L_i) + P(\neg T_i)L_i}$$

We deliberately chose to consult subjects' subjective likelihood, L, provided *after* the focal statement-signal pairing (rather than before) to allow for the fact that subjects' subjective likelihood ratio may instantaneously change upon receipt of the focal signal; affecting their belief updating on the focal statement (Cheng & Hsiaw, 2017). Concretely, suppose L after trial A is 99%; that is, the subject believes signals are highly diagnostic. If they subsequently receive a signal on trial B that contradicts their strong belief about the truth of statement B, it is possible they become instantaneously more uncertain about the diagnosticity of the signals (i.e., L decreases); affecting their updating on statement B. After computing the Bayesian posterior beliefs using the above method, the two DVs were calculated as in Study 1 (difference index, ratio index).

3.1.4. Post-Task Measures

Following the belief update task, subjects completed the 7-item CRT as in Study 1 (M = 3.67, SD = 2.07, α = .76, 95% CI [.74, .79], average inter-item r = .31), and provided demographic information. There was no memory test or other exploratory measures in Study 2.

3.2. Results

3.2.1. Data Exclusions

As in Study 1, for all hypothesis tests, we exclude subjects with duplicate IP addresses (N = 12, 1.20%); retaining the earliest responses only. For H1 and H2 tests, we also exclude trials on which subjects' prior belief was provided as 0 or 100, and the signal they received was FALSE or TRUE, respectively (N trials = 346, 2.18%). Trials with missing data for prior beliefs, posterior beliefs, or likelihood judgments are excluded from H1 and H2 tests (N = 0). Trials with missing data for signal accuracy ratings are excluded from the test of H3 (N = 0). Finally, subjects who do not report a US political party preference are excluded from H2 and H3 tests (N = 0). As in Study 1, these exclusion criteria were preregistered.

3.2.2. H1: High CRT Scorers Deviate Less from Bayesian Updating

After data exclusions, we retained N = 992 for the preregistered test of H1. The analysis plan was the same as in Study 1.

3.2.2.1. Preregistered Tests

CRT performance was again negatively correlated with absolute deviation from Bayesian posterior beliefs, though the association was larger for the difference index: $\tau = -.18$, $Z = -8.04$, $p < .001$, than the ratio index: $\tau = -.07$, $Z = -3.14$, $p = .002$. The raw data are displayed in Figure 4 (A).

As in Study 1, the magnitude of the negative correlation was similar across politically concordant and politically discordant evidence: Politically discordant: $\tau = -.14$, $Z = -6.25$, $p < .001$ (difference index), $\tau = -.06$, $Z = -2.59$, $p = .010$ (ratio index), politically concordant: $\tau = -.17$, $Z = -7.66$, $p < .001$ (difference index), $\tau = -.08$, $Z = -3.40$, $p < .001$ (ratio index). As in Study 1, the latter analyses—split by political concordance—were conducted *post-hoc*.

The results thus replicate Study 1 (H1): Subjects who scored higher on the CRT deviated less from Bayesian posterior beliefs overall, combining their prior beliefs with new politically-relevant evidence in a more normative (less biased) manner.

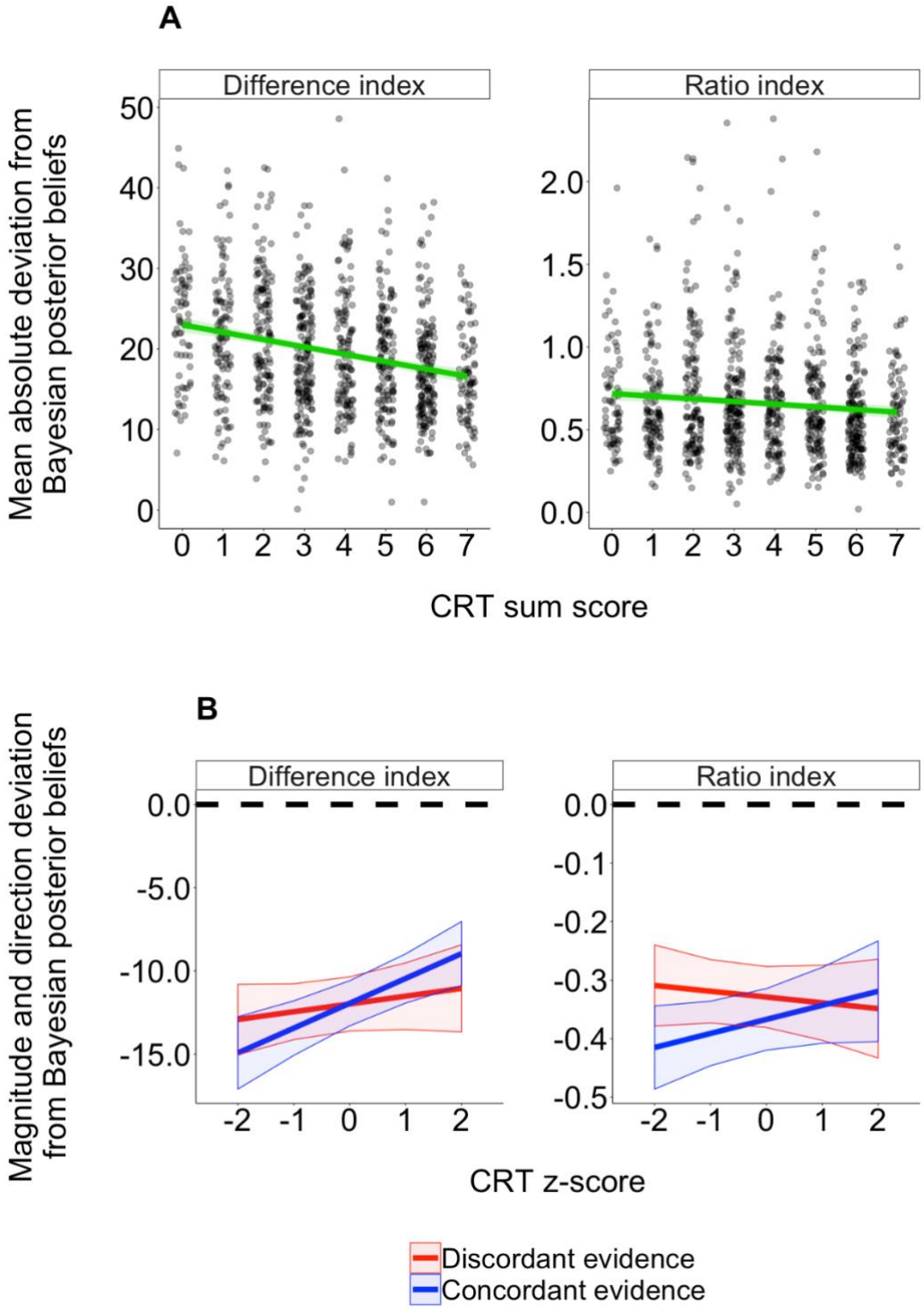


Figure 4. Deviation from Bayesian posterior beliefs as a function of CRT performance and the political concordance of evidence in Study 2. A*, mean absolute deviation from Bayesian posterior beliefs (represented at y = 0). Each point corresponds to the mean absolute deviation of one subject. Data points have slight jitter for visibility.* ***B****, predicted magnitude and direction deviation from Bayesian posterior beliefs (represented at y = 0). Predicted values are obtained from the models reported in Table 4. Shaded regions are 95% CI. CRT = Cognitive Reflection Test.*

3.2.2.2. Exploratory Tests

We repeated all the exploratory analyses conducted in Study 1 (H1) as a robustness check on the above results (reported in the SM). As before, these analyses suggest that the preregistered H1 result is robust to a variety of analytic specifications and potential confounds.

3.2.3. H2: High CRT Scorers Deviate from Bayesian Updating Conditional on their Political Identities

After data exclusions, we retained N = 992 for the preregistered test of H2. The analysis plan was the same as in Study 1. Recall that, in this analysis, DV values greater than zero imply over-updating; values smaller than zero imply under-updating; and values of exactly zero imply the normatively correct updating of a Bayesian agent. The critical test of H2 is on the interaction between CRT performance and the political concordance of the evidence in predicting these values.

3.2.3.1. Preregistered Tests

In the difference index model, the critical LRT missed the significance threshold (albeit only just), χ^2 (1) = 3.76, p =.053. That is, the interaction between CRT performance and the political concordance of evidence did not improve model fit at p <.05. Predicted values from the

difference index maximal model are displayed in Figure 4 (left panel B). Parameter estimates from this model are reported in Table 4.

Due to convergence issues, the ratio index model was not fitted with a maximal random effects structure, and we estimated degrees of freedom and p-value for the interaction term via the Satterthwaite approximation, using the *lmerTest* package in R (Kuznetsova et al., 2017). These analysis contingencies were preregistered. The critical test just missed the significance threshold, $t_{922} = 1.93$, $p = .054$. That is, the interaction between CRT performance and the political concordance of evidence did not improve model fit at $p < .05$. Predicted values from the final model are displayed in Figure 4 (right panel B). Parameter estimates are reported in Table 4.

Table 4. Linear Mixed Effects Model Output in Study 2 (Hypothesis 2).

	Difference index			Ratio index		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
Fixed Effects						
(Intercept)	-11.99	0.83	<.001	-0.33	0.03	<.001
CRT	0.47	0.44	.291	-0.01	0.01	.498
Political concordance	0.04	0.52	.939	-0.04	0.02	.030
CRT x Political concordance	1.02	0.52	.048	0.03	0.02	.053
Observations	15526			15526		
$N_{Subject}$	992			992		
$N_{Stimuli}$	16			16		
R^2 / Ω_0^2	.002 / .170			.001 / .099		
Deviance	140739.37			43212.82		

Note. For brevity, only the fixed effect estimates are displayed. Random effects estimates are reported in the SM. The ratio index model is estimated via ML because REML did not converge (the difference index model is estimated via REML). P-values in the table are estimated via Wald test (note that the Wald test p-value for the interaction in the difference index model is .048, but the preregistered LRT p-value—reported in text—was .053). CRT = Cognitive Reflection Test (z-score).

While the LRTs bearing on H2 (for both DVs) failed to meet the preregistered threshold for statistical significance in Study 2, the results do not appear to be an emphatic rejection of H2 because the tests only just missed this threshold; and the predicted values of both models are qualitatively consistent with a cross-over interaction pattern (Figure 4B). As in Study 1, however, the simple slopes of CRT performance tended towards Bayesian posterior beliefs in three of four cases. Thus, also as in Study 1, the pattern appears broadly *in*consistent with the notion that high CRT performance is associated with more *biased* belief updating conditional on political identity.

3.2.3.2. Exploratory Tests

As in Study 1, to increase the sensitivity of the H2 test we also fitted models where political identity was represented continuously (not dichotomized into a preference for Democratic or Republican Party). The details of this exploratory analysis are reported in the SM. The results were consistent with the pattern of results reported above.

3.2.4. H3: High CRT Scorers Evaluate Evidence Conditional on their Political Identities

After data exclusions, we retained N = 992 for the preregistered test of H3. In testing H3, we turn our focus from belief updating to reasoning about the accuracy of the evidence (i.e., "evidence evaluation"). In particular, H3 is that high CRT scorers' ratings of signal accuracy are conditional on their political identity to a greater extent than low CRT scorers'. In other words, subjects who score higher on the CRT are more polarized in their evaluation of evidence that is concordant vs. discordant with their political identity (cf. Kahan, 2013).

3.2.4.1. Preregistered Tests

We fitted a maximal linear mixed effects model on the trial-level data, with CRT performance and the political concordance of evidence—and their interaction—as IVs. The DV was signal accuracy ratings, provided by subjects for each signal they received in P1 of the belief update task. Recall that these ratings represent judgments about whether each signal is accurate (from

1 = definitely NOT, to 5 = definitely YES). The critical LRT on the interaction missed the statistical significance threshold, χ^2 (1) = 3.56, p =.059. That is, the interaction between CRT scores and the political concordance of evidence did not improve model fit at p <.05. Model parameter estimates are reported in Table 5. The interaction estimate is positive, however, meaning that the result is directionally consistent with H3—but just missed the preregistered threshold for statistical significance. We thus conducted exploratory sensitivity analyses (see below).

Table 5. Linear Mixed Effects Model Output in Study 2 (Hypothesis 3).

	Signal accuracy ratings		
	B	*std. Error*	*p*
Fixed Effects			
(Intercept)	3.08	0.03	<.001
CRT	-0.03	0.02	.134
Political concordance	0.61	0.05	<.001
CRT x Political concordance	0.06	0.03	.054
Observations		15872	
$N_{Subject}$		992	
$N_{Stimuli}$		16	
R^2 / Ω_0^2		.067 / .227	
Deviance		48296.82	

Note. For brevity, only the fixed effect estimates are displayed. Random effects estimates are reported in the SM. The model is estimated via ML because REML did not converge. P-values in the table are estimated via Wald test. CRT = Cognitive Reflection Test (z-score).

3.2.4.2. Exploratory Tests

To increase the sensitivity of the H3 test we fitted a model where political identity was instead represented continuously (see SM for details). Specifically, in this model we fitted a three-way interaction between (i) CRT performance, (ii) political party identity (scored 1 = Strong Democrat, to 7 = Strong Republican, midpoint-centered and standardized for analysis), and (iii) whether the signal favored Democrats or Republicans. The latter variable is computed by combination of the partisanship of the political statement (Democrat-favor or Republican-favor), and the signal type (TRUE or FALSE) (e.g., see Table 2).

In contrast to the preregistered LRT result in H3, this interaction *did* significantly improve model fit at $p < .05$: χ^2 (1) = 4.76, p =.029, b = 0.07 (SE = 0.03). The predicted values from this model are displayed in Figure 5 (A). They show a pattern consistent with H3 and the IPC facilitation hypothesis as it relates to reasoning about evidence: Namely, subjects who scored higher on the CRT were more polarized in their evaluation of evidence that was concordant vs. discordant with their political identity. Notwithstanding type I error inflation due to multiple testing, this suggests the preregistered test of H3—which treated political identity as a dichotomous variable—was somewhat obscuring the association between CRT performance, political identity, and evidence evaluation (expected under the IPC facilitation hypothesis).

However, we explored whether this pattern was better accounted for by high CRT scorers deferring more to their *prior beliefs*, rather than their political identities *per se*. To do so, we first coded whether each signal was concordant or discordant with subjects' prior belief about the truth of the political statement in question, as well as the strength of that prior belief (see SM for further details). We then *jointly* modelled the three-way interaction between (i) CRT performance and these two new prior belief variables, and (ii) CRT performance, political party identity, and whether the signal (evidence) favored Democrats or Republicans (as in the previous model).

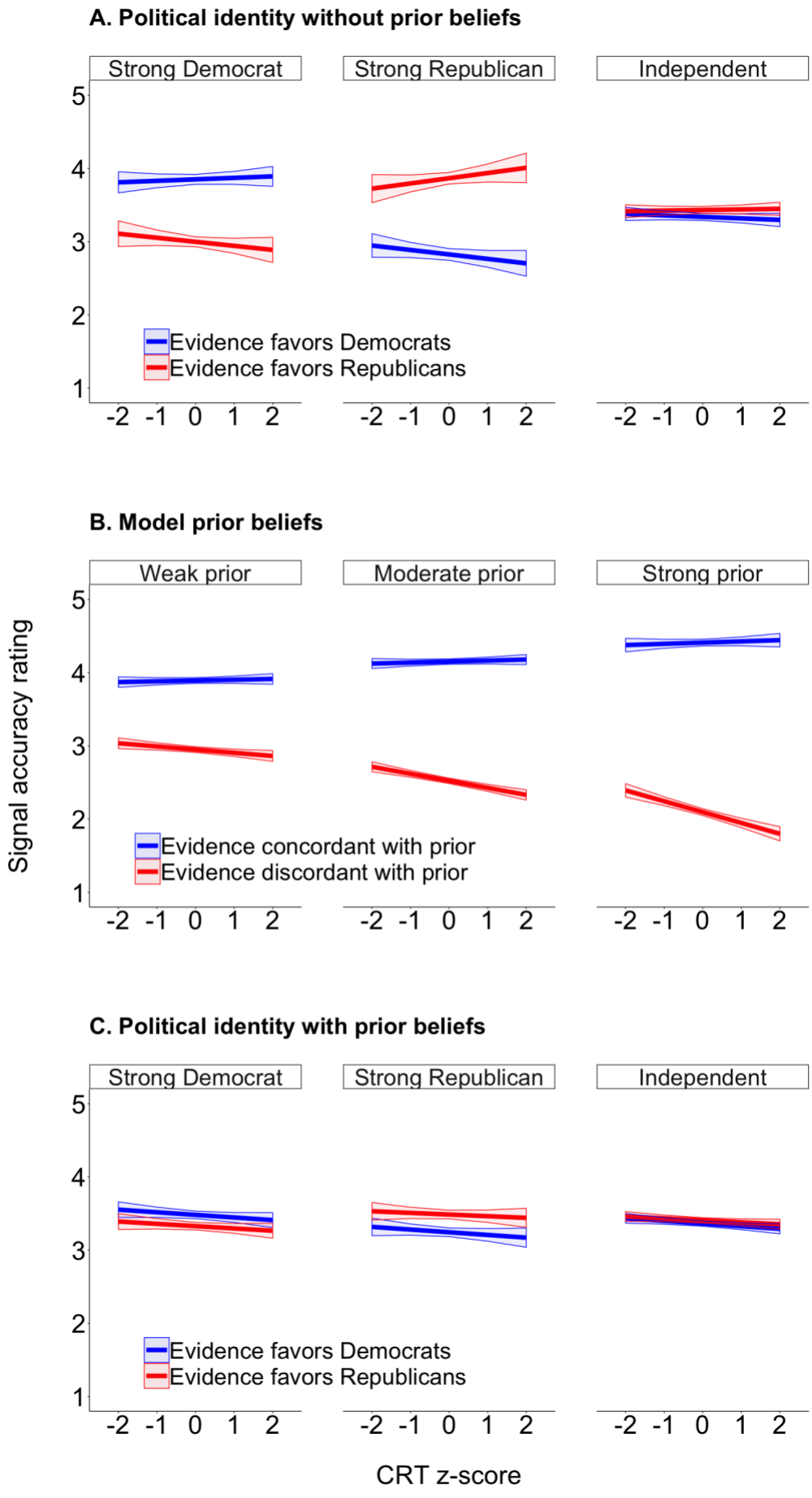

A. Political identity without prior beliefs
Strong Democrat
Strong Republican
Independent
Evidence favors Democrats
Evidence favors Republicans
B. Model prior beliefs
Weak prior
Moderate prior
Strong prior
Evidence concordant with prior
Evidence discordant with prior
C. Political identity with prior beliefs
Strong Democrat
Strong Republican
Independent
Evidence favors Democrats
Evidence favors Republicans
Signal accuracy rating
CRT z-score

Figure 5. Predicted signal (evidence) accuracy ratings in Study 2. A, C, *Strong Democrat, Strong Republican, and Independent correspond to values of -1.5, +1.5, and 0 on the midpoint-centered and standardized party identity variable, respectively.* ***B,*** *Weak, Moderate, and Strong priors correspond to values of 15, 30, and 45 on the prior belief strength variable, respectively (range: 1-50). Predicted values are obtained from the linear mixed effects models reported in text. Shaded regions are 95% CI.*

In this model, the three-way interaction with prior beliefs improved model fit, χ^2 (1) = 14.07, p <.001, b = 0.004 (SE = 0.001). This interaction is displayed in the predicted values in Figure 5 (B). The pattern illustrates that subjects scoring higher on the CRT tended to defer more strongly to their prior beliefs when rating the signals ("evidence"). In contrast to the model without prior beliefs, in this model the three-way interaction with political identity no longer improved model fit, χ^2 (1) = 0.05, p =.819, b = 0.003 (SE = 0.015). The predicted values are displayed in Figure 5 (C). They illustrate that the association between CRT performance, political identity and evidence evaluation—expected under the IPC facilitation hypothesis—is eliminated in the model with prior beliefs.

3.3. Discussion

There are three key results from Study 2. First, as in Study 1, we found that subjects who scored higher on the CRT deviated less from Bayesian posterior beliefs overall; combining their prior beliefs with new politically-relevant information in a less biased (more normative) manner (H1). This result is more compelling for the fact that subjects reported their *subjective* likelihoods on the evidence; isolating belief updating from reasoning about the evidence in our analysis. This suggests that differences in reasoning about the diagnosticity of the evidence—between subjects scoring high and low on the CRT—do not account for the association between CRT performance and deviation from Bayesian posterior beliefs that we observed.

However, a noteworthy limitation in the design of Study 2 (and Study 1) was the lack of a control group when measuring belief updating. Recall that, in these studies, subjects reported their beliefs in a first phase (P1), subsequently received evidence (signals), and then reported

their beliefs again in a second phase (P2). When taking repeated measurements of a quantity with natural variability and/or susceptibility to measurement error—in this case, beliefs about the truth of various political statements—inclusion of a control group is necessary to ensure regression to the mean does not bias estimates of belief updating (Yu & Chen, 2015).

Regression to the mean (RTM) describes the phenomenon where more extreme measurements at time 1 tend to regress towards the mean when measured again at time 2. Thus, in the case of Study 1 and Study 2, more extreme prior beliefs (i.e., closer to the likelihood scale ends of 0% and 100%) may have been associated with greater RTM measured in the posterior beliefs. If the extremity of prior beliefs differed systematically over CRT performance, differences in RTM—rather than differences in belief updating behavior *per se*—could have accounted for the association between CRT performance and deviation from Bayesian posterior beliefs. In the SM, however, we show that subjects who scored higher on the CRT did not have more extreme prior beliefs (on average) than subjects who scored lower on the CRT. Furthermore, we find that the association between CRT performance and deviation from Bayesian posterior beliefs remains after statistically adjusting for the extremity of prior beliefs (see SM). Thus, while inclusion of a control group—subjects who receive no (or unrelated) evidence—is the gold standard for separating RTM from estimates of belief updating, our supplemental analyses provide some evidence that RTM does not account for the association between CRT performance and belief updating that we observe in Studies 1 and 2.

The second key result is that, as in Study 1, there was little evidence that subjects who scored higher on the CRT deviated from Bayesian posterior beliefs conditional on the political concordance of the new information (H2). Indeed, once again, three of four simple slopes of CRT performance tended towards the posterior beliefs of a Bayesian agent (Figure 4, B); more consistent with *H1* than H2. The most striking result from Study 2 is that, after modelling individuals' subjective likelihoods, updating is estimated to be substantially less than that prescribed by the Bayesian agent. This is consistent with past work on human belief updating, where "conservatism" in updating is regularly observed (Hahn & Harris, 2014), and suggests that subjects did not uniformly apply the given likelihood ratio (of 2, i.e., 2/3 ÷ 1/3) in Study 1.

As in Study 1 we examined whether subjects interpreted the task instruction about signal accuracy as $P(T|S_{TRUE}) = .67$, violating our assumption that they interpreted it as $P(S_{TRUE}|T) = .67$. Accordingly, in Figure S4 in the SM we plot the distributions of posterior beliefs as a function of political statement stimuli and signal received ("TRUE" or "FALSE"). As in Study 1, there is little evidence that the posterior beliefs stack on the values of 67% or 33%, respectively. This suggests that subjects did not interpret the task instructions about signal accuracy as $P(T|S_{TRUE}) = .67$ (or, analogously, as $P(T|S_{FALSE}) = .33$).

The third key result is that the association between CRT performance, political identity, and evidence evaluation (H3)—expected under the IPC facilitation hypothesis—all but disappeared after modelling prior beliefs. In particular, in an exploratory sensitivity analysis, we found evidence to suggest that our preregistered use of a binary political identity measure for testing H3 was insensitive to the association between CRT performance, political identity, and evidence evaluation (Figure 5A). In a second exploratory analysis, however, we found evidence to suggest this association was subsumed by an association between CRT performance, *prior beliefs*, and evidence evaluation (Figure 5B and 5C).

These results indicate that the association between CRT performance and political identity in predicting reasoning about evidence may be confounded by prior beliefs. This is intuitive given that prior beliefs (about political issues) are often correlated with political identity; for example, because of selective exposure to information (Bolin & Hamilton, 2018; Rodriguez et al., 2017). Our data support this conjecture: The mean correlation between (i) the prior belief that a statement was true and (ii) political party affiliation, over the 8 pro-Democratic political statements was -.35 (SD = 0.07, range = [-.45, -.26]). In other words, Republican subjects believed pro-Democratic statements were more likely to be false than did Democrats. For the 8 pro-Republican statements, the reverse was true ($M_{correlation} = .33$, SD = 0.06, range = [.26, .42]. Explicit modelling of this shared variance suggested that the interaction between CRT performance and *prior beliefs* was decisive in predicting reasoning about the validity of the evidence. However, these results were exploratory and must be replicated in confirmatory analysis to carry greater inferential weight.

Study 3

In Study 3, therefore, we conducted a conceptual replication of the experiment reported by Kahan (2013); the results of which are cited as evidence that cognitive sophistication facilitates identity-protective reasoning in political belief formation. Importantly, unlike in Kahan's (2013) experiment, we measured and modelled relevant prior beliefs, alongside political identity.

In the original experiment (Kahan, 2013), subjects completed the CRT before being randomly assigned to treatments where they were asked to evaluate the validity of the CRT itself. Across treatments, information about the CRT was manipulated. Specifically, in both experimental treatments, the CRT was said to measure "open-minded and reflective thinking", but, in one treatment, it was stated that climate change *skeptics* tended to score higher in the test. In the other experimental treatment, in contrast, it was stated that climate change *believers* tended to score higher in the test. Kahan (2013) found that subjects' evaluation of the validity of the CRT in measuring "open-minded and reflective thinking" was conditional on their political identities and treatment assignment: "Liberal Democrats" rated the test as more valid in the treatment where believers were said to score higher (vs. the treatment were skeptics were said to score higher), and vice versa for "conservative Republicans". Crucially, the data indicated that this conditioning on political identity when reasoning was *greatest* among those who scored highest on the CRT; consistent with the IPC facilitation hypothesis.

Based on the alternative hypothesis outlined in the introduction—and suggested by the exploratory results of Study 2—we tested whether Kahan's (2013) results are confounded by an association between CRT performance and conditioning on *prior beliefs* when reasoning about the evidence. That is, whether cognitively sophisticated subjects defer more strongly to their prior beliefs—rather than their political identities *per se*—when reasoning about the validity of the psychological test.

4.1. Methods

4.1.1. Sample

We sought to collect the approximate N-per-treatment used in Kahan (2013, N = 583); our target sample size was thus N = 1200 (i.e., N-per-treatment = 600). A total of N = 1215 subjects completed the study.

4.1.2. Design & Procedure

To measure and model prior beliefs alongside political identity, we simplified Kahan's (2013) original design. Specifically, in the original design, we identified (at least) three prior beliefs that may conceivably have affected variance in ratings of the CRT's validity as a measure of open-mindedness: Subjects' beliefs about (i) the relative open-mindedness of climate-skeptics vs. climate-believers, (ii) how open-minded they themselves are, and (iii) how they just performed on the CRT.

To reduce the number of relevant beliefs (and thus streamline their measurement and modelling) in our design, subjects rated the validity of a different but related test – which we labeled the "Open-Mindedness Test" – comprised of three self-report questions taken from the *Actively Open-minded Thinking* scale (Baron, 2008; Haran et al., 2013). Importantly, subjects did *not* complete this test themselves. This modification served to remove the influence of prior belief (iii), and reduce the influence of prior belief (ii), described above. To measure prior belief (i) – subjects' belief about the relative open-mindedness of climate-skeptics vs. climate-believers – we asked them who they considered to be more open-minded: Someone who believes climate change is happening vs. someone who is skeptical climate change is happening (scored from 1 to 7, anchored 1 = believer is definitely more open-minded, 4 = neither is more open-minded than the other, 7 = skeptic is definitely more open-minded). This question was embedded within a list of nine additional (distractor) questions related to other targets' open-mindedness; for example, the relative open-mindedness of females vs. males. The verbatim task instructions are available on the OSF: https://osf.io/yt3kd/.

When rating the validity of the "Open-Mindedness Test", subjects were randomly assigned to one of two treatments. In both treatments, subjects were told that psychologists were still evaluating the validity of the test, but that a higher score is taken to indicate greater open-mindedness. In one treatment – *believers are open-minded* – subjects were asked whether they agreed that this test supplied good evidence of how open-minded someone is, on the assumption that future research finds that individuals who *believe* climate change is happening tend to score higher than individuals who are *skeptical* climate change is happening. In the other treatment – *skeptics are open-minded* – subjects provided the same judgment, but on the reverse assumption: That future research finds that climate change *skeptics* tend to score higher

than those who *believe* climate change is happening. These ratings were provided on a 1-7 scale, as in Kahan (2013) (anchored: 1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = neither agree nor disagree, 5 = somewhat agree, 6 = agree, 7 = strongly agree).

To recap, our modified design comprised three key components: (i) ratings of the validity of the "Open-Mindedness Test" (experimental component), (ii) ratings of how open-minded climate-skeptics vs. climate-believers are (prior beliefs), and (iii) completion of the 7-item CRT ($M = 3.43$, $SD = 2.09$, $\alpha = .77$, 95% CI [.75, .79], average inter-item $r = .32$). The order of (i) and (ii) was counterbalanced across subjects, and (iii) was always completed in between (i) and (ii). At the end of the study, subjects provided demographic information, including their political party affiliation (from 1-7, anchored 1 = Strong Democrat, 7 = Strong Republican) and political ideology (from 1-5, anchored 1 = Very liberal, 5 = Very conservative), following Kahan (2013).

4.2. Results

4.2.1. Data Exclusions

For all hypothesis tests, we excluded N = 14 (1.15%) subjects who were duplicate respondents (determined by their unique MTurk ID/IP address); retaining the earliest responses only. This exclusion criterion was preregistered. In addition to the preregistered exclusion criterion, N = 1 (0.08%) subject did not report their political party affiliation or political ideology, and was thus unable to be included in the tests of H1 and H3.

4.2.2. H1: High CRT Scorers Evaluate Evidence Conditional on their Political Identities

After data exclusions, we retained N = 1200 for the preregistered test of H1. H1 is a conceptual replication of Kahan (2013), and states that subjects who score higher on the CRT are more polarized in their evaluation of the "open-mindedness test", conditional on their political identities and treatment assignment.

4.2.2.1. Preregistered Test

We fitted a linear regression model with three variables: (i) CRT performance (z-score), (ii) treatment assignment [0 = believer-is-open-minded, 1 = skeptic-is-open-minded], and (iii) political identity. The DV was agreement that the test supplies good evidence of how open-minded someone is (higher scores = greater agreement). To compute the political identity variable, we summed the standardized and midpoint-centered party affiliation and political ideology variables (following Kahan, 2013). The political identity variable thus ranged from -3.75 (very liberal/strong Democrat) through 0 (moderate/Independent) to +3.75 (very conservative/strong Republican). The three-way interaction between (i), (ii), and (iii) is the test of theoretical interest.

As per the model parameter estimates reported in Table 6, the three-way interaction was not statistically significant. The predicted values from this model (2) are displayed in Figure 6 (A). They indicate that subjects with high and low scores on the CRT deferred to their political identities to the same extent when evaluating the test, conditional on treatment assignment. This result is inconsistent with H1 and the results reported by Kahan (2013).

4.2.2.2. Exploratory Test

We also fitted an ordered logistic regression model with the same variables as a check on the robustness of the preregistered H1 result. In this model, similarly, the key three-way interaction did not improve fit (at $p < .05$): χ^2 (1) = 0.05, p =.830, b = 0.01 (SE = 0.06).

Table 6. Linear Regression Model Output in Study 3 (Hypothesis 1).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	4.53	0.07	<.001	4.53	0.07	<.001
Treatment	-0.53	0.09	<.001	-0.53	0.09	<.001
Political identity	-0.19	0.03	<.001	-0.19	0.03	<.001
CRT	-0.18	0.07	.007	-0.18	0.07	.007
Treatment x Pol ID	0.37	0.05	<.001	0.37	0.05	<.001
Treatment x CRT	0.01	0.09	.951	0.01	0.09	.881
Pol ID x CRT	-0.01	0.02	.792	-0.01	0.03	.673
Treatment x Pol ID x CRT				0.02	0.05	.740
Observations	1200			1200		
R^2 / adj. R^2	.109 / .105			.109 / .104		
Deviance	2859.86			2859.60		

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

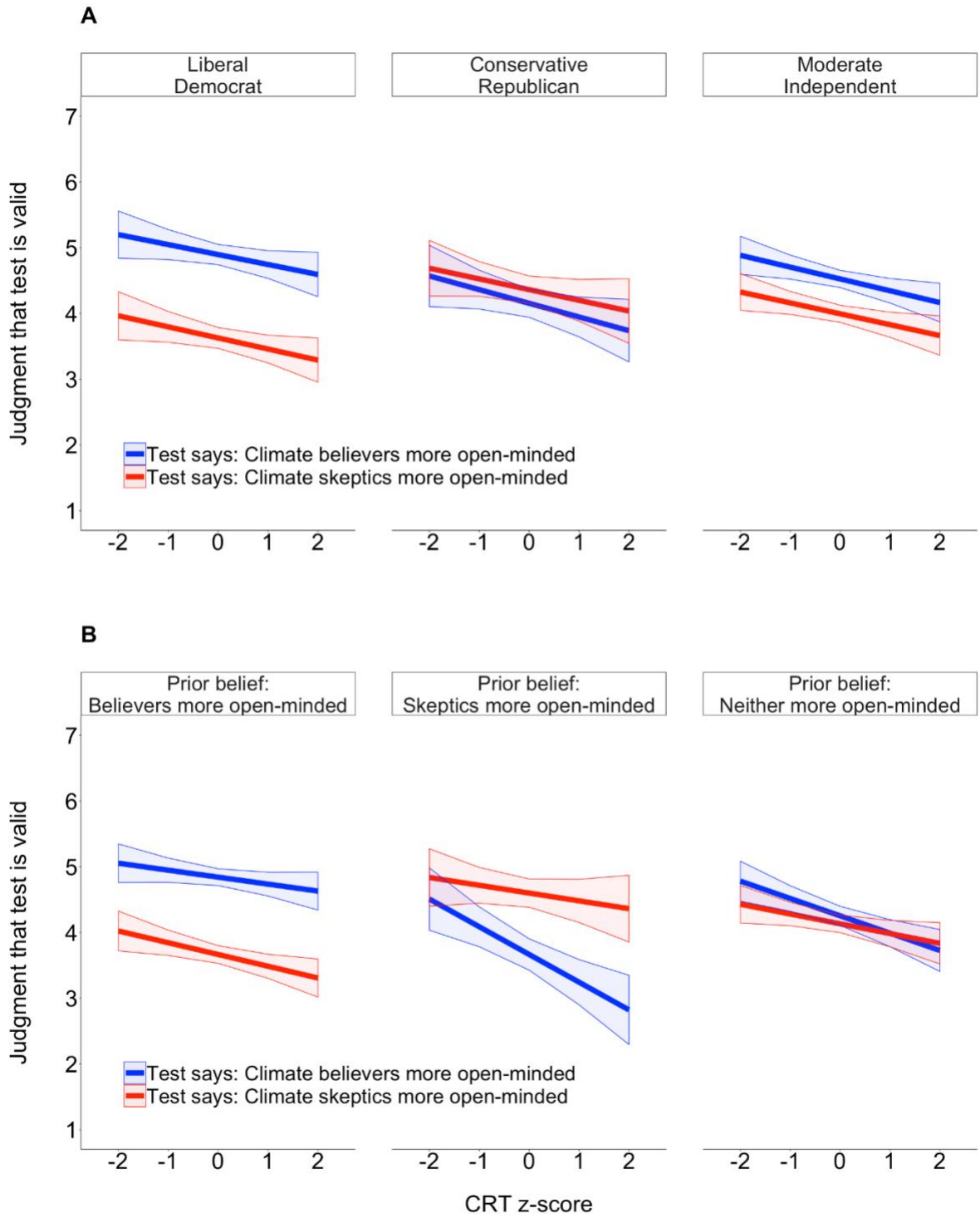


Figure 6. Predicted test validity judgments in Study 3. A*, Liberal Democrat, Conservative Republican, and Moderate Independent correspond to values of -2, +2, and 0 on the political identity variable, respectively (range: -3.75, +3.75).* ***B****, Prior belief: Believers, Skeptics, and Neither more open-minded correspond to values of -1, +1, and 0 on the midpoint-centered and standardized prior belief variable, respectively. Predicted values are obtained from the models*

(2) reported in Tables 6 (A) and 7 (B). Shaded regions are 95% CI. CRT = Cognitive Reflection Test.

4.2.3. H2: High CRT Scorers Evaluate Evidence Conditional on their Prior Beliefs

After data exclusions, we retained N = 1201 for the preregistered test of H2. H2 states that subjects who score higher on the CRT are more polarized in their evaluation of the "open-mindedness test" conditional on their prior beliefs and treatment assignment (political identity is not modelled in the test of H2, see H3).

4.2.3.1. Preregistered Test

We fitted a linear regression model with three variables: (i) CRT performance (z-score), (ii) treatment assignment [0 = believer-is-open-minded, 1 = skeptic-is-open-minded], and (iii) prior belief in the relative open-mindedness of climate-believers vs. climate-skeptics (standardized and midpoint-centered, original scale: 1 = believer is definitely more open-minded, 7 = skeptic is definitely more open-minded). The DV was the same as in the test of H1.

As per the model parameters reported in Table 7, the three-way interaction was statistically significant. The predicted values from this model (2) are displayed in Figure 6 (B). They indicate that subjects who scored higher on the CRT deferred to their prior beliefs to a greater extent when evaluating the test (conditional on treatment assignment); driven by individuals who believed that climate *skeptics* are more open-minded (central panel B). This result is broadly consistent with H2 and the exploratory results of Study 2 regarding prior beliefs.

4.2.3.2. Exploratory Test

As in H1, we fitted an ordered logistic regression model with the same variables as a check on the robustness of the preregistered H2 result. In this model, the key three-way interaction improved model fit (at $p < .05$): χ^2 (1) = 4.28, p =.039, b = 0.22 (SE = 0.10).

Table 7. Linear Regression Model Output in Study 3 (Hypothesis 2).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	4.27	0.07	<.001	4.25	0.07	<.001
Treatment	-0.15	0.10	.135	-0.12	0.10	.229
Prior belief	-0.57	0.06	<.001	-0.59	0.06	<.001
CRT	-0.21	0.07	.002	-0.26	0.07	<.001
Treatment x Prior	1.03	0.09	<.001	1.06	0.09	<.001
Treatment x CRT	0.01	0.09	.881	0.12	0.10	.240
Prior x CRT	-0.06	0.04	.145	-0.16	0.06	.009
Treatment x Prior x CRT				0.19	0.08	.025
Observations	1201			1201		
R^2 / adj. R^2	.163 / .159			.166 / .162		
Deviance	2687.11			2675.78		

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

4.2.4. H3: High CRT Scorers Do Not Evaluate Evidence Conditional on their Political Identities More than Low CRT Scorers after Modelling Prior Beliefs

After data exclusions, we retained N = 1200 for the preregistered test of H3. H3 states that the association between CRT performance, political identity, and evidence evaluation is better accounted for by an association between CRT performance, prior beliefs, and evidence evaluation. Since H1 was rejected, the test of H3 is somewhat moot. Nevertheless, there is value in testing whether the association with prior beliefs holds after modelling the shared variance between political identity and prior beliefs; especially given that these two variables

are moderately correlated (replicating the correlations in Study 2), r (1198) = .42, 95% CI [.37, .47], p <.001.

4.2.4.1. Preregistered Test

We jointly modelled the association between (i) CRT performance (z-score), prior beliefs, and treatment assignment, and (ii) CRT performance (z-score), political identity, and treatment assignment. The DV was the same as in the test of H1 and H2. The results of the linear regression model showed that the interaction with prior beliefs remained of similar size, positive, and statistically significant [b = 0.20, SE = 0.09, p =.035]. The interaction with political identity also remained similarly sized and statistically non-significant [b = -0.03, SE = 0.05, p =.621] (full model parameter estimates are reported in the SM). This result is consistent with H3 and the exploratory results of Study 2 regarding prior beliefs.

4.2.4.2. Exploratory Test

In an ordered logistic regression model, the three-way interaction with prior beliefs just missed the p <.05 significance threshold, χ^2 (1) = 3.70, p =.054, b = 0.23 (SE = 0.12). In addition, as in the preregistered analyses, the interaction with political identity did not improve model fit, χ^2 (1) = 0.22, p =.638, b = -0.03 (SE = 0.06).

4.3. Discussion

The key result from Study 3 is that subjects who scored higher on the CRT deferred more strongly to their prior beliefs when reasoning about evidence concordant vs. discordant with those beliefs (H2, H3); a result driven primarily by subjects who believed climate change skeptics are more open-minded and who received evidence to the contrary (Figure 6, central panel B). In contrast, we found no evidence that subjects who scored higher on the CRT deferred more strongly to their political identities when reasoning about the evidence (i.e., rejection of H1, Figure 6A). In other words, we did not replicate the result reported by Kahan (2013) and expected under the IPC facilitation hypothesis.

While our design was slightly modified from the original experiment (Kahan, 2013) to simplify the measurement and modelling of prior beliefs, the IPC facilitation hypothesis makes a clear prediction about the pattern of results expected in Study 3. That is, individuals scoring higher on the CRT would "*use* their distinctive analytic proficiencies to form identity-congruent assessments of [the] evidence" (p.3, Kahan, 2017, emphasis in original). According to LeBel and colleagues' (2017) replication taxonomy, our study would be classified somewhere between a "very close" and "close" replication (p. 256). Moreover, we had a slightly larger sample size than the original experiment, and, in addition, the distribution of CRT scores in our data (reported in SM) were less skewed than in the original, where over a third of subjects scored zero. This provided favorable conditions for the expected result to emerge. Finally, we *did* observe the hypothesized interaction between CRT performance and prior beliefs (H2, H3), which suggests that prior beliefs may have confounded the result reported by Kahan (2013).

However, another dimension on which our study differed from that reported in Kahan (2013) was the sampling population. In particular, we recruited a convenience sample via Amazon's Mechanical Turk (MTurk). It is well documented that MTurk samples skew more liberal/Democrat than the general US population (Chandler & Shapiro, 2016), and have substantial prior experience with the CRT (Chandler et al., 2014). We consider it unlikely that experience with the CRT accounts for the difference in our results, for reasons described above regarding the theoretically favorable distribution of scores in our sample, as well as work indicating that the predictive validity of the CRT is robust to multiple exposures (Bialek & Pennycook, 2017; Meyer et al., 2018; Stagnaro et al., 2018). However, it is possible that the skew in our political identity variable prevented us from detecting the pattern of results expected under the IPC facilitation hypothesis (Kahan, 2013; Kahan, 2016a).

Study 4

In Study 4, we therefore conducted a further replication of the experiment reported in Kahan (2013). In this study, we drew our sample from a population more demographically representative of the general US population and (likely to be) less experienced with the CRT than the MTurk population.

5.1. Methods

5.1.1. Sample

We recruited the Study 4 sample from Lucid (https://luc.id/), a marketplace for online survey research that uses quota sampling to match respondents to census demographics. Importantly, compared to MTurk, samples of US adults recruited via Lucid are (a) closer to national benchmarks in reported political party identification, and (b) much closer in reported political ideology (Coppock & McClellan, 2017). Samples recruited via Lucid are also better matched to national benchmarks than MTurk on demographics such as age, gender, education, and ethnicity, as well as on personality traits (Coppock & McClellan, 2017).

We sought to collect N = 2000, increasing the target N-per-treatment from N = 600 (in Study 3) to N = 1000 in the current study. A total of N = 2060 subjects completed the study. The sample was slightly left-of-center on a 7-point party identification scale (M = 3.87, SD = 1.80, scored from 1 = Strong Democrat to 7 = Strong Republican), and slightly right-of-center on a 5-point ideology scale (M = 3.04, SD = 1.04, scored from 1 = Very liberal to 5 = Very conservative) (distributions reported in the SM). This pattern—lean Democrat, lean conservative—is qualitatively similar to the pattern observed in the 2008 and 2012 *American National Election Survey*, and is distinct from the pattern observed in MTurk samples, which skew Democrat and liberal (for a direct comparison, see Table 2 in Coppock & McClellan, 2017). Furthermore, replicating the representative US sample in Kahan (2013), the correlation between CRT performance and political identity (combined party and ideology variables) in our Study 4 sample was not significantly different from zero, r (2050) = .005, 95% CI [-.04, .05], p =.831.

5.1.2. Design & Procedure

The design and procedure were identical to Study 3, except for the following two adjustments. First, we made two changes to the list of distractor questions that asked about other targets' open-mindedness (i.e., targets *other than* climate skeptics vs. believers). Specifically, we replaced the gun control and abortion targets with supporters/opponents of (a) genetically modified food and (b) driverless cars, respectively. We did this to avoid priming the political identity of respondents in areas unrelated to our focus (climate change).

Second, we slightly changed the wording of the experimental treatment delivered to subjects. Specifically, we more closely followed the treatment wording administered by Kahan (2013). Recall that, in Study 3, we asked subjects to rate the validity of the "open-mindedness test", *assuming* future research finds that climate change skeptics [believers] tend to score higher in the test. In Study 4, we removed this conditional statement. Specifically, we asked subjects to rate the validity of the test after simply informing them that "among a group of subjects in one recent study, the researchers found that people who reject [accept] evidence of climate change tend to score higher on the test than people who accept [reject] evidence of climate change." This more closely reflects the wording administered by Kahan (2013). The verbatim task instructions are available on the OSF: https://osf.io/yt3kd/.

All other aspects of the design and procedure were identical to Study 3. We note that, as in Kahan (2013), CRT scores were skewed towards the lower end of the scale in the Study 4 sample (7-item CRT: M = 1.99, SD = 1.74, α = .68, 95% CI [.66, .70], average inter-item r = .24) (distribution reported in the SM).

5.2. Results

5.2.1. Data Exclusions

For all hypothesis tests, we excluded N = 7 (0.34%) subjects who were duplicate respondents (determined by IP address); retaining the earliest responses only. This exclusion criterion was preregistered. In addition to the preregistered exclusion criterion, N = 1 (0.05%) subject did not report their political party affiliation, and was thus unable to be included in the tests of H1 and H3.

5.2.2. H1: High CRT Scorers Evaluate Evidence Conditional on their Political Identities

After data exclusions, we retained N = 2052 for the preregistered test of H1. As in Study 3, H1 is a conceptual replication of Kahan (2013), and states that subjects who score higher on the CRT are more polarized in their evaluation of the "open-mindedness test", conditional on their political identities and treatment assignment.

5.2.2.1. Preregistered Test

As in Study 3, we fitted a linear regression model with three variables: (i) CRT performance (z-score), (ii) treatment assignment [0 = believer-is-open-minded, 1 = skeptic-is-open-minded], and (iii) political identity. The DV was agreement that the test supplies good evidence of how open-minded someone is. The political identity variable was calculated as in Study 3 (higher values = more conservative/Republican).

As per the model parameter estimates reported in Table 8, in contrast to Study 3 this time the key three-way interaction was statistically significant. The predicted values from this model (2) are displayed in Figure 7 (A). They show that subjects who scored higher on the CRT deferred to their political identities to a greater extent when evaluating the test (conditional on treatment assignment); driven by subjects who identified as liberal/Democrat (Figure 7, left panel A). This result is broadly consistent with the IPC facilitation hypothesis and the results reported in Kahan (2013).

5.2.2.2. Exploratory Test

We fitted an ordered logistic regression model with the same variables as a check on the robustness of the preregistered H1 result. In this model, similarly, the key three-way interaction improved fit: χ^2 (1) = 5.74, p =.017, b = 0.10 (SE = 0.04).

Table 8. Linear Regression Model Output in Study 4 (Hypothesis 1).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	4.70	0.05	<.001	4.70	0.05	<.001
Treatment	-0.29	0.07	<.001	-0.30	0.07	<.001
Political identity	-0.13	0.03	<.001	-0.13	0.03	<.001
CRT	-0.07	0.05	.116	-0.08	0.05	.095
Treatment x Pol ID	0.23	0.04	<.001	0.23	0.04	<.001
Treatment x CRT	-0.20	0.07	.003	-0.19	0.07	.007
Pol ID x CRT	0.02	0.02	.329	-0.03	0.03	.189
Treatment x Pol ID x CRT				0.10	0.04	.004
Observations	2052			2052		
R^2 / adj. R^2	.044 / .041			.047 / .044		
Deviance	4858.41			4839.00		

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

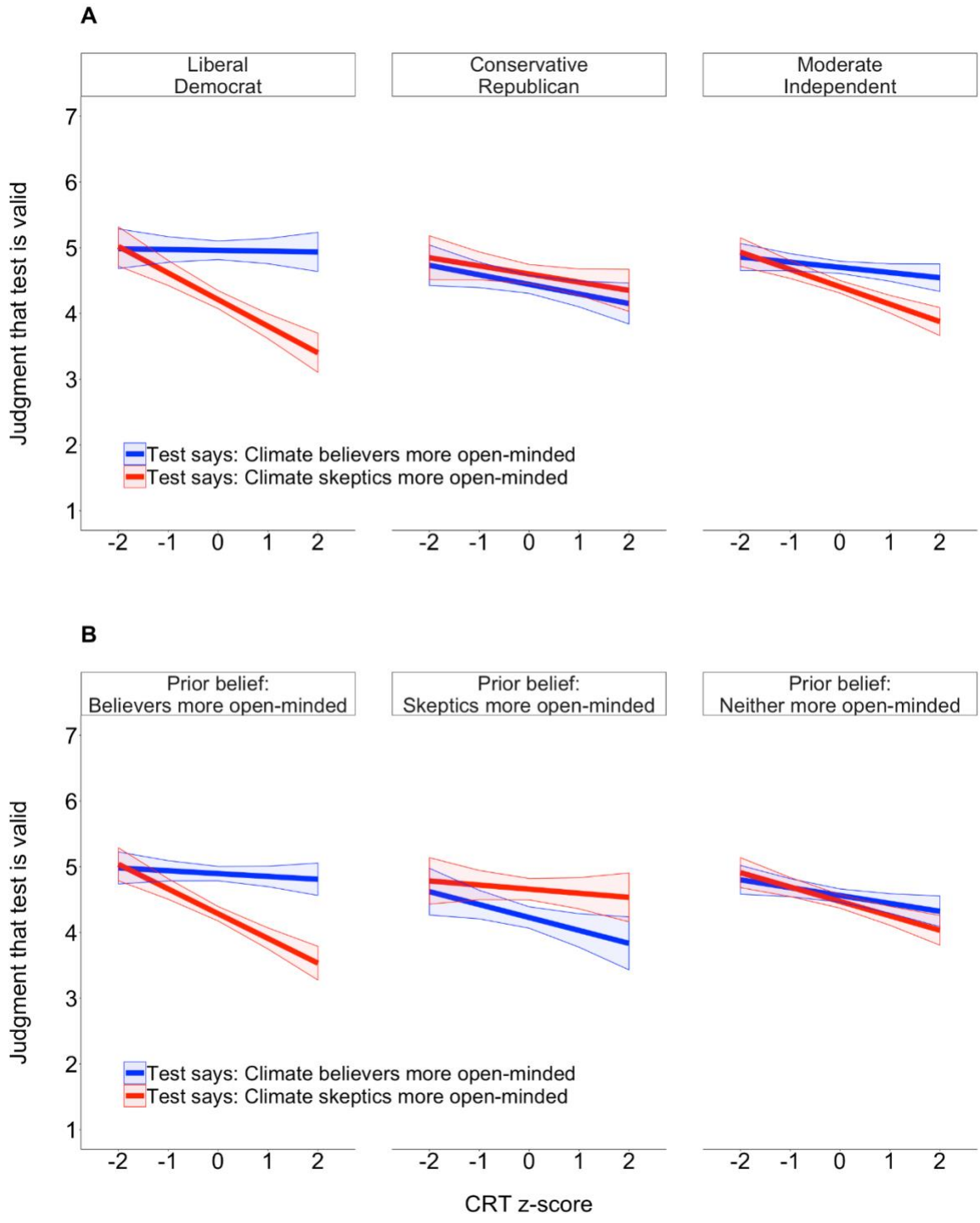


Figure 7. Predicted test validity judgments in Study 4. A, *Liberal Democrat, Conservative Republican, and Moderate Independent correspond to values of -2, +2, and 0 on the political identity variable, respectively (range: -3.58, +3.58).* ***B,*** *Prior belief: Believers, Skeptics, and Neither more open-minded correspond to values of -1, +1, and 0 on the midpoint-centered and standardized prior belief variable, respectively. Predicted values are obtained from the models*

(2) reported in Tables 8 (A) and 9 (B). Shaded regions are 95% CI. CRT = Cognitive Reflection Test.

5.2.3. H2: High CRT Scorers Evaluate Evidence Conditional on their Prior Beliefs

After data exclusions, we retained N = 2053 for the preregistered test of H2. As in Study 3, H2 is that individuals who score higher on the CRT are more polarized in their evaluation of the "open-mindedness test" conditional on their prior beliefs and treatment assignment.

5.2.3.1. Preregistered Test

We fitted a linear regression model with three variables: (i) CRT performance (z-score), (ii) treatment assignment [0 = believer-is-open-minded, 1 = skeptic-is-open-minded], and (iii) prior belief in the relative open-mindedness of climate-believers vs. climate-skeptics (higher values = skeptic relatively more open-minded). The DV was the same as in the test of H1.

As per the model parameter estimates reported in Table 9, the three-way interaction was statistically significant. The predicted values from this model (2) are displayed in Figure 7 (B). They show that subjects who scored higher on the CRT deferred to their prior beliefs to a greater extent when evaluating the test (conditional on treatment assignment); though, in contrast to Study 3, this effect was primarily driven by subjects who considered climate change *believers* to be more open-minded (Figure 7, left panel B). This result is consistent with H2 and replicates the general pattern observed in Study 3 regarding prior beliefs.

5.2.3.2. Exploratory Test

As in H1, we fitted an ordered logistic regression model with the same variables as a check on the robustness of the preregistered H2 result. In this model, the key three-way interaction improved model fit: χ^2 (1) = 9.80, p =.002, b = 0.26 (SE = 0.08).

Table 9. Linear Regression Model Output in Study 4 (Hypothesis 2).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	4.58	0.05	<.001	4.56	0.05	<.001
Treatment	-0.11	0.07	.144	-0.09	0.07	.220
Prior belief	-0.32	0.05	<.001	-0.33	0.05	<.001
CRT	-0.07	0.05	.165	-0.12	0.05	.021
Treatment x Prior	0.49	0.07	<.001	0.52	0.07	<.001
Treatment x CRT	-0.19	0.07	.005	-0.10	0.07	.174
Prior x CRT	0.05	0.04	.199	-0.08	0.05	.141
Treatment x Prior x CRT				0.23	0.07	.001
Observations	2053			2053		
R^2 / adj. R^2	.052 / .050			.057 / .054		
Deviance	4814.429			4789.695		

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

5.2.4. H3: High CRT Scorers Do Not Evaluate Evidence Conditional on their Political Identities More than Low CRT Scorers after Modelling Prior Beliefs

After data exclusions, we retained N = 2052 for the preregistered test of H3. As in Study 3, H3 states that the association between CRT performance, political identity, and evidence evaluation is accounted for by an association between CRT performance, prior beliefs, and evidence evaluation. For comparison with Study 3, we note that the correlation between (i) prior belief in the relative open-mindedness of climate change skeptics vs. believers, and (ii) political identity is moderate, r (2050) = .28, 95% CI [.24, .32], p <.001.

5.2.4.1. Preregistered Test

As in Study 3, we jointly modelled the association between (i) CRT performance (z-score), prior beliefs, and treatment assignment, and (ii) CRT performance (z-score), political identity, and treatment assignment. The DV was the same as before.

The results of the joint linear regression model showed that both interactions reduced in size. Consequently, the interaction with political identity no longer improved model fit (at $p < .05$) [$b = 0.06$, $SE = 0.04$, $p = .111$]. The interaction with prior beliefs remained below the preregistered threshold of statistical significance (albeit on the borderline) [$b = 0.15$, $SE = 0.08$, $p = .049$] (full model parameter estimates are reported in the SM). This result is thus (weakly) consistent with H3.

5.2.4.2. Exploratory Test

In an ordered logistic regression model, the three-way interaction with prior beliefs contributed to model fit (at $p < .05$), χ^2 (1) = 4.20, $p = .040$, $b = 0.18$ ($SE = 0.09$). The three-way interaction with political identity, as in the preregistered H3 test, did not, χ^2 (1) = 1.44, $p = .230$, $b = 0.05$ ($SE = 0.04$).

5.3. Discussion

There are three key findings from Study 4. The first is that—contrary to Study 3—we successfully replicated the result reported by Kahan (2013) and expected under the IPC facilitation hypothesis. Specifically, subjects who scored higher on the CRT deferred more strongly to their political identity when reasoning about evidence concordant vs. discordant with that identity (H1). This result was driven entirely by subjects who identified as liberal/Democrat (Figure 7, left panel A). While this difference in the corroboration of H1—compared with the result of Study 3—is plausibly due to the change in sampling population, it might also be due to other (albeit minor) changes we made to the design/procedure in Study 4 (see Methods).

The second key result is that we replicated the general pattern observed in Study 2 and Study 3 regarding prior beliefs: Subjects who scored higher on the CRT deferred more strongly to

their prior beliefs when reasoning about evidence concordant vs. discordant with those beliefs (H2); a result driven primarily by subjects who considered believers in climate change to be more open-minded (Figure 7, left panel B). While the specifics of this interaction differed from that in Study 3—where the effect localized to those who considered climate *skeptics* to be more open-minded (Figure 6, central panel B)—the general pattern is again consistent with H2.

The third key result is that, after modelling both three-way interactions together—political identity *and* prior beliefs—their independent contributions to model fit were both reduced; though this reduction appeared more emphatic for political identity (H3). Overall, this result is consistent with the preregistered result of Study 3 (H3) and the exploratory results of Study 2 regarding prior beliefs. That is, the interaction between CRT performance, political identity, and reasoning about evidence—expected under the IPC facilitation hypothesis—does not survive the counterpart modelling of prior beliefs.

Study 5

In Study 5, we conducted a final empirical test of the hypothesis that individuals who score higher on the CRT defer more to their prior beliefs when reasoning about information in the political domain. In particular, we re-analyzed previously published data collected during the 2016 US presidential election (Tappin et al., 2017). In that study, supporters of then presidential candidates Donald Trump and Hillary Clinton read about election polls where it was highlighted that *either* some polls predicted Trump was more likely to win the election, *or* some polls predicted Clinton was more likely to win the election. Before the polling information, subjects reported their prior beliefs about which candidate was more likely to win. After reading the polling information, subjects were asked whether they considered polls (in general) informative. As reported in the paper (Tappin et al., 2017, supplementary material), subjects' responses to this question were conditional on whether the just-presented polling results were concordant or discordant with their prior belief about who would win (in the expected fashion).

While the CRT was not administered in that particular study, many of the subjects had taken (or had yet to take) part in unrelated studies—conducted in the senior authors' lab between 2012 and 2017—which *did* administer the CRT (these studies are aggregated in Stagnaro et

al., 2018). As a result, we could retroactively match the CRT performance of subjects in those studies to the prior beliefs and polling evaluations provided by the same subjects in Tappin et al. (2017). This allowed us to again test the hypothesis that people who score higher on the CRT defer more strongly to their prior beliefs when reasoning about the validity/accuracy of evidence that is concordant vs. discordant with those beliefs.

In addition, as a function of the design used in Tappin et al. (2017), subjects' prior belief about which candidate was more likely to win the election was only weakly correlated with who they *preferred* to win. Assuming that candidate preference is a proxy for political identity allowed us to test the association between CRT performance and prior beliefs largely independent of political identity—in contrast to Study 2-4, where prior beliefs and political identity were moderately or strongly correlated.

6.1. Methods

6.1.1. Sample

A total of N = 900 subjects completed the study reported in Tappin et al. (2017). As in the original study, we excluded N = 1 (0.11%) subjects for duplicate responding; leaving N = 899 to match with the CRT performance data. The CRT data was aggregated over 11 studies conducted using Amazon's Mechanical Turk (MTurk) by the senior author's lab between 2012 and 2017 (for the studies, see Stagnaro et al., 2018). These data comprised a total of N = 23,743 observed CRT sum scores. Of these 23,743 observations, using subjects' unique MTurk IDs, we identified N = 443 *unique* observations that could be matched to the study responses in Tappin et al. (2017). N = 443 was therefore our sample size for analysis.

6.1.2. Evidence Evaluation

In the study reported in Tappin et al. (2017), subjects first indicated (i) their preferred presidential candidate (dichotomous: Donald Trump or Hillary Clinton), and (ii) their prior belief about which candidate was more likely to win the 2016 presidential election. Prior beliefs were provided on a bipolar sliding scale from 0 (labelled "Hillary Clinton") to 100 ("Donald Trump"), sensitive to three decimal places. Subjects were instructed that dragging the slider closer to either name indicated they were more confident that that candidate would win the

election (subjects who responded with exactly 50—complete uncertainty—were unable to take part in the study). Subjects fell into one of two groups: Those whose prior belief was *consistent* with their candidate preference—for example, a preference for Trump to win, and the belief that he was more likely to win—and those for whom it was *inconsistent* (e.g., a preference for Trump to win, but the belief that Clinton was more likely to win). An equal number of each group was recruited—meaning that, in the aggregate sample of the study, prior belief and candidate preference were (approximately) uncorrelated.

Subjects were then randomly assigned to read a short paragraph about election polls that emphasized that *either* some polls predicted Trump was more likely to win, *or* some polls predicted Clinton was more likely to win the election. Random assignment was blocked on whether subjects initially believed either Trump or Clinton was more likely to win. This ensured that (roughly) equal numbers of subjects received a polling treatment concordant vs. discordant with their prior belief (and preference). In addition, there was no deception: At the time of data collection, the candidates were within several percentage points of each other and some polls had one or the other candidate ahead (verbatim study instructions are available in Tappin et al., 2017, Supplemental Material). On the same survey page as the polling information, subjects were asked: "In general, do you think polling data is informative?" They provided their response on a 1-7 scale, anchored 1 = Not at all, 7 = Very much so. This constitutes the measure of evidence evaluation for the analysis we conduct here.

6.1.3. Prior Belief and Political Identity Variables

In our sample of N = 443, we first determined whether the polling treatment subjects received was pro-Clinton or pro-Trump. A total of N = 211 (47.63%) subjects received the treatment that emphasized polls reporting that Clinton was more likely to win; for the remaining subjects, the treatment emphasized polls reporting that Trump was more likely to win (variable: poll treatment). Second, we determined whether subjects initially believed either that Clinton was more likely to win (indicated by < 50 on the prior belief scale, N = 329, 74.27%), or that Trump was more likely to win (> 50 on the prior belief scale, N = 114, 25.73%) (variable: prior belief). Third, we computed the strength of subjects' prior belief as the distance from the scale midpoint (50) (variable: prior belief strength). These values thus ranged from 0.237 (minimum strength prior) to 50 (maximum strength prior) (M = 30.15, SD = 16.03). Finally, we determined which political candidate subjects preferred to win (as a proxy for their political identity) (variable:

preference): N = 201 (45.37%) preferred Clinton to win; the remaining subjects preferred Trump to win. The sample sizes for each combination of preferred candidate, prior belief, and treatment assignment are displayed in Table 10.

Due to the design used in Tappin et al. (2017), receiving evidence—that is, a poll treatment—*concordant* with prior belief was only weakly (and non-significantly) correlated with receiving evidence concordant with candidate preference, r (441) = .08, p =.098, 95% CI [-.01, .17].

Table 10. Group Sample Sizes in Study 5.

Preference	Prior belief	Poll treatment	N
Clinton	Clinton	Pro-Clinton	77
		Pro-Trump	86
	Trump	Pro-Clinton	17
		Pro-Trump	21
Trump	Clinton	Pro-Clinton	78
		Pro-Trump	88
	Trump	Pro-Clinton	39
		Pro-Trump	37

6.1.4. CRT Performance

The CRT performance data came from studies which administered the original 3-item CRT (Frederick, 2005), not the combined 7-item CRT used here in Studies 1-4. Thus, in Study 5, CRT sum scores ranged from 0 to 3 (correct responses were summed as usual). For subjects with repeated CRT measurements in the Stagnaro et al. (2018) dataset, we computed their *mean* CRT sum score over the repeated measurements. The mean CRT sum score over the N = 443 sample was 1.22 (SD = 1.13). The distribution of mean CRT sum scores over the sample is reported in the SM (we note there is strong positive skew, consistent with Kahan, 2013).

6.2. Results

6.2.1. Prior Beliefs Analysis

We first fitted a linear regression model with four IVs: (i) poll treatment assignment, (ii) prior belief, (iii) prior belief strength, and (iv) CRT performance. The DV was the measure of evidence evaluation (higher values = greater endorsement of evidence). The critical test is on the four-way interaction between (i), (ii), (iii) and (iv).

The four-way interaction was statistically significant and positive [b = 0.05, SE = 0.02, p =.007] (full model parameter estimates are reported in the SM, Table S8). The predicted values from this model are displayed in Figure 8. They indicate that, among subjects who believed Trump was more likely to win the election, those who scored higher on the CRT deferred more strongly to their prior beliefs in evidence evaluation (Figure 8B). Unexpectedly, however, also among these subjects, those who had *weak* prior beliefs and who scored higher on the CRT appeared to defer *less* to their prior beliefs; evaluating the evidence almost at parity across prior belief concordance (Figure 8B, left-most panel). In other words, in these data, Trump-believing subjects who scored higher on the CRT were more *sensitive* to how the evidence squared with their prior beliefs; rather than blindly deferring to their prior beliefs *per se*. In contrast, subjects who believed Clinton was more likely to win the election and who scored higher on the CRT were relatively less sensitive to their prior beliefs in evidence evaluation (Figure 8A). The four-way interaction remained of similar size and statistically significant in an ordered logistic regression model: χ^2 (1) = 8.68, p =.003, b = 0.06 (SE = 0.02).

A possible confound in this result concerns the asymmetric group sample sizes used in the analysis. In particular, among subjects who believed Trump was more likely to win, approximately two-thirds also *preferred* him to win (see Table 10). Since the critical four-way interaction with prior beliefs was mainly driven by Trump-believing subjects (Figure 8), this interaction may, in fact, reflect an interaction with candidate preference (rather than prior beliefs). To rule out this possibility, we balanced the group sample sizes used in the analysis. Specifically, we balanced the number of subjects who preferred Trump *and* believed Trump would win, with those who preferred Clinton but believed Trump would win. To do so, we drew N random subjects from the former group of subjects, where N was equal to the latter group of subjects (N = 38). Thus, both groups of subjects now contained N = 38. We then refitted the critical four-way interaction in a linear regression model (exactly as before).

We performed this random-sample-refitting procedure 10,000 times, and recorded the estimate, standard error, and p-value of the four-way interaction after each iteration. We also recorded the correlation between receiving evidence concordant with prior beliefs and receiving evidence concordant with candidate preference (as a check on the balancing procedure). After 10,000 iterations, the median correlation coefficient was -.01; indicating that prior beliefs and candidate preference were now (approximately) fully decoupled, as intended. The median estimate of the critical four-way interaction was 0.05; the median standard error was 0.02; and the median p-value was .017. In other words, the results of this procedure reproduced the results of the foregoing analysis that contained asymmetric group sample sizes.

6.2.2. Political Identity and Joint Analysis

We then proceeded to model the three-way interaction between (i) poll treatment assignment, (ii) subjects' candidate preference (as a proxy for their political identity) and (iii) CRT scores. The interaction was not statistically significant [b = -0.37, SE = 0.25, p =.133]. The full model parameter estimates are reported in Table S9 in the SM (model 1). The predicted values are displayed in Figure 9 (A).

Because of the positive—albeit weak—correlation between subjects receiving evidence concordant with prior beliefs and receiving evidence concordant with candidate preference (r = .08, reported in Methods), we also fitted a *joint* model—that is, estimating both target interactions together. In this joint model (parameter estimates reported in full in Table S9 in the SM, model 2), the three-way interaction between treatment, candidate preference, and CRT performance was statistically significant—but in the *opposite* direction to that expected under the IPC facilitation hypothesis [b = -0.71, SE = 0.26, p =.006]. That is, subjects who scored higher on the CRT deferred relatively *less* to their candidate preference in evidence evaluation after modelling prior beliefs. The predicted values from this model are displayed in Figure 9 (B). In contrast to these results, in the joint model, the four-way interaction with prior beliefs remained of similar size, positive, and statistically significant [b = 0.06, SE = 0.02, p <.001].

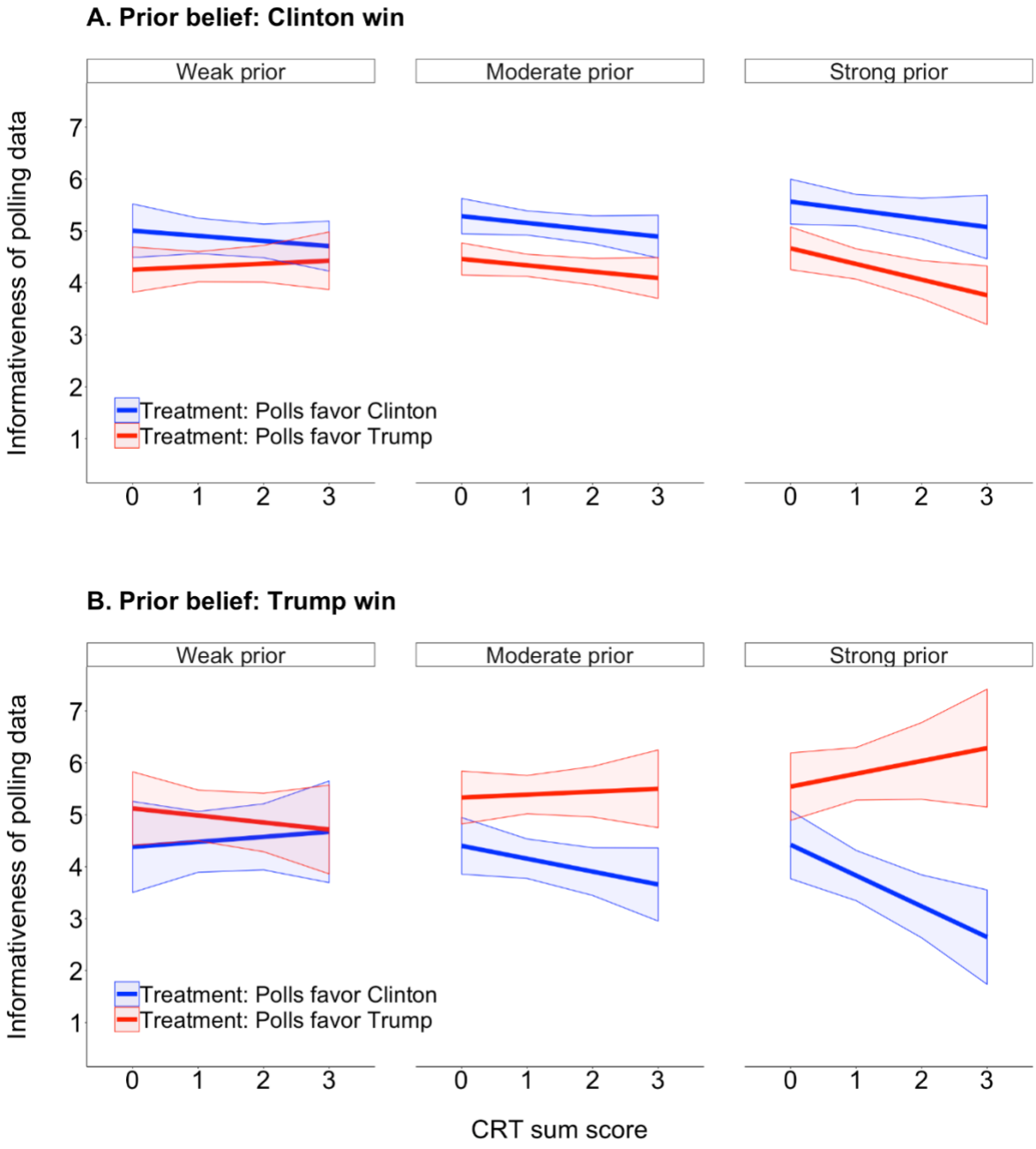


Figure 8. Predicted judgments on the informativeness of polling data in Study 5. ***A****, subjects who believed Clinton was more likely to win the election (than Trump).* ***B****, subjects who believed Trump was more likely to win the election (than Clinton). Weak, Moderate, and Strong priors correspond to values of 15, 30, and 45 on the prior belief strength variable, respectively (range: 0.237-50.000). Predicted values are obtained from the linear regression model (2) reported in Table S8 in the SM. Shaded regions are 95% CI. CRT = Cognitive Reflection Test.*

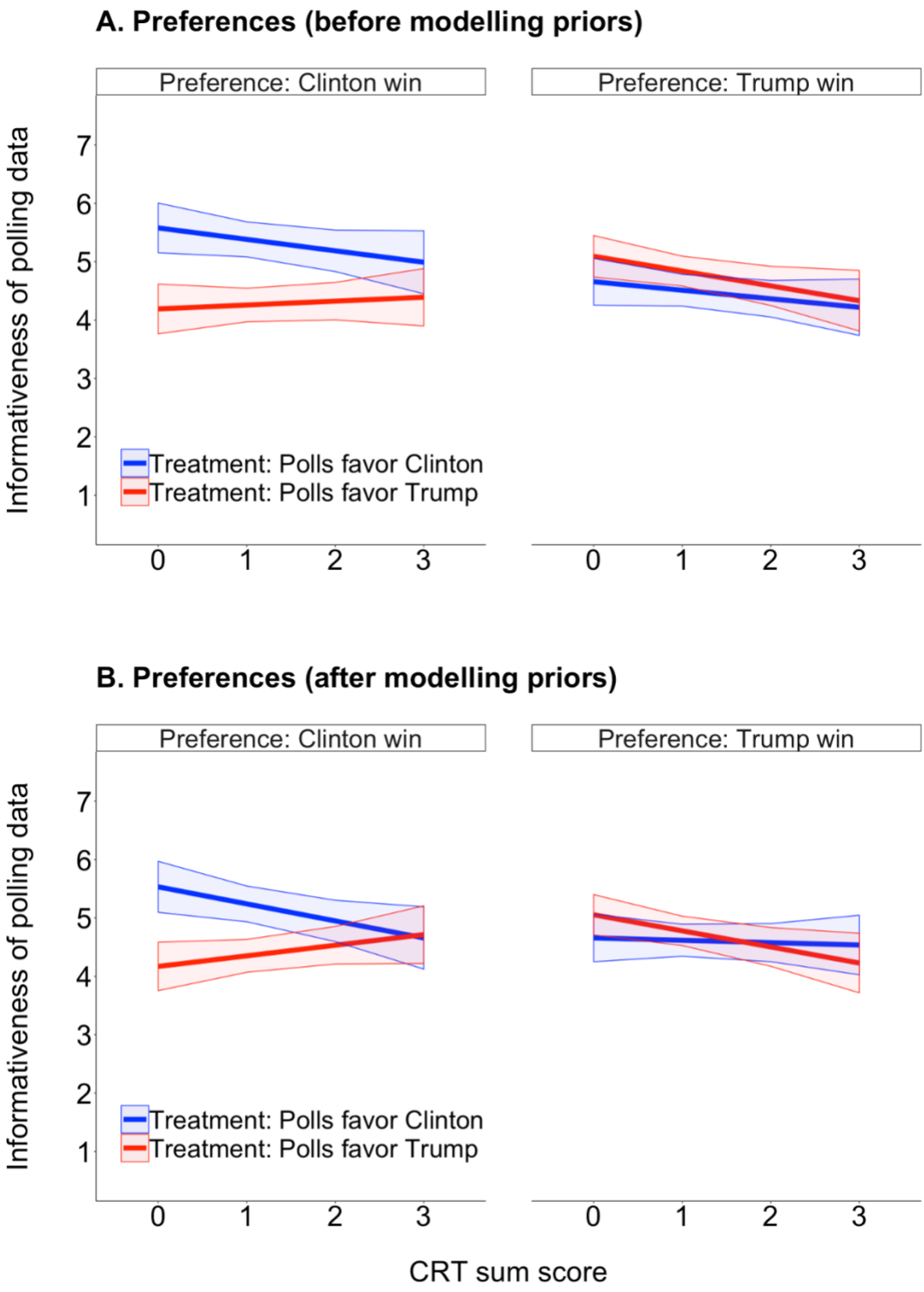


Figure 9. Predicted judgments on the informativeness of polling data in Study 5. A*, preferences in the model without any prior belief variables.* ***B****, preferences in the model with all prior belief variables. Predicted values in panel A are obtained from linear regression*

model (1), and in panel B from model (2), both reported in Table S9 in the SM. Shaded regions are 95% CI. CRT = Cognitive Reflection Test.

6.3. Discussion

The main result from Study 5 is that we again found evidence that subjects who scored higher on the CRT deferred more to their prior beliefs when reasoning about the validity of evidence; driven primarily by subjects who believed Trump was more likely to win the election. This result is notable, for two reasons. First, as a function of the design used in Tappin et al. (2017), prior beliefs and candidate preference were only weakly correlated in the sample we analyzed. That we nevertheless observed an interaction between CRT performance and prior beliefs meshes well with other evidence that suggests this interaction exists independent of political identity (e.g., Trippas et al., 2015) (insofar as candidate preference is a proxy for political identity). Furthermore, the interaction between CRT performance and prior beliefs was evident even after adjusting for the residual (weak) correlation between prior beliefs and candidate preference (via the random-sample-refitting procedure, and joint modelling of the two target interactions). The second reason this result is notable is the circumstances of data collection: We retroactively matched CRT data collected over a 5-year period—under various study designs and research aims—to subject responses in a different study again, conducted during the 2016 US presidential election. The broad consistency of the result with that observed in the evidence evaluation results of Study 2, 3, and 4 is a testament to its robustness.

That said, there were some differences from the results we previously observed. In particular, while (Trump-believing) subjects who scored higher on the CRT and had strong prior beliefs appeared to defer more to those beliefs when reasoning about the evidence, we also observed that—among subjects with *weak* prior beliefs—those who scored higher on the CRT seemed to defer somewhat *less* to their prior beliefs (a trend in evidence for both Trump-believing *and* Clinton-believing subjects; Figure 9A and 9B, left-most panels). In other words, the results of Study 5 suggested that CRT performance correlated with greater *sensitivity* to how evidence squared with prior beliefs; rather than blind deference to those beliefs. While unexpected, this pattern seems generally consistent with the exploratory results of Study 2 regarding prior beliefs and evidence evaluation (Figure 5B). There, the association between CRT performance

and deference to prior beliefs increased in size in conjunction with the strength of those beliefs (i.e., through weak-moderate-strong prior beliefs).

A secondary result from Study 5 is that subjects who scored higher on the CRT appeared to defer *less* to their political identity in evidence evaluation; a pattern opposite to that expected under the IPC facilitation hypothesis. However, this result is based on a *proxy* for political identity—preference for US presidential candidate—which was measured dichotomously rather than continuously. These factors limit the sensitivity of the test.

General Discussion

The influence of political identity on the formation of factual beliefs is advocated by several overlapping theoretical accounts (Bermúdez, 2018; Kahan, 2016a; Van Bavel & Pereira, 2018). The central proposition common to these accounts is that political identity biases cognition toward the formation of politically concordant factual beliefs, and away from politically discordant factual beliefs. We referred to this biasing influence as identity-protective cognition (IPC) (cf. Kahan, 2017). An alarming hypothesis derived from the logic of IPC is that cognitive sophistication *facilitates* identity-protective processing in political belief formation (Kahan, 2013; Kahan et al., 2017). By extension, cognitive sophistication may counterintuitively increase bias in this domain; polarizing—rather than unifying—the beliefs of people who identify with opposing political groups. We referred to this as the "IPC facilitation hypothesis", and noted its rather ominous implications for the prospect of achieving convergence on true beliefs in politics. Chiefly, that the distinct proficiencies of cognitively sophisticated partisans will be deployed to disregard and resist evidence that threatens their political identities.

In the current paper, we investigated this hypothesis, focusing on two distinct processes involved in political belief formation: (1) *belief updating*—the integration of new information with prior beliefs—and (2) *reasoning*—evaluations about the validity or accuracy of new information. Our key results are twofold. First, with regard to *belief updating*, when benchmarked against a normative Bayesian agent we found that cognitively sophisticated subjects appeared to be less—not more—biased in their belief updating after receipt of political information. Second, with regard to reasoning, we found evidence to suggest that more

cognitively sophisticated subjects deferred more to their *prior beliefs*—rather than to their political identities *per se*—when reasoning about the validity of political information.

In particular, we investigated *belief updating* in Studies 1 and 2, and found that subjects who scored higher on the Cognitive Reflection Test (CRT) – a behavioral measure of reflective/analytic thinking – deviated less from the posterior beliefs of a Bayesian agent overall. That is, more analytic subjects combined their prior beliefs with new politically-relevant information in a more normative (i.e., less biased) manner. In contrast, we observed scant evidence that higher CRT scores were associated with *more* biased belief updating conditional on political identity, as would be consistent with the IPC facilitation hypothesis. Indeed, in most cases, subjects who scored higher on the CRT tended towards the posterior beliefs of the Bayesian agent—irrespective of the political identity concordance of the evidence they received.

We then investigated *reasoning* about information in Studies 2 through 5, and found that subjects who scored higher on the CRT were more sensitive to their prior beliefs when *reasoning* about the validity of evidence that was concordant vs. discordant with those beliefs. Often, this manifested in these subjects deferring *more* to their prior beliefs in evidence evaluation, but not always. For example, in Studies 2 and 5, the *strength* of subjects' prior beliefs seemed to moderate the extent to which CRT performance was associated with deference to those beliefs. Concretely, when prior beliefs were weak, there was some evidence that high CRT scorers deferred to their prior beliefs to the same extent (e.g., Study 2, Figure 5, left panel B) or less (Study 5, Figure 8A and 8B, left-most panels) than low CRT scorers. When prior beliefs were strong, in contrast, we found that high CRT scorers generally deferred more strongly to these beliefs when reasoning about the information (and never less than low CRT scorers). We note, however, that not all subjects exhibited this pattern: across studies a common observation was that *either* one group of prior believers *or* the opposite group displayed the hypothesized interaction between CRT performance and prior beliefs—but rarely both.

The data were less consistent with the results expected under the IPC facilitation hypothesis: Namely, that high CRT scorers would defer more strongly to their *political identities* when reasoning about the validity of the new information. In particular, whereas in some cases we observed evidence consistent with this hypothesis (Study 2, Figure 5A; and Study 4, Figure 7, left panel A), in other cases we did not (Study 3, Figure 6A; and Study 5, Figure 9). Perhaps

more importantly, though, across all studies the interaction between CRT performance and political identity in predicting evidence evaluations never survived the modelling of prior beliefs. In other words, we observed no evidence that cognitively sophisticated subjects deferred more strongly to their political identities *per se* when reasoning about evidence concordant vs. discordant with those identities. Overall, therefore, while our results do not challenge the existence of identity-protective cognition *in general*, they challenge the notion that cognitive sophistication *facilitates* IPC. Table 11 displays a summary of the studies and their results as they relate to the relevant hypotheses.

7.1. Theoretical Implications

The theoretical implications of our investigation are twofold. First, the result that subjects who scored higher on the CRT deviated less from Bayesian updating (Studies 1 and 2) is *in*consistent with the notion that individuals update in a fashion that is more *biased* in favor of their political identities; where bias is represented explicitly as deviation from a formal normative or optimality criterion, in this case Bayesian rationality (cf. Hahn & Harris, 2014). While IPC is generally conceived of as biasing belief formation (Bermúdez, 2018; Kahan, 2016a; Van Bavel & Pereira, 2018), we find that *belief updating* is more normative (i.e., less biased) among subjects who score higher on the CRT. This result appears somewhat at odds with the notion that cognitive sophistication facilitates identity-protective cognition. Of course, there are numerous notions of "bias" (e.g., reviewed in Hahn & Harris, 2014); not all of which relate to (or imply) deviation from Bayesian rationality. For example, the results of Studies 1 and 2 are silent over the question of whether cognitive sophistication facilitates bias where bias means "double standards" (or hypocrisy), as it may be understood by the average person (Ditto et al., 2018b).

Furthermore, it is important to note that different models of IPC make different predictions about which cognitive processes are biased by political identity; and presumably, therefore, which biases may be exacerbated by cognitive sophistication. For example, Kahan's (2016a) model of IPC focuses exclusively on the process of *reasoning* about the validity of new evidence—and explicitly *not* on the process of integrating that evidence with prior beliefs (i.e., belief updating). Other models assume a much broader range of cognitive processes are influenced by IPC (e.g., Van Bavel & Pereira, 2018). With respect to the latter models, our results suggest that belief updating is one cognitive process where cognitive sophistication

seemingly does not facilitate bias in favor of political identity; on the contrary, evaluated against a Bayesian agent, it is associated with *less* bias.

The second implication of our investigation is based on the result that subjects who scored higher on the CRT deferred more strongly to their prior beliefs when reasoning about the validity of new information (Studies 2 through 5). In particular, our results suggest that past evidence taken as corroborative of the IPC facilitation hypothesis in the context of reasoning—whereby cognitive sophistication putatively facilitates deference to *political identity* when individuals reason about information (e.g., Kahan, 2013)—is confounded by deference to prior beliefs. That is, our results highlight that cognitively sophisticated individuals may not, in fact, defer more to their political identities *per se* in these cases; but, rather, they defer more to their prior beliefs. The distinction is subtle—it appears we are merely trading one "bias" for another—but, theoretically speaking, it is crucial. It is crucial because models of identity-protective cognition posit a *particular* theoretical mechanism; specifically, an *identity*-driven bias in reasoning, and not bias driven by the prior beliefs of the reasoner. To facilitate clear exposition of this implication of our results, in Figure 10 we reproduce the IPC model from Kahan (2016a) (with some additions, which are explained below), and map out precisely how our results relate to this model.

Table 11. Summary of Studies and Corresponding Results.

Study	Cognitive Process	Hypothesis	Empirical Prediction	Result
#1	Belief Updating	IPC Facilitation	High CRT_p = Greater deviation from Bayesian posterior beliefs conditional on political identity	-
		Alternative	High CRT_p = Lesser deviation from Bayesian posterior beliefs	+
#2	Belief Updating	IPC Facilitation	High CRT_p = Greater deviation from Bayesian posterior beliefs conditional on political identity	-
		Alternative	High CRT_p = Lesser deviation from Bayesian posterior beliefs	+
	Reasoning	IPC Facilitation	High CRT_p = Evaluation of evidence more strongly conditional on political identity	+/-
		Alternative (exploratory)	High CRT_p = Evaluation of evidence more strongly conditional on prior beliefs	+
#3	Reasoning	IPC Facilitation	High CRT_p = Evaluation of evidence more strongly conditional on political identity	-
		Alternative	High CRT_p = Evaluation of evidence more strongly conditional on prior beliefs	+
#4	Reasoning	IPC Facilitation	High CRT_p = Evaluation of evidence more strongly conditional on political identity	+/-
		Alternative	High CRT_p = Evaluation of evidence more strongly conditional on prior beliefs	+
#5	Reasoning	IPC Facilitation	High CRT_p = Evaluation of evidence more strongly conditional on candidate preference	-*
		Alternative	High CRT_p = Evaluation of evidence more strongly conditional on prior beliefs	+

*Note. The negative sign (-) indicates that the prediction was inconsistent with the data; the positive sign (+) indicates that the prediction was consistent with the data. Sign (+/-) indicates that the prediction was initially consistent with the data, but did not remain so after including the prior belief variables. * indicates that this particular result does not bear strongly on the IPC Facilitation Hypothesis (see Study 5 discussion). CRT_p = Cognitive Reflection Test performance.*

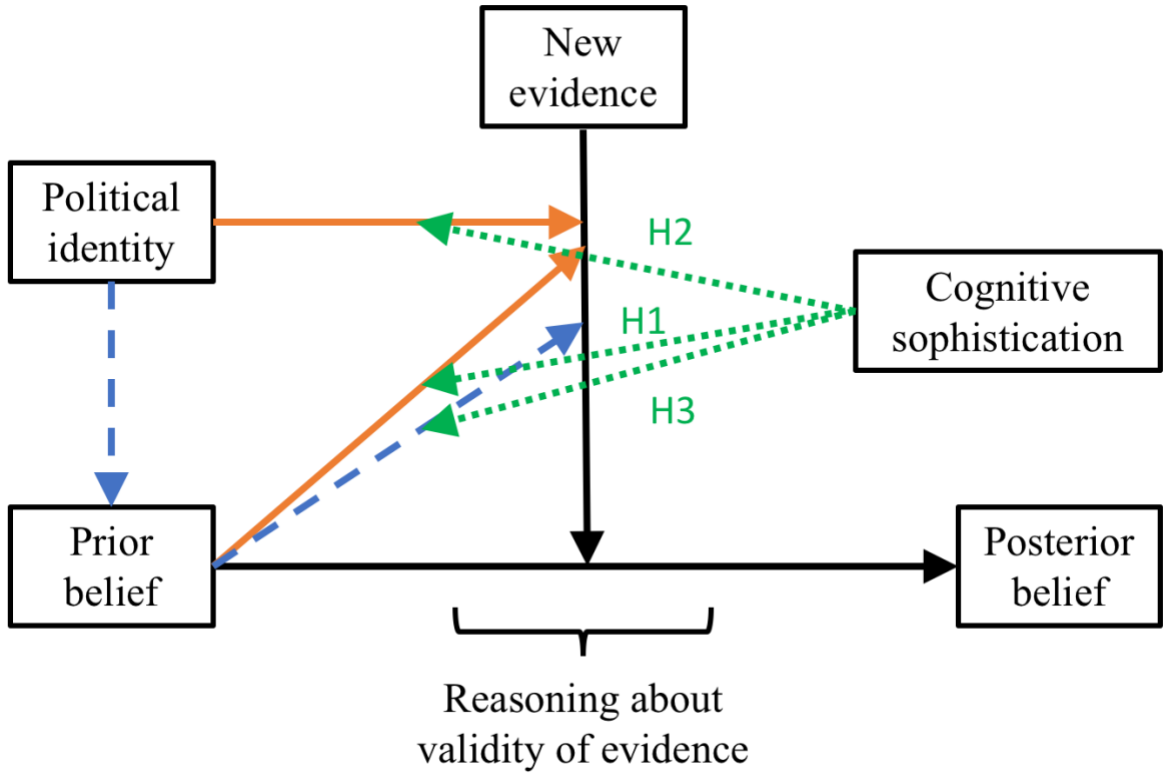


Figure 10. Model of (Political) Identity-Protective Reasoning. *The model is adapted from Kahan (2016a, in which it is referred to as "Politically Motivated Reasoning"). The model shows how political identity, prior beliefs and cognitive sophistication theoretically relate to reasoning about the validity of new evidence. Solid orange edges denote direct effects; dotted blue edges denote the indirect effect of political identity; and dotted green edges denote the facilitating effect of cognitive sophistication. H1 represents the hypothesis that cognitive sophistication facilitates a direct effect of prior beliefs on reasoning; H2 represents the hypothesis that cognitive sophistication facilitates a direct effect of political identity on reasoning; and H3 represents the hypothesis that cognitive sophistication facilitates an indirect effect of political identity on reasoning.*

As described in Kahan (2016a), the model depicted in Figure 10 shows how political identity and prior beliefs theoretically relate to reasoning about the validity of new evidence. In particular, that political identity exerts (i) a direct effect on such reasoning (solid orange edge from political identity → reasoning), and (ii) an indirect effect on such reasoning, via prior

beliefs (dotted blue edge from political identity → prior belief → reasoning). Our first addition to this model is a direct effect of prior beliefs on reasoning (solid orange edge from prior belief → reasoning); implying that people's prior beliefs can influence their reasoning independently of political identity. The inclusion of this effect is well supported by evidence suggesting that people's prior beliefs affect their reasoning in distinctly non-political domains; for example, the numerous "belief bias" experiments discussed in the introduction (e.g., Evans et al., 1983; Klauer et al., 2000; Markovits & Nantel, 1989). For simplicity, we exclude the possibility that prior beliefs may also influence political identity—and, thus, reasoning indirectly[35]. Our second addition to the model is the putative facilitation effect of cognitive sophistication (dotted green edges from cognitive sophistication → …, labelled H1, H2, and H3). Accordingly, H1 represents the hypothesis that cognitive sophistication facilitates a direct effect of prior beliefs on reasoning; H2 represents the hypothesis that cognitive sophistication facilitates a direct effect of political identity on reasoning; and, finally, H3 represents the hypothesis that cognitive sophistication facilitates an indirect of political identity (via prior beliefs) on reasoning. As is clear, both H2 and H3 represent forms of the IPC facilitation hypothesis.

The results from Studies 2 through 5 (concerning reasoning) are inconsistent with H2, for the reason that we observed no evidence that the interaction between CRT and political identity survived the modelling of the counterpart interaction between CRT and prior beliefs. In other words, there was no evidence that subjects who scored higher on the CRT conditioned more on their political identities *per se*—that is, independent of conditioning on their prior beliefs—when evaluating new information. To put it yet another way: assuming there is a direct effect of political identity on reasoning—that is, independent of any effect of prior beliefs—we found no evidence that such an effect is facilitated by analytic thinking. This result cannot be explained by contesting the assumption of a direct effect of political identity (*per se*); since, in all the relevant studies, we observed evidence *consistent* with such a direct effect. Specifically, in all of the models where political identity and prior beliefs were entered simultaneously as predictors, the coefficient on the [treatment x political identity] term was statistically

[35] While it is almost certainly the case that prior beliefs influence political identity to some extent, this directional effect would seem difficult to distinguish from the reverse directional relationship (i.e., political identity → prior beliefs). Furthermore, the model in Kahan (2016a) is silent regarding the influence of prior beliefs on political identity; thus, we ignore it here to focus solely on how our results relate to that model (but we note that future modelling work should recognize the distinction).

significant, large, and in the direction predicted by IPC (see Tables S6, S7 and S9 in the SM)[36]. This underscores our earlier point that our results do not challenge identity-protective cognition *in general*, but, rather, the notion that cognitive sophistication *facilitates* IPC. Given that influential studies bearing on the IPC facilitation hypothesis did not measure, model or otherwise account for prior beliefs (e.g., Kahan, 2013; Kahan et al., 2017), it is reasonable to conclude that there is currently no compelling evidence that cognitive sophistication facilitates a direct effect of political identity on reasoning about the validity of new information.

This leaves H1 vs. H3, and the question of which is more consistent with our results. Recall that H3 assumes that political identity causes prior beliefs, which, in turn, influence reasoning (Kahan, 2016a). Thus, H3 implies that what looks like stronger conditioning on prior beliefs among high CRT scorers is, in fact, stronger conditioning on political identity. Or, put another way, analytic thinking facilitates the indirect effect of political identity → prior beliefs → reasoning. In contrast, H1 implies that stronger conditioning on prior beliefs among high CRT scorers is independent of political identity. Of course, it is reasonable to assume that the prior beliefs subjects "bring with them" into the experiment are *to some extent* dependent on their political identities; for example, because of partisan selective exposure to news media (Bolin & Hamilton, 2018; Rodriguez et al., 2017), or because people tend to learn from others who are politically *like them* (Kahan, 2015). Insofar as prior beliefs are not totally determined by political identity, however, it is clear that H1 and H3 are not mutually exclusive. Given this, the question is one of degree: that is, is there evidence—here or elsewhere—to suggest that H1 *can* (to any extent) account for our results? We consider the following evidence in favor of H1:

- In Studies 3, 4 and 5, the interaction between CRT and prior beliefs in predicting evidence evaluations remained of similar size (and statistically significant) after modelling the counterpart interaction between CRT and political identity. This provides some evidence that subjects who scored higher on the CRT conditioned on their prior beliefs independent of their political identity. However, due to imperfect measurement of the relevant constructs—and the associated type I error inflation that results (Westfall & Yarkoni, 2016)—this is weak evidence in favor of H1.

[36] The same is true of the [treatment x prior beliefs] terms, consistent with a direct effect of both political identity and prior beliefs on reasoning about the validity of new information.

- In Study 5, as a function of study design, receiving information concordant with prior beliefs was only trivially correlated with receiving information concordant with candidate preference (a proxy for political identity). In other words, for approximately half of the subjects, if the information was concordant with their prior belief it was explicitly discordant with their candidate preference. If analytic thinking only facilitates an indirect effect of political identity (via prior beliefs), we ought to observe limited—or no—conditioning on prior beliefs in this case; because prior beliefs and candidate preference are decoupled in the aggregate sample of the study. Despite this decoupling, we nevertheless observed evidence that subjects who scored higher on the CRT conditioned more strongly on their prior beliefs in evidence evaluation (e.g., see Figure 8B).
- Across studies, the interaction between CRT and prior beliefs in predicting reasoning about evidence appeared more robust than the corresponding interaction between CRT and political identity. From a purely pragmatic perspective, this suggests that increased "bias" among the cognitively sophisticated may be more reliably diagnosed by assuming (and measuring) deference to prior beliefs, rather than political identity.
- Perhaps the strongest evidence in favor of H1 comes from recent studies conducted by Trippas and colleagues (2015, 2018). As described in the introduction, these authors investigated the association between cognitive ability, CRT scores and performance on belief bias tasks. They found that the reasoning performance of subjects with greater cognitive ability—as well as those scoring higher on the CRT—while better overall, appeared more influenced by whether the evidence was concordant (vs. discordant) with their prior beliefs. Since all the belief bias tasks were explicitly non-political, their results imply support for H1: that cognitively sophisticated individuals' condition more strongly on their prior beliefs (per se) when reasoning.

Despite these points, of course, we are unable to satisfactorily discriminate between the relative importance of H1 and H3—given the non-experimental nature of the relevant data[37]. Nevertheless, on the basis of our investigation, it is ill-advised to dismiss the possibility that

[37] Some readers might wonder why we did not conduct mediation analysis. We did not conduct mediation analysis because (i) our data are cross-sectional, and (ii) we did not randomly assign political identity, prior beliefs or cognitive sophistication. The validity of the inferences that can be gleaned from mediation analysis is extremely limited in this case (Bullock et al., 2010; Stone-Romero & Rosopa, 2008).

evidence taken to corroborate H3—if not H2—is, in fact, amenable to explanation by H1 (e.g., Kahan, 2013).

Furthermore, there is a related and more general issue highlighted by the above discussion; one that, for simplicity of exposition, we have thus far skirted over. That is, the type of experimental design often used to infer identity-protective cognition (Kahan, 2016a; for a meta-analysis see Ditto et al., 2018a) prohibits causal inferences about the effect of political identity—or, for that matter, prior beliefs or cognitive sophistication (see also Kim, 2018). This follows directly from the fact that the critical inferential test in the design is a treatment (evidence) by covariate (identity) interaction (Gerber & Green, 2012). As the current investigation has highlighted, this allows for the random assignment of evidence to not only alter the concordance of the evidence with subjects' political identities—but, also, its concordance with their prior beliefs (and a range of correlates besides); an "empirical catch-22", as noted by Ditto and colleagues (2018a, p. 14; see also Ditto et al., 2018b). To convincingly isolate a causal effect of political identity in this case, it would be necessary to randomly assign *both* political identity *and* evidence concordance[38]. It seems unlikely that such a design is feasible.

Possible workarounds might include designs that equalize prior beliefs across subjects (for a review of such attempts in primarily apolitical domains, see Ditto, 2009), random assignment of political party *cues*, rather than political identity *per se* (e.g., Cohen, 2003), or random assignment of threat to (or affirmation of) political identity (e.g., Nyhan & Reifler, 2018). Unfortunately, the first workaround is extremely difficult in practice given that the relevant prior beliefs are likely to be numerous and embedded in a network of possibly interdependent beliefs (e.g., Brandt et al., 2018); all of which may be brought to bear on reasoning (Gershman, 2018). Regarding the second case (party cues), as noted by Ditto and colleagues (2018b) it may be entirely reasonable for people to rely on whether their political party endorse (or oppose) the information at hand. In fact, such reliance may *reflect* the role of prior beliefs—for example, beliefs about the trustworthiness of one's in-party elites—rather than ruling them out. The final case—threatening or affirming identity—strikes us as a promising design (in theory) to isolate causal effects of political identity on evidence evaluation. Though, in practice, recent attempts at this have met with mixed results (e.g., Nyhan & Reifler, 2018; but see Kim, 2018). Overall,

[38] Such a design would still prohibit the inference that cognitive sophistication causally affects reasoning about the evidence conditional on political identity, because cognitive sophistication is also not randomly assigned (e.g., see Gerber & Green, 2012, pp. 301-303).

the above discussion highlights the fundamental difficulty in isolating a causal effect of political identity—or indeed prior beliefs or cognitive sophistication—on reasoning about evidence (see also Ditto, 2009).

7.2. Implications for Convergence on True Beliefs in Politics

In Studies 1 and 2, we found that subjects who scored higher on the CRT tended to update their beliefs more normatively; in particular, deviating less from the posterior beliefs of a Bayesian agent. In Studies 2-5, we did not assess subjects' reasoning against a normative benchmark, but others have shown that deference to prior beliefs in evaluation of new evidence is consistent with Bayesian rationality (Koehler, 1993). Relatedly, it has been argued that such deference is rational and appropriate (Baron & Jost, 2018; Gerber & Green, 1999; Jern et al., 2014); in fact, that it is "essential for any organism to make sense of, and respond adaptively to, its environment" (Lord et al., 1979, p. 2107).

However, whether reasoning or belief updating are consistent with Bayesian rationality does not (by itself) mean that individuals will converge in their beliefs, that their beliefs will be more accurate (Cook & Lewandowsky, 2016; Jern et al., 2014; Kahan, 2016a) or that they will be less prone to undesirable double standards (Ditto et al., 2018b). On the one hand, the notion highlighted by our results—that cognitive sophistication may be deployed to assess new evidence in light of what the person currently believes to be true—strikes us as more optimistic than the alternative: That cognitive sophistication is deployed to disregard and resist identity-threatening evidence. Yet, practically speaking, insufficient independence between prior beliefs and reasoning about new evidence may be similarly problematic if the goal is to achieve convergence on true beliefs in politics (Kahan, 2016a). For example, insufficient independence may increase belief polarization (Taber et al., 2009), prevent correction of false beliefs (Rabin & Schrag, 1999), and otherwise violate desirable standards of judgment; for example, in various legal, political or scientific contexts (Ditto et al., 2018b; Koehler, 1993).

Ultimately, in a world where the precise reliability of evidence is often unknown and must be inferred, conditioning on prior beliefs when evaluating new evidence may be one of several imperfect strategies available to individuals: Recent simulations show that such a strategy can confer (limited) accuracy gains in belief formation given certain assumptions, but, also, that it may incur significant epistemic costs (Hahn et al., 2018). Regarding the current investigation,

whether and to what extent individuals who scored higher on the CRT deferred *too much* to their prior beliefs when evaluating new evidence—compromising the accuracy of their beliefs in the long run—is not straightforward to determine (e.g., see Hahn & Harris, 2014). Given this, we concur with Hahn and colleagues (2018) that it is of paramount importance to safeguard the integrity of the information environment to which individuals are exposed over the long-term. Insofar as this is achieved, deference to prior beliefs is less likely to compromise the accuracy of belief formation; since those prior beliefs are more likely to be based on accurate information in the first place.

7.3. Strength & Limitations

The primary strength of our investigation was our diverse approach. Specifically, we conducted five studies—comprising a range of designs, stimuli, analytic approaches, and dependent variables—that triangulated on a recurring empirical theme. While our approach was clearly not exhaustive, it is an improvement over investigations that rely on a single approach or experimental design to answer the research question (Munafò & Smith, 2018).

That said, we limited our operationalization of cognitive "sophistication" to performance on the Cognitive Reflection Test (CRT); a behavioral measure of reflective/analytic thinking (Frederick, 2005; Pennycook et al., 2016). This choice was guided primarily by past work that has investigated the relationship between identity-protective cognition and analytic thinking (Kahan, 2013). This previous work was central to the investigation we conducted here. Nevertheless, our reliance on this measure as an index of cognitive sophistication is a limitation of the current investigation insofar as the propensity to think analytically—and rational thinking more broadly—is distinct from general cognitive ability (Stanovich et al., 2016). Though CRT performance is reliably correlated with measures of the latter (Toplak et al., 2011), whether our findings generalize to established measures of cognitive sophistication—for example, a standardized IQ test—is clearly an important question to address.

The two foremost experiments consistent with the idea that cognitive sophistication facilitates identity-protective belief formation are reported by Kahan (2013) and Kahan et al. (2017). While our findings suggest the former experimental result is confounded by prior beliefs, the latter experiment used a different measure of cognitive sophistication than the CRT—i.e., a numeracy test—and a different design. Specifically, in that experiment, subjects were tasked

with finding the correct answer to a covariance-detection problem. Across the relevant experimental treatments, the correct answer to the problem was manipulated to be concordant or discordant with subjects' political identities. Importantly, as a function of the design of the experiment, the "heuristic" (most obvious) answer was always incorrect. When the heuristic answer was identity-*con*cordant, subjects high in numeracy only did slightly better in identifying the correct answer than subjects low in numeracy. In contrast, when the heuristic answer was identity-*dis*cordant, subjects high in numeracy did much better in identifying the correct response than subjects low in numeracy.

The key inference was that high numeracy subjects *selectively* engaged their superior numeracy to solve the problem: When the heuristic answer was concordant with political identity, there was less motivation to engage—whereas, when this answer was discordant with political identity, there was more motivation to engage (Kahan et al., 2017). Crucially, it was assumed that these subjects' motivation to engage was to protect their political identities; consistent with the logic of the IPC facilitation hypothesis.

However, this pattern of results bears close resemblance to that reported in the belief bias experiments of Trippas and colleagues (Trippas et al., 2015; Trippas et al., 2018). As described above, subjects higher in cognitive ability outperformed subjects lower in cognitive ability at discriminating whether a given conclusion logically followed from a set of premises. In particular, this superior discrimination among high ability subjects was greater when conclusions were *dis*cordant (vs. concordant) with prior beliefs. In other words, high ability subjects appeared to *reason better* in the face of evidence that was discordant (vs. concordant) with their prior beliefs[39]. This is qualitatively similar to the pattern observed by Kahan and colleagues (2017) in their experiment, but swapping out "political identity" for "prior beliefs". Indeed, as already mentioned, there was *no* leverage of political identity in the experiment of Trippas and colleagues (2015), which exposed subjects to stimuli such as "some animals are cats" (p. 435). Considering this similarity between experimental results, the identity-selective, or "motivated" reasoning of high ability subjects reported by Kahan and colleagues (2017) is perhaps also confounded by an effect of prior beliefs. This is only conjecture, of course, but resonates with the results of the current investigation.

[39] Perhaps because the mismatch induced cognitive conflict; the detection of which has been associated with individual differences in analytic thinking (Pennycook, Fugelsang, & Koehler, 2015b).

7.3. Conclusion

In this paper, we investigated the hypothesis that cognitive sophistication facilitates identity-protective processing in political belief formation. Our findings suggested that cognitively sophisticated individuals deferred more to their prior beliefs—rather than to their political identities *per se*—when reasoning about information in the political domain. Furthermore, benchmarked against a Bayesian agent, we found evidence that these individuals were overall less—not more—biased in their belief updating after receipt of such information. These results highlight a somewhat more optimistic perspective on the role of cognitive sophistication in political belief formation than prior work: That cognitive sophistication may be deployed to assess and integrate new evidence in light of what the person currently believes to be true, rather than to disregard and resist identity-threatening evidence per se. From a practical perspective, however, deference to prior beliefs may be similarly problematic for the prospect of achieving convergence on true beliefs in politics. One factor determining this assessment is the quality of the information environment (e.g., on social media, claims made by politicians). Where prior beliefs about political issues are constructed on the basis of misinformation and bad evidence, deference to prior beliefs will cement false beliefs. The upshot highlights the paramount importance of safeguarding the integrity of the information environment to which people are exposed over the long-term.

Supplemental Material

Political Statement Stimuli Selection (Studies 1 and 2)

As described in the main text, we selected political statement stimuli via a three-step pre-testing procedure. These stimuli were used in Study 1 and Study 2.

In step 1, we identified 53 political statements from fact-checking websites *politifact.com* and *factcheck.org*. We focused on two selection criteria in step 1: (i) statements had to be classified as unambiguously true or false by the fact-checking websites, and (ii) the pool of statements had to be distributed such that we had an approximately equal number that were true and false, and pro-Democratic and pro-Republican.

In step 2, we pre-tested the 53 political statements to obtain ratings over four features of each statement:

1. The likelihood that the statement was true (from 0-100 in whole integers, anchored from "certainly false" to "certainly true")
2. Assuming the statement was true, how favorable it would be for Democrats vs. Republicans (measured on a 1-5 scale, anchored "More favorable for Democrats", "Somewhat more favorable for Democrats", "Equally favorable for Democrats and Republicans", "Somewhat more favorable for Republicans", "More favorable for Republicans")
3. Assuming the statement was true, how favorable it would be for President Donald Trump (measured on a 1-5 scale, anchored "Very unfavorable", "Somewhat unfavorable", "Neither favorable nor unfavorable", "Somewhat favorable", "Very favorable")
4. Whether the statement was familiar; had participants seen or heard it before ("Yes", "Unsure", "No")

For the stimuli pre-test, we recruited N = 201 subjects from Amazon's Mechanical Turk to rate each statement according to the above features. In step 3, we selected a subset of 16 statements

from the pre-tested set of 53 statements. Our aims were threefold in this step of statement selection.

First, we aimed to select 8 true statements, and 8 false statements; half of which were pro-Democratic and half of which were pro-Republican (giving 4 categories of statement overall: False + pro-Republican; False + pro-Democratic; True + Republican; True + Democratic). Second, we aimed to select statements whose likelihood values were close to the midpoint of 50 (i.e., not *obviously* true or false), and whose partisanship values were close to the extremes of 1 or 5 (i.e., unambiguously partisan). Thirdly, we aimed to select statements such that subjects who identified with either the Democratic Party or Republican Party rated those statements that were politically *favorable* approximately similar to their political opponents on likelihood and extremity of partisanship. In other words, we aimed to ensure that supporters of one party did not receive politically favorable statements that were clearly more likely to be true or more likely to be false, or that were clearly more partisan, than supporters of the other party.

To select the 16 statements according to the aims above, subjects were first categorized as either Democrat-leaning or Republican-leaning based on a forced choice between the two parties (Democratic and Republican). Mean likelihood and partisanship scores were computed separately for subjects identifying as Democrat and Republican. Based on these scores, we then selected 16 statements bearing in mind our aims above. In Table S1, we display the mean score for each category (i.e., False + pro-Republican; False + pro-Democratic; True + Republican; True + Democratic) over the 16 statements on the key pre-test variables. The 16 statements themselves (subset by category) are listed below. After each statement listed below, we provide the link to the fact-checking webpage from which the statement was obtained.

False + Pro-Republican

1. CNN’s ratings have decreased due to their poor coverage of President Donald Trump. https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2017/08/trumps-phoenix-fiction//index.html_gzip
2. Expansion of Medicaid through "Obamacare" in Ohio left nearly 60,000 disabled citizens stuck on waiting lists for months. https://www.politifact.com/truth-o-

meter/statements/2017/jul/17/mike-pence/pence-falsely-ties-medicaid-expansion-disability-w/

3. Only 10 cents on every dollar from the Clinton Foundation goes to charitable causes. https://www.politifact.com/truth-o-meter/statements/2016/oct/04/mike-pence/pence-repeats-false-claim-clinton-foundations-limi/
4. President Donald Trump signed more bills through the legislature in his first 178 days in office than any president in history. https://www.politifact.com/wisconsin/statements/2017/jul/20/donald-trump/donald-trump-not-close-claiming-he-has-signed-more/

False + Pro-Democratic

1. During a 2017 summer event highlighting problems with "Obamacare", President Donald Trump ignored a disabled child who tried to shake his hand. https://www.factcheck.org/2017/08/trump-didnt-ignore-disabled-child/
2. By 2011, former President Barack Obama had only increased US debt by 16%, compared to his Republican predecessor George W. Bush who increased US debt by 115%. https://www.politifact.com/truth-o-meter/statements/2011/may/19/nancy-pelosi/nancy-pelosi-posts-questionable-chart-debt-accumul/
3. Democrat Nancy Pelosi was right when she claimed in 2012 that under "Obamacare" everybody in the US will have lower rates, better quality, and better access to health care. https://www.politifact.com/truth-o-meter/statements/2012/jul/06/nancy-pelosi/nancy-pelosi-says-everybody-will-get-more-and-pay-/
4. More jobs were created in the private sector during the first year of the Obama administration than during the previous eight years of his Republican predecessor George W. Bush. https://www.politifact.com/truth-o-meter/statements/2011/may/17/nancy-pelosi/nancy-pelosi-says-more-jobs-created-obamas-first-y/

True + Pro-Republican

1. During the first six months of President Donald Trump's administration, average weekly earnings for all private sector workers went up.

https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2017/08/trumps-phoenix-fiction//index.html_gzip

2. Under President Donald Trump's administration, unemployment has fallen to a 17-year low. https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2017/08/trumps-phoenix-fiction//index.html_gzip
3. Under the Trump administration's first six months in charge, over 1 million jobs were added to the economy. https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2017/08/trumps-phoenix-fiction//index.html_gzip
4. During President Donald Trump's first quarter in office, American exports of coal were up by almost 60% on the previous year. https://www.factcheck.org/2017/08/factchecking-trumps-west-virginia-rally/

True + Pro-Democratic

1. During former President Barack Obama's final 4 years in office, wages of the average American worker went up. https://www.factcheck.org/wp-content/cache/wp-rocket/www.factcheck.org/2017/08/trumps-phoenix-fiction//index.html_gzip
2. For approximately six years following the signing of "Obamacare" into law, American businesses created new jobs every single month - a new record at the time. https://www.politifact.com/truth-o-meter/statements/2016/jan/12/barack-obama/business-has-created-jobs-every-month-obamacare-be/
3. After six years of the Obama administration, US job growth was at its fastest pace since before the millennium. https://www.politifact.com/truth-o-meter/statements/2015/jan/21/barack-obama/barack-obama-says-us-economy-creating-jobs-fastest/
4. House speaker Republican Paul Ryan was wrong when he claimed in 2016 that Medicare is going broke because of "Obamacare." Medicare was in better shape because of Obamacare. https://www.politifact.com/wisconsin/statements/2016/dec/23/paul-ryan/repeal-and-replace-works-paul-ryan-says-obamacare-/

Table S1. Mean Likelihood, Partisanship, Trump favourability, and Familiarity Scores for the Four Categories of Political Statements Used in Studies 1 and 2 (16 Statements Total).

		Likelihood		Partisanship		Trump favourability		Familiarity
		Mean	Diff. from midpoint	Mean	Diff. from midpoint	Mean	Diff. from midpoint	Mean
Pro-Democratic	True	39.15	10.85	2.04	0.96	2.40	0.60	2.51
	False	59.57	9.57	1.95	1.05	2.13	0.87	2.41
Pro-Republican	True	40.68	9.32	4.02	1.02	4.05	1.05	2.53
	False	59.89	9.89	4.00	1.00	4.03	1.03	2.31

Note. Likelihood judgments were provided on a scale from 0-100; Partisanship was provided on a scale from 1-5 (1=more favorable for Democrats, 5=more favorable for Republicans); Trump favourability was provided on a scale from 1-5 (1=very unfavorable for Trump, 5=very favorable for Trump); Familiarity was provided on a scale 1-3 (1=Yes, 2=Unsure, 3=No).

Study 1

Methods

Sample

Subjects were paid $2 for taking part in Study 1. Subjects who completed the political statement pre-test were prevented from taking part in Study 1.

Belief Update Task

Subjects read instructions detailing the task, and were then asked two comprehension questions to ensure they understood that signals were accurate, on average, two out of three times. Subjects could not begin the task without correctly answering the two comprehension questions. During the task instructions, it was emphasized we were interested only in the personal opinion of subjects, and thus they were asked *not* to look up the truth or falsity of the statements online. The verbatim task instructions and comprehension questions are available on the OSF: https://osf.io/yt3kd/. On P1 and P2 trials, subjects had unlimited time to provide their likelihood judgments. Signals were presented for an enforced minimum of 5 seconds, at which point subjects were free to continue onto the next trial.

Comparison with Bayesian Agent

Before computing Bayesian posterior beliefs—as detailed in the main text—we recoded all *prior* beliefs (provided by subjects in P1) of 0 and 100 to 0.5 and 99.5, respectively. This is because probabilities of 0 and 1 prevent computation of Bayesian posterior beliefs. We did the same recoding for subjects' *posterior* beliefs (provided in P2). This recoding plan was preregistered prior to data collection. To illustrate the maximum raw magnitude of update we can expect from a Bayesian agent, assume a subject is totally uncertain whether a particular statement is true; that is, their prior belief is P = .5 (50%). They receive a signal that states that particular statement is TRUE. The Bayesian posterior belief that this statement is true $\approx$ .67 (67%).

Post-Task Measures

The 7 items of the combined CRT administered in Study 1-4 are reported below. The distribution of CRT sum scores (summed for each subject) in Study 1-5 are displayed in Figure S1. CRT items were presented one per page, and responses provided in open-ended format:

<u>From Shenhav et al. (2012)</u>

1. The ages of Mark and Adam add up to 28 years' total. Mark is 20 years older than Adam. How many years old is Adam?
2. If it takes 10 seconds for 10 printers to print out 10 pages of paper, how many seconds will it take for 50 printers to print out 50 pages of paper?
3. On a loaf of bread, there is a patch of mold. Every day, the patch doubles in size. If it takes 40 days for the patch to cover the entire loaf of bread, how many days would it take for the patch to cover half of the loaf of bread?

<u>From Thomson and Oppenheimer (2016)</u>

4. If you're running a race and you pass the person in second place, what place are you in?
5. A farmer had 15 sheep and all but 8 died. How many are left?
6. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?
7. How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?

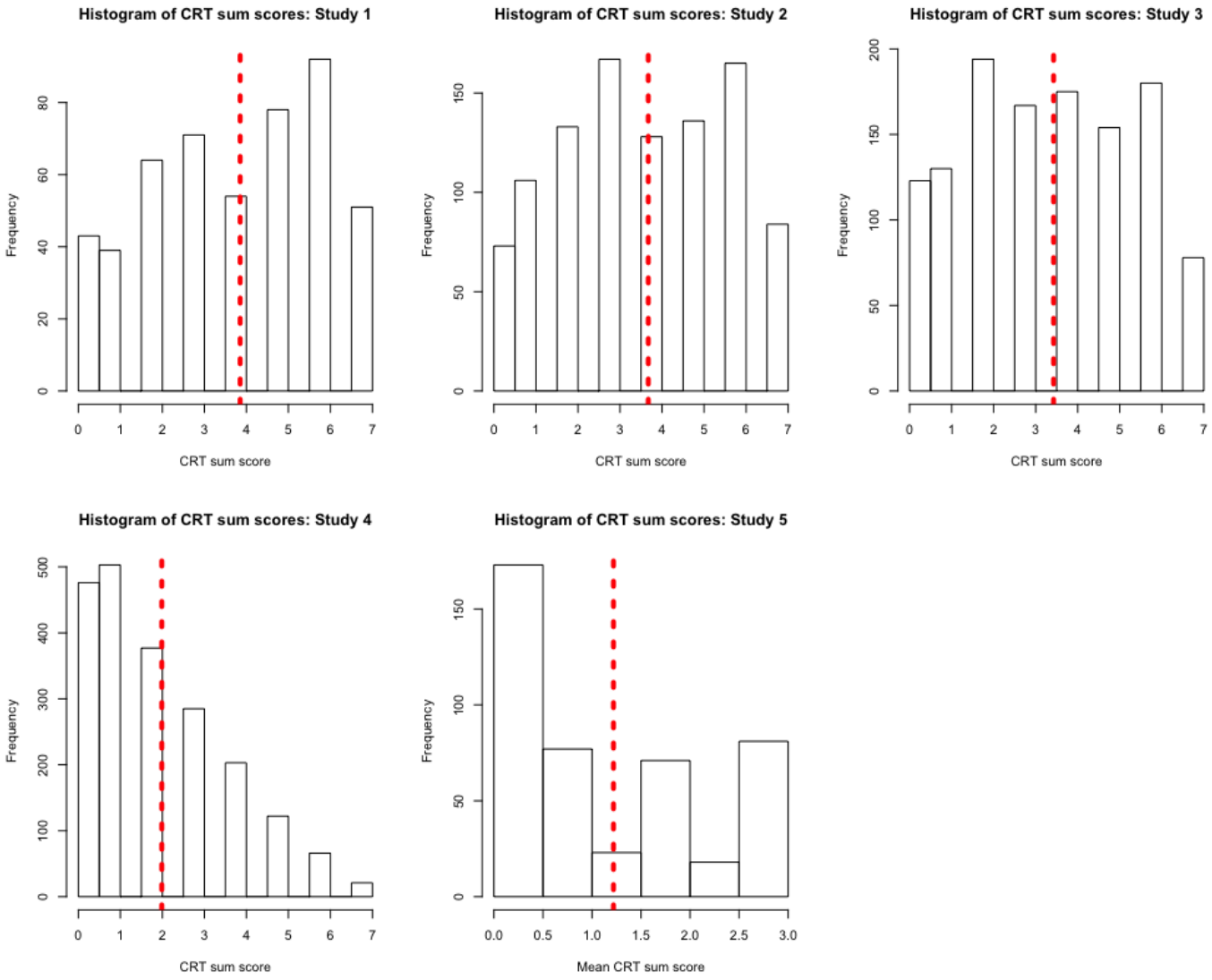


Figure S1. Distribution of CRT Sum Scores in Studies 1-5. *The dashed red line indicates the mean. CRT = Cognitive Reflection Test.*

Results

H1: High CRT Scorers Deviate Less from Bayesian Updating

Exploratory Tests

As described in the main text, we conducted a range of exploratory analyses as a check on the robustness of the preregistered test of H1.

Memory errors. Subjects completed a memory test immediately after the belief update task, where they were presented with each political statement again and were asked to respond whether they saw a TRUE or FALSE signal for that statement. We computed the proportion of memory errors for each subject by taking the mean of their incorrect responses over trials [correct = 0, incorrect = 1]. Subjects' proportion of memory errors were correlated with their CRT scores (analysis conducted at the level of subjects), τ = -.18, Z = -5.46, p <.001; individuals who scored higher on the CRT tended to make fewer memory errors. It is possible this difference—rather than differences in updating behavior *per se*—accounts for the relationship between CRT performance and absolute deviation from Bayesian posterior beliefs.

We thus conducted nonparametric partial correlations between (i) CRT scores and (ii) mean absolute deviation from Bayesian posterior beliefs, adjusting for (iii) proportion memory errors (using the *ppcor* package in R, Kim, 2015). The results of this analysis are reported in the main text.

Prior beliefs and regression to the mean. Regression to the mean (RTM) describes the phenomenon whereby more extreme measurements at Time 1 tend to approach the mean when measured again at Time 2. Translated here, more extreme prior beliefs (i.e., closer to the likelihood scale ends of 0% and 100%) may be associated with greater RTM measured in the posterior beliefs (Yu & Chen, 2015). If the extremity of prior beliefs differs systematically over CRT performance, differences in RTM—rather than differences in updating behavior *per se*—could account for the association between CRT performance and deviation from Bayesian posterior beliefs. We thus computed a variable indexing the *extremity* of prior beliefs, by calculating the distance between the prior belief and the likelihood scale midpoint (50%) on each trial. Prior belief extremity could thus range from 0 to 50, where 0 = the least extreme prior (i.e., 50% on the likelihood scale) and 50 = the most extreme prior (i.e., 0% or 100% on likelihood scale). We computed the mean extremity value for each subject over their 16 trials.

The correlation between mean extremity and CRT performance was small and missed the significance threshold, τ = -.02, Z = 0.63, p =.530, suggesting a trivial difference in prior belief extremity between subjects who scored higher vs. lower on the CRT. Indeed, conducting

nonparametric partial correlations between (i) CRT scores and (ii) mean absolute deviation from Bayesian posterior beliefs, adjusting for (iii) the extremity of prior beliefs, closely reproduced the results reported in the main text: $\tau = -.21$, $Z = -6.80$, $p < .001$ (difference index), and $\tau = -.07$, $Z = -2.18$, $p = .029$ (ratio index). These analyses were conducted using the *ppcor* package in R (Kim, 2015). Though this suggests that RTM is not responsible for the difference in updating between high and low CRT scoring subjects, we note that our design did not include a control group (the design gold standard for ensuring RTM does not confound repeated measurements).

Median absolute deviation. Computing a *mean* value for the absolute deviation from Bayesian posterior beliefs for each subject is arguably inappropriate. Specifically, because many subjects had non-normally distributed absolute deviation scores over their 16 trials. We thus repeated the preregistered nonparametric correlations at the level of subjects, but using the *median* absolute deviation from Bayesian posterior beliefs for each subject, rather than the mean. The results reproduced those of the preregistered test of H1: $\tau = -.16$, $Z = -4.94$, $p < .001$ (difference index), $\tau = -.08$, $Z = -2.55$, $p = .011$ (ratio index).

Linear mixed effects modeling. We anticipated large positive skew in the scores indexing subjects' absolute deviation from Bayesian posterior beliefs, and, thus, preregistered nonparametric tests as our primary tests of H1. As a further robustness check, however, we also fitted two exploratory linear mixed effects models to the data—one for each DV—at the trial-level. Fitting a maximal model on the difference index showed that CRT scores—converted to z-scores for fitting—were negatively associated with absolute deviation from Bayesian posterior beliefs, $b = -2.93$ ($SE = 0.54$), $t_{211.57} = -5.47$, $p < .001$. Mirroring the pattern of results reported in the main text, this association was smaller on the ratio index, $b = -0.04$ ($SE = 0.03$), $t_{277.22} = -1.31$, $p = .192$. Degrees of freedom and p-values were estimated using the Satterthwaite approximation via the *lmerTest* package in R (Kuznetsova et al., 2017).

Political concordance of evidence. A further question of interest is whether the negative correlation between CRT performance and absolute deviation from Bayesian posterior beliefs holds across the political concordance of the signals (evidence) subjects received. To explore this question, we computed *two* absolute deviation scores per subject; one mean computed over trials where the signals they received were politically *concordant*, and the other mean over trials where the signals were politically *discordant*. This variable was computed as a function

of the subject's political party preference (Democrats, Republicans), the partisanship of the statement (Democrat-favor, Republican-favor), and the evidence (signal) they received on that trial (see Table 2 in the main text). Two subjects did not report their political party preference, so these analyses are based on N = 490. The results of this analysis are reported in the main text alongside the preregistered test of H1.

H2: High CRT Scorers Deviate from Bayesian Updating Conditional on their Political Identities

Preregistered Tests

For both DVs, model fitting proceeded in the following steps:

1. We fitted random intercepts on participants and political statement stimuli.
2. We added separate fixed effects denoting CRT z-scores (zCRT) and political concordance of the evidence (i.e., the main effects).
3. We then added the key interaction term: zCRT x political concordance.
4. We statistically evaluated – by Likelihood Ratio Test (LRT) – the reduction in model deviance (i.e., improvement in model fit) between steps 2 and 3.
5. We then fitted correlated random slopes – a maximal model (Barr et al., 2013) – to the step 3 model.
6. Finally, we dropped the key fixed-effect interaction term from the step 5 model, and conducted an LRT comparison between this model and the step 5 model. This test evaluated reduction in model deviance contributed by the key fixed-effect interaction term in the presence of a maximal random effects structure, and, thus, constitutes the critical inferential LRT bearing on H2 (reported in the main text).

As reported in the main text, the random effects structure for both DV maximal models was slightly mis-specified in the preregistered protocol for the test of H2. In particular, we specified the [zCRT x political concordance] interaction as a random slope on subjects in the maximal model, but this is incorrect; only the within-subjects main effect—political concordance—should be specified as a random slope on subjects (Brauer & Curtin, 2017). However, for full transparency, we also fitted these incorrect models and reported the relevant results in the main

text. Parameter estimates and model fit indices from the correctly-specified models are reported in Table S2.

Table S2. Linear Mixed Effects Model Output in Study 1.

	Difference index			Ratio index		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
Fixed Effects						
(Intercept)	2.29	1.34	.087	0.14	0.05	.009
CRT	-0.76	0.89	.391	-0.05	0.04	.234
Political concordance	-3.51	0.93	<.001	-0.06	0.05	.202
CRT x Political concordance	1.74	0.88	.047	0.10	0.04	.011
Random Effects						
σ^2		782.953			1.366	
$\tau_{00,\ Subject}$		271.741			0.485	
$\tau_{00,\ Stimuli}$		16.462			0.025	
ρ_{01}		-0.755			-0.631	
$N_{Subject}$		490			490	
$N_{Stimuli}$		16			16	
$ICC_{Subject}$		0.254			0.258	
$ICC_{Stimuli}$		0.015			0.013	
Observations		7586			7586	
R^2 / Ω_0^2		.004 / .217			.002 / .239	
Deviance		73038.555			24978.641	

Note. The ratio index model is estimated via ML because REML did not converge (the difference index model is estimated via REML). P-values in the table are estimated via Wald test. CRT = Cognitive Reflection Test (z-score).

Exploratory Tests

As reported in the main text, to increase the sensitivity of the H2 test we also fitted models where political identity was represented continuously (i.e., not dichotomized into a preference for Democratic or Republican Party). To define political identity as a continuous variable, we referred to subjects' social and economic political views (each provided on scales from 1 = strongly liberal to 5 = strongly conservative) and political party affiliation (provided on a scale from 1 = strong Democrat to 7 = strong Republican). Social and economic views were combined by computing the mean across the two responses; we then standardized this combined variable and centered it at the scale midpoint (new variable: *conserv_ideology_z*). We also standardized and midpoint-centered the party affiliation variable (new variable: *party_z*). Finally, these two new variables were summed; creating a variable (*conserv_rep*) ranging from -3.48 to +3.48 where values greater than zero denote more conservative/Republican identity, and values less than zero denote more liberal/Democrat identity (M = -0.76, SD = 1.90).

We fitted two exploratory linear mixed effects models at the trial-level – one for each DV (difference index, ratio index) – with three IVs. The first IV was CRT performance (z-score). The second IV was evidence type i.e., whether the signal favored Democrats or Republicans, and was computed via combination of signal received (TRUE, FALSE) and the partisanship of the statement (Democrat-favor, Republican-favor) (see Table 2 in the main text). The third IV was the (continuous) political identity of subjects i.e., the new variable, *conserv_rep*. The focal exploratory test bearing on H2 is a three-way interaction between these variables.

We attempted to fit maximal models for both DVs, but these would not converge; the models were too complex for the data, given the extra variable. Thus, both models were estimated with random intercepts, as well as the random slope of evidence type on subjects and stimuli, and the random slope of CRT x evidence type x political identity—the critical three-way interaction—on stimuli. The models were fitted with restricted maximal likelihood, and degrees of freedom and p-values for the three-way interaction are estimated using the *lmerTest* package (Kuznetsova et al., 2017).

In the difference index model, the three-way interaction missed the significance threshold, $t_{108.30} = 1.51$, p =.134, b = 0.70 (SE = 0.46). In the ratio index model, the three-way interaction was statistically significant, $t_{114.70} = 2.10$, p =.038, b = 0.04 (SE = 0.02). The predicted values from these models are displayed in Figure S2. The predicted values are estimated with the

political identity variable set at -3 (i.e., strong liberal/Democrat) and +3 (strong conservative/Republican), respectively. Visual inspection of the values shows a pattern largely consistent with that produced by the models which treated political identity as a dichotomous variable (i.e., as a preference for Democrats or Republicans). Thus, we conclude that the preregistered use of a binary political identity variable (for the test of H2 reported in the main text) is not obscuring identity-protective deviation from Bayesian posterior beliefs among those with a high CRT score/strong political identity combination.

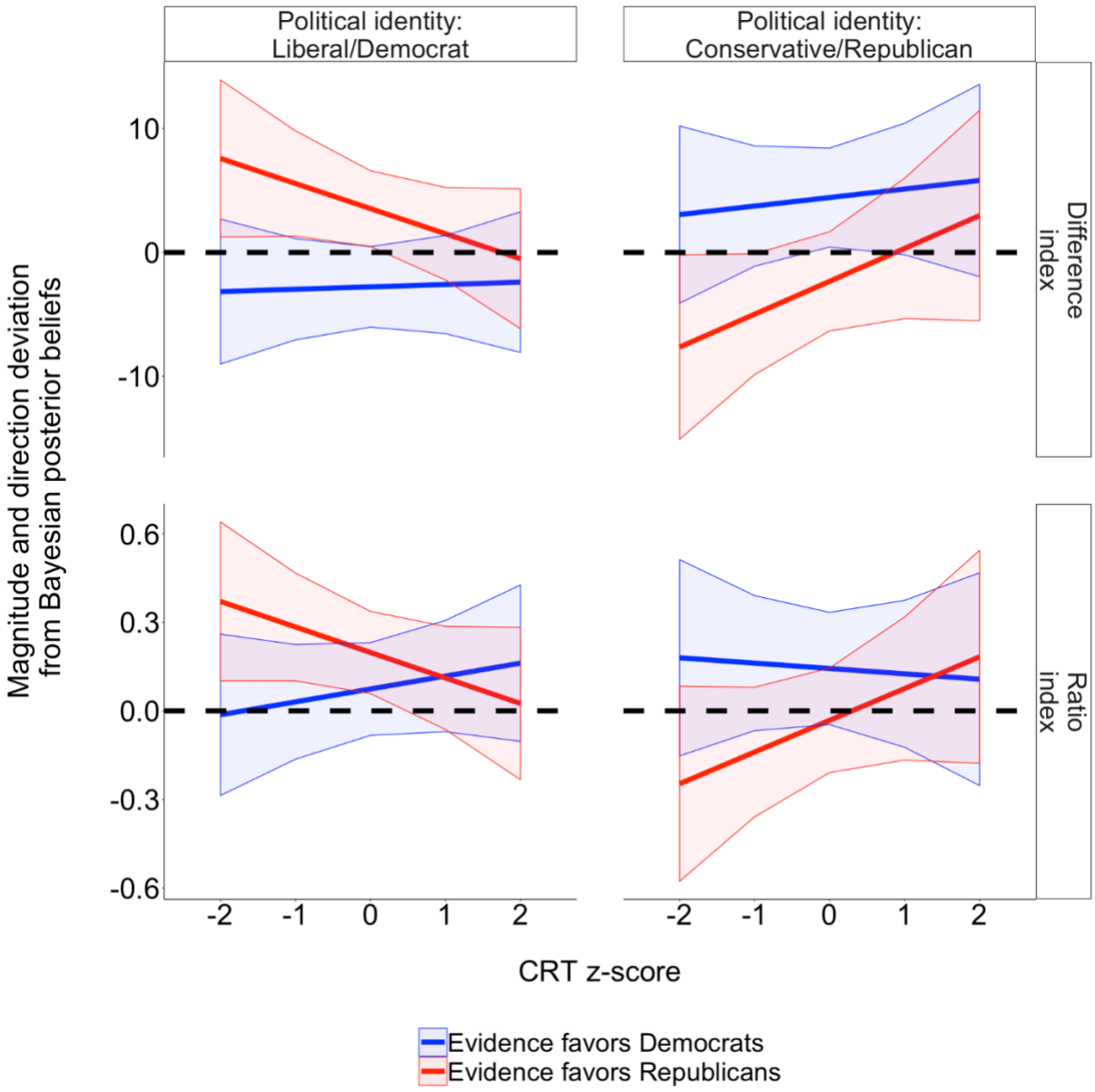


Figure S2. Predicted magnitude and direction deviation from Bayesian posterior beliefs as a function of CRT performance, evidence type, and political identity (Study 1). *Predicted values are estimated from the two models (one per DV) with political identity as a continuous variable, fitted as an exploratory test of H2 (Study 1). Political identity is set at -3 (strong liberal/Democrat) and 3 (strong conservative/Republican). The dashed line at y = 0 indicates Bayesian posterior beliefs: Relative to Bayesian, y-axis values greater than zero imply over-updating, and values less than zero imply under-updating.*

Distributions of Posterior Beliefs

In figure S3, we plot the distributions of posterior beliefs as a function of the political statement stimuli (panels 1-16) and signal received (TRUE or FALSE, denoted by colour). As the figure shows, there is little evidence that the posterior beliefs are stacking on 33% or 67% following receipt of a signal saying "FALSE" or "TRUE", respectively.

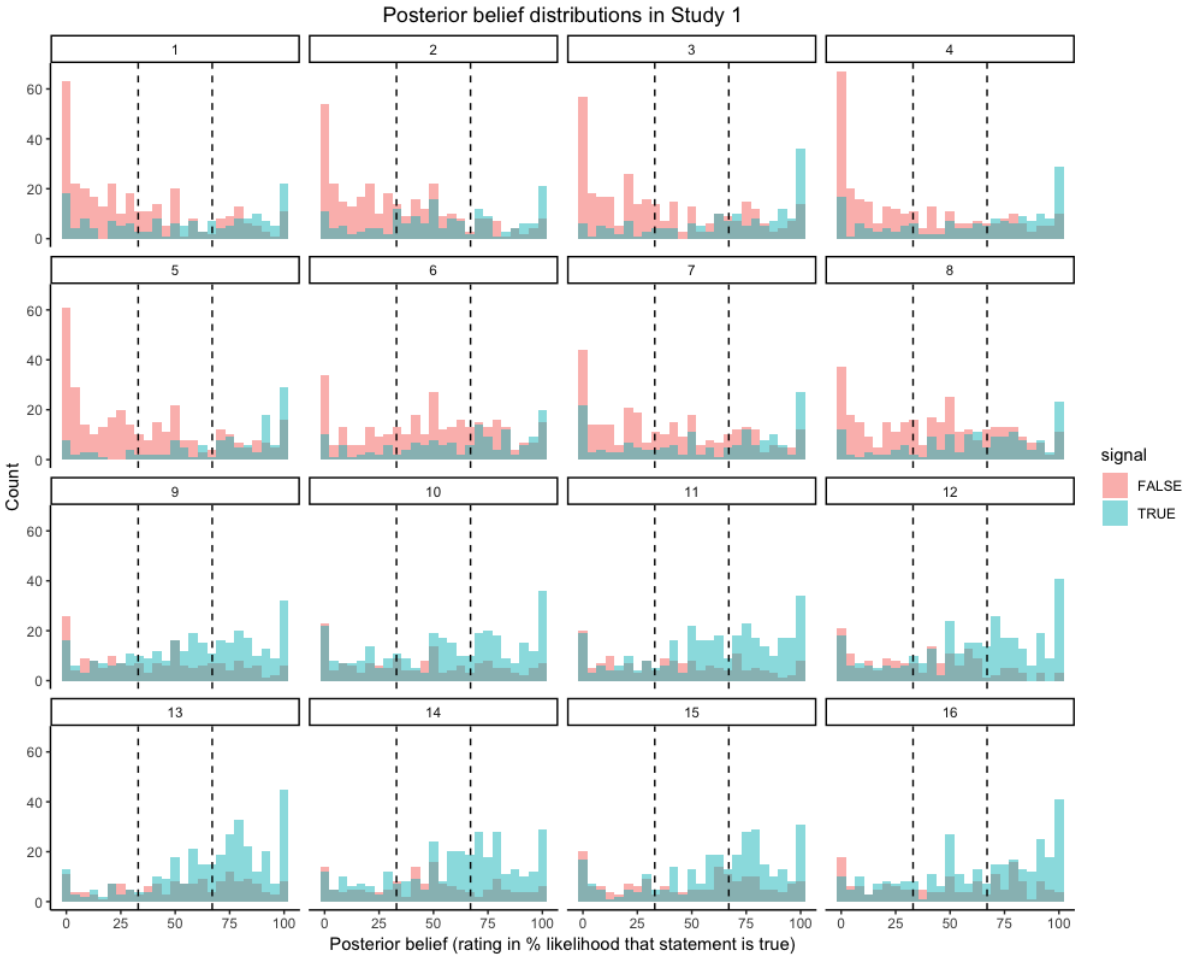


Figure S3. Distributions of posterior beliefs as a function of political statement stimuli (panels 1-16) and signal received (colour) in Study 1. The dashed black lines intersect the x-axis at 33% and 67%.

Study 2

Methods

Sample

Subjects received $1.50 for taking part in Study 2. The discrepancy in fees between Study 1 and 2 is because Study 2 was shorter. Subjects who completed the political statement pre-test or Study 1 were prevented from taking part in Study 2.

Belief Update Task

The task was identical to Study 1, except for the key adjustments outlined in the main text. There were several other minor differences. Specifically, signals were described as "clues" (rather than "signals") and, upon receipt of each signal, subjects were free to continue onto the next trial as soon as they had provided their signal accuracy rating (from 1 to 5) on that particular signal (i.e., they were not required to wait for 5 seconds). The verbatim task instructions and comprehension questions are available on the OSF: https://osf.io/yt3kd/.

Results

H1: High CRT Scorers Deviate Less from Bayesian Updating

Exploratory Tests

As described in the main text, we repeated the exploratory analyses conducted in Study 1 (H1) as a check on the robustness of the preregistered test of H1 (Study 2); except for memory errors, since those data were not collected in Study 2.

Prior beliefs and regression to the mean. The correlation between mean prior belief extremity and CRT performance was small but below the significance threshold, $\tau = -.05$, $Z = -2.31$, $p = .021$, suggesting that subjects who scored higher on the CRT had slightly less extreme prior beliefs on average. As in Study 1 (H1), however, nonparametric partial correlations between

CRT performance and absolute deviation from Bayesian posterior beliefs—adjusting for prior belief extremity—closely reproduced the preregistered results: τ = -.19, Z = -8.77, p <.001 (difference index), and τ = -.06, Z = -2.77, p =.006 (ratio index).

Median absolute deviation. Taking the median—rather than the mean—absolute deviation from Bayesian posterior beliefs over subjects' 16 trials reproduced the preregistered results: τ = -.14, Z = -6.23, p <.001 (difference index), τ = -.07, Z = -3.08, p =.002 (ratio index).

Linear mixed effects modelling. Fitting a maximal model on the trial-level data showed that CRT z-scores were negatively associated with absolute deviation from Bayesian posterior beliefs, b = -1.91 (SE = 0.25), $t_{75.55}$ = -7.71, p <.001 (difference index), b = -0.03 (SE = 0.01), $t_{57.27}$ = -2.79, p =.007 (ratio index); reproducing the preregistered results.

Political concordance of evidence. Nonparametric correlations showed that the negative correlation between CRT performance and absolute deviation from Bayesian posterior beliefs is present across both evidence types; these results are reported in the main text. Two subjects did not receive any signals that could be classified as politically concordant; thus, the politically concordant analyses are based on N = 990 (not N = 992).

H2: High CRT Scorers Deviate from Bayesian Updating Conditional on their Political Identities

Preregistered Tests

For both DVs, model fitting proceeded in the same steps as in Study 1 (unlike in Study 1, however, in Study 2 the maximal random effects structure was specified *correctly* in the preregistered analysis plan).

As stated in the main text, the maximal model fitted on the ratio index DV failed to converge. In line with the preregistered analysis plan, we thus iteratively modified the random effects structure to attain convergence while keeping the model as maximal as possible. After achieving convergence, we attempted model comparison by dropping the key fixed-effect interaction term and performing a LRT. However, dropping this term caused the model to fail convergence. In line with our preregistered analysis contingencies, we thus estimated the df

and p-value on the key fixed-effect interaction term using the *lmerTest* package in R (Kuznetsova et al., 2017). The final models (with parameter estimates and model fit indices) are presented in Table S4.

Exploratory Tests

As reported in the main text, to increase the sensitivity of the Study 2 H2 test we fitted models where political identity was instead represented continuously (i.e., not dichotomized into a preference for Democratic or Republican Party). In contrast to Study 1, in Study 2 we did not collect social or economic political views from participants; only their political party affiliation (provided on a scale from 1 = strong Democrat to 7 = strong Republican). As in Study 1, we standardized and midpoint-centered this variable (new variable: *party_z*, M = -0.20, SD = 1).

We fitted two exploratory linear mixed effects models at the trial-level – one for each DV (difference index, ratio index) – with three IVs. The first IV was CRT performance (z-score). The second IV was evidence type i.e., whether the signal favored Democrats or Republicans, and was computed via combination of signal received (TRUE, FALSE) and the partisanship of the statement (Democrat-favor, Republican-favor) (see Table 2 in the main text). The third IV was *party_z*. The focal exploratory test bearing on H2 is a three-way interaction between these variables.

Both models were estimated with random intercepts, as well as random slope of [evidence type] on subjects and stimuli, and the random slope of [CRT x evidence type x political identity]—the critical three-way interaction—on stimuli. The models were fitted with restricted maximal likelihood, and degrees of freedom and p-values for the three-way interaction are estimated via the *lmerTest* package (Kuznetsova et al., 2017).

In the difference index model, the three-way interaction missed the significance threshold of $p < .05$, $t_{60.10} = 1.85$, $p = .069$, $b = 0.84$ (SE = 0.46). Similarly, in the ratio index model, the three-way interaction missed this significance threshold, $t_{251.80} = 1.70$, $p = .091$, $b = 0.03$ (SE = 0.02). Thus, as in Study 1, we conclude that the preregistered test of H2 is not obscuring identity-protective deviation from Bayesian posterior beliefs by treating political identity dichotomously.

Table S4. Linear Mixed Effects Model Output in Study 2 (Hypothesis 2).

	Difference index			Ratio index		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
Fixed Effects						
(Intercept)	-11.99	0.83	<.001	-0.33	0.03	<.001
CRT	0.47	0.44	.291	-0.01	0.01	.498
Political concordance	0.04	0.52	.939	-0.04	0.02	.030
CRT x Political concordance	1.02	0.52	.048	0.03	0.02	.053
Random Effects						
σ^2	457.424			0.882		
$\tau_{00,\ Subject}$	99.647			0.084		
$\tau_{00,\ Stimuli}$	8.486			0.008		
ρ_{01}	-0.554			-0.403		
$N_{Subject}$	992			992		
$N_{Stimuli}$	16			16		
$ICC_{Subject}$	0.176			0.086		
$ICC_{Stimuli}$	0.015			0.008		
Observations	15526			15526		
R^2 / Ω_0^2	.002 / .170			.001 / .099		
Deviance	140739.370			43212.822		

Note. The ratio index model is estimated via ML because REML did not converge (the difference index model is estimated via REML). P-values in the table are estimated via Wald test (note that the Wald test p-value for the interaction in the difference index model is .048, but the preregistered LRT p-value—reported in the main text—was .053). CRT = Cognitive Reflection Test (z-score).

H3: High CRT Scorers Evaluate Evidence Conditional on their Political Identities

Preregistered Tests

The model fitting procedure was the same as in the test of H2 (except that, in H3, the DV was signal accuracy ratings). The maximal model parameter estimates and fit indices are presented in Table S5.

Table S5. Linear Mixed Effects Model Output in Study 2 (Hypothesis 3).

	Signal accuracy ratings		
	B	*std. Error*	*p*
Fixed Effects			
(Intercept)	3.08	0.03	<.001
CRT	-0.03	0.02	.134
Political concordance	0.61	0.05	<.001
CRT x Political concordance	0.06	0.03	.054
Random Effects			
σ^2		1.089	
$\tau_{00,\ Subject}$		0.251	
$\tau_{00,\ Stimuli}$		0.011	
ρ_{01}		-0.749	
$N_{Subject}$		992	
$N_{Stimuli}$		16	
$ICC_{Subject}$		0.186	
$ICC_{Stimuli}$		0.008	
Observations		15872	
R^2 / Ω_0^2		.067 / .227	
Deviance		48296.821	

Note. The model is estimated via ML because REML did not converge. P-values in the table are estimated via Wald test. CRT = Cognitive Reflection Test (z-score).

Exploratory Tests

Sensitivity analysis. As described in the main text, we fitted a linear mixed effects model where political identity was represented continuously rather than dichotomously (as in the preregistered test of H3). For this analysis, we midpoint-centered and standardized the 1-7 political party affiliation variable (new variable: *party_z*) (original scoring: 1=strong Democrat, 2=Democrat, 3=lean Democrat, 4=Independent, 5=lean Republican, 6=Republican, 7=strong Republican).

The model was estimated with three IVs: CRT z-score, party_z, and whether the signals favored Democrats or Republicans (signal type) (see Table 2 in the main text). The DV was signal accuracy ratings. The test of interest is on the three-way interaction between the IVs. The maximal model did not converge, and so the model was fitted with random intercepts on both subjects and stimuli, random slope of [signal type] on subjects, and random slope of [CRT x party_z x signal type] on stimuli. A Likelihood Ratio Test showed that, unlike in the preregistered test of H3, in this model the fixed-effect interaction improved model fit at $p < .05$, χ^2 (1) = 4.76, p =.029, b = 0.07 (SE = 0.03). These results are reported in the main text.

Prior beliefs analysis. For this analysis, we first determined whether signals were concordant or discordant with subjects' prior beliefs by the following method (recall that prior beliefs were provided on a 0-100 scale):

<u>Signal concordant with prior belief</u>
Prior belief < 50 & signal = FALSE; OR prior belief > 50 & signal = TRUE

<u>Signal discordant with prior belief</u>
Prior belief < 50 & signal = TRUE; OR prior belief > 50 & signal = FALSE

This provided a dichotomous variable [0=discordant, 1=concordant]. Trials on which the prior belief was exactly 50 were excluded from this analysis (N trials = 1024, 6.45%). To represent how extreme (strong) subjects' prior beliefs were, we used the prior belief extremity variable previously computed in Study 1. Recall that this variable was computed by taking the absolute distance of subjects' prior beliefs on each trial from the scale midpoint. Thus, values ranged from 1 to 50, where 1=minimum prior belief extremity (i.e., 49 or 51 on the likelihood scale:

min. strength prior) and 50=maximum prior belief extremity (i.e., 0 or 100 on the likelihood scale: max. strength prior).

We then proceeded to jointly model the three-way interaction between (i) CRT scores and these two new prior belief variables, and (ii) CRT scores, political party affiliation [party_z], and whether the signals favored Democrats or Republicans [signal type]. Given the complexity of this joint model, we fitted only random intercepts on subjects and stimuli (i.e., no random slopes).

Prior to fitting the joint model, we checked that the three-way interaction between CRT scores, political party affiliation [party_z], and signal type significantly improved fit in this new joint model. This was necessary because, in the new joint model, we had to exclude trials on which prior beliefs were equal to 50 (see preceding paragraph) and, in addition, the random effects structure was very different than in the sensitivity analysis reported above (i.e., we no longer modelled random slopes). It was thus necessary to rule out that *these* model changes were not responsible for the elimination of the [CRT x party_z x signal type] interaction reported in the main text (rather than the modelling of prior beliefs per se). The results suggested they were not. In the new joint model—before modelling prior beliefs—consistent with the sensitivity analysis reported above, the three-way interaction [CRT x party_z x signal type] was statistically significant and positive, $t_{14417} = 3.42$, $p <. 001$, $b = 0.06$ ($SE = 0.02$). Thus, we proceeded with fitting the joint model that also included prior beliefs. The relevant results of this model are reported in the main text.

Distributions of Posterior Beliefs

In figure S4, we plot the distributions of posterior beliefs as a function of the political statement stimuli (panels 1-16) and signal received (TRUE or FALSE, denoted by colour). As in Study 1, the figure shows there is little evidence that the posterior beliefs are stacking on 33% or 67% following receipt of a signal saying “FALSE” or “TRUE”, respectively.

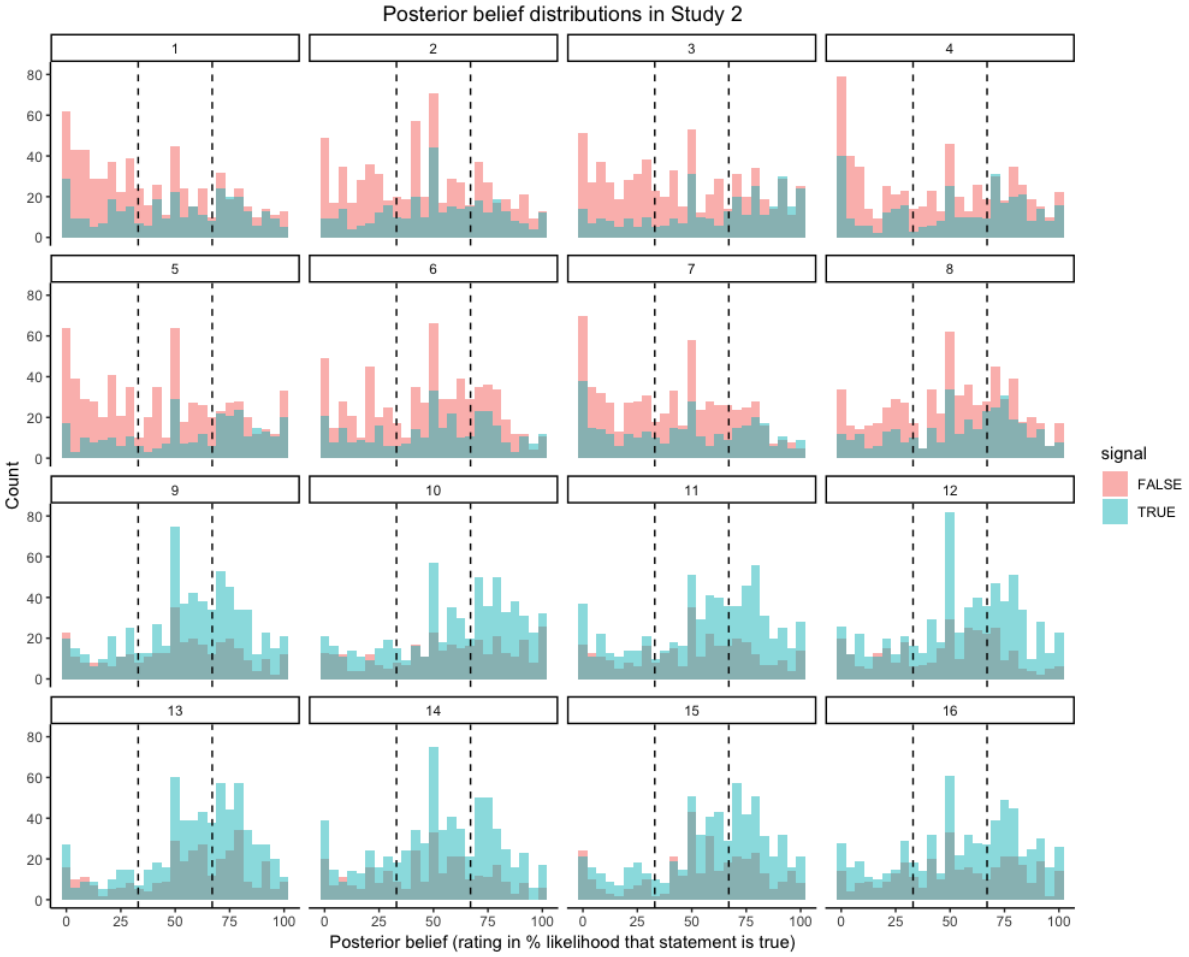


Figure S4. Distributions of posterior beliefs as a function of political statement stimuli (panels 1-16) and signal received (colour) in Study 2. The dashed black lines intersect the x-axis at 33% and 67%.

Study 3

Methods

Sample

Subjects received $0.60 for completing the study. Given that Study 3 comprised a different design, stimuli, and variables than Study 1 and 2, we did not prevent subjects who took part in those studies taking part in Study 3.

Results

Parameter estimates from the preregistered model in H3 are reported in Table S6.

Table S6. Linear Regression Model Output in Study 3 (Hypothesis 3).

	(1)		
	B	*std. Error*	*p*
(Intercept)	4.25	0.07	<.001
Treatment	-0.10	0.10	.300
Political identity	-0.08	0.04	.024
CRT	-0.27	0.07	<.001
Prior belief	-0.53	0.07	<.001
Treatment x Pol ID	0.17	0.05	<.001
Treatment x CRT	0.13	0.10	.179
Pol ID x CRT	0.01	0.04	.732
Treatment x Prior	0.92	0.10	<.001
CRT x Prior	-0.17	0.07	.014
Treatment x Pol ID x CRT	-0.03	0.05	.621
Treatment x CRT x Prior	0.20	0.09	.035
Observations		1200	
R^2 / adj. R^2		.175 / .167	
Deviance		2648.314	

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

Study 4

Methods

Sample

Subjects received $1 for completing the study. The distributions of the political party identification and political ideology variables are displayed in Figure S5.

Results

Parameter estimates from the preregistered model in H3 are reported in Table S7.

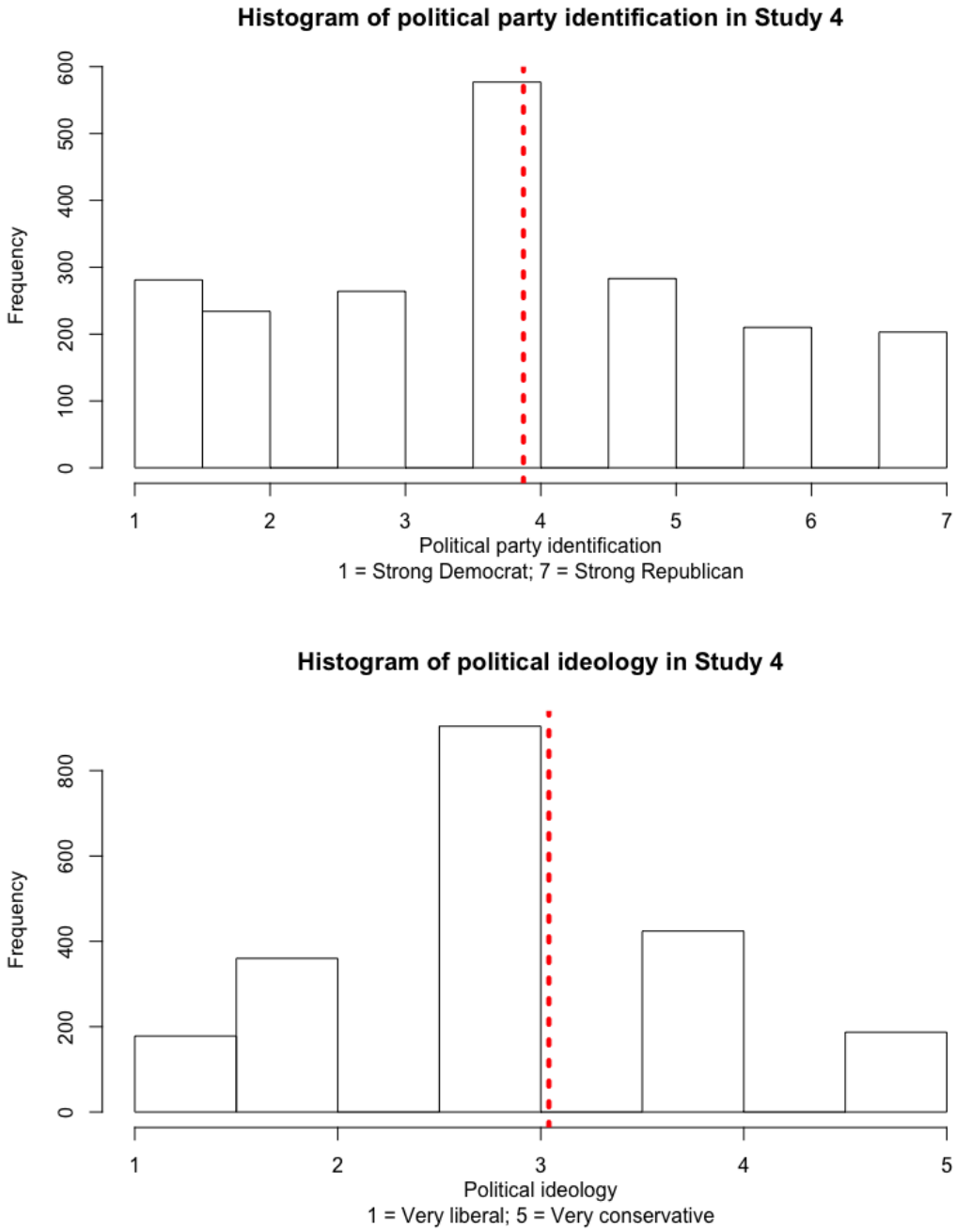


Figure S5. Histogram of political party identification and political ideology variables in the Study 4 sample. *N = 2052 for political party identification; N = 2053 for political ideology. The dashed red line indicates the mean.*

Table S7. Linear Regression Model Output in Study 4 (Hypothesis 3).

	(1)		
	B	*std. Error*	*p*
(Intercept)	4.59	0.05	<.001
Treatment	-0.13	0.07	.082
Political identity	-0.09	0.03	.002
CRT	-0.11	0.05	.037
Prior belief	-0.28	0.05	<.001
Treatment x PID	0.16	0.04	<.001
Treatment x CRT	-0.12	0.07	.093
PID x CRT	-0.02	0.03	.521
Treatment x Prior	0.42	0.07	<.001
CRT x Prior	-0.04	0.06	.423
Treatment x PID x CRT	0.06	0.04	.111
Treatment x CRT x Prior	0.15	0.08	.049
Observations		2052	
R^2 / adj. R^2		.066 / .061	
Deviance		4744.732	

Note. The DV is agreement that the test supplies good evidence of how open-minded someone is (higher values = greater agreement). CRT = Cognitive Reflection Test (z-score).

Study 5

Results

Parameter estimates from the regression models reported in the main text—prior beliefs analysis (Table S8) and political identity/joint analysis (Table S9) are displayed below.

Table S8. Linear Regression Model Output in Study 5 (Prior Beliefs Analysis).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	4.45	0.40	<.001	4.73	0.41	<.001
Treatment	-0.17	0.50	.732	-0.68	0.53	.204
Prior belief	0.71	0.70	.312	-0.37	0.80	.648
Prior belief strength	0.03	0.01	.012	0.02	0.01	.099
CRT	0.12	0.20	.556	-0.07	0.21	.751
Treatment x Prior	-0.56	0.79	.481	1.23	1.03	.232
Treatment x Strength	-0.02	0.01	.131	-0.00	0.01	.740
Treatment x CRT	-0.07	0.27	.782	0.30	0.30	.313
Prior x Strength	-0.05	0.02	.008	-0.02	0.02	.428
Prior x CRT	-0.25	0.34	.461	0.51	0.44	.245
Strength x CRT	-0.01	0.01	.151	-0.00	0.01	.734
Treatment x Prior x Strength	0.07	0.02	<.001	0.02	0.03	.535
Treatment x Prior x CRT	0.31	0.28	.271	-1.07	0.58	.067
Treatment x Strength x CRT	0.00	0.01	.708	-0.01	0.01	.273
Prior x Strength x CRT	0.00	0.01	.746	-0.02	0.01	.088
Treatment x Prior x Strength x CRT				0.05	0.02	.007
Observations	443			443		
R^2 / adj. R^2	.149 / .121			.163 / .134		
Deviance	852.383			838.184		

Note. The DV is judgment that polling data are informative (higher values = more informative). CRT = Cognitive Reflection Test (sum score).

Table S9. Linear Regression Model Output in Study 5 (Political Identity and Joint Analysis).

	(1)			(2)		
	B	*std. Error*	*p*	*B*	*std. Error*	*p*
(Intercept)	5.58	0.22	<.001	5.55	0.50	<.001
Treatment	-1.39	0.31	<.001	-2.11	0.63	<.001
Preference	-0.92	0.30	.002	-0.87	0.32	.006
CRT	-0.20	0.13	.130	-0.31	0.26	.226
Treatment x Preference	1.82	0.41	<.001	1.76	0.42	<.001
Treatment x CRT	0.26	0.18	.143	0.86	0.36	.016
Preference x CRT	0.05	0.18	.778	0.25	0.19	.185
Treatment x Preference x CRT	-0.37	0.25	.133	-0.71	0.26	.006
Prior belief				-1.14	0.84	.173
Prior belief strength				0.01	0.01	.570
Treatment x Prior				2.44	1.06	.022
Treatment x Strength				0.01	0.02	.489
Prior x Strength				0.01	0.02	.689
CRT x Prior				0.80	0.46	.079
CRT x Strength				0.00	0.01	.812
Treatment x Prior x Strength				-0.03	0.03	.394
Treatment x CRT x Prior				-1.59	0.60	.008
Treatment x CRT x Strength				-0.02	0.01	.090
CRT x Prior x Strength				-0.03	0.01	.026
Treatment x CRT x Prior x Strength				0.06	0.02	<.001
Observations	443			443		
R^2 / adj. R^2	.084 / .069			.199 / .163		
Deviance	917.712			802.151		

Note. The DV is judgment that polling data are informative (higher values = more informative). CRT = Cognitive Reflection Test (sum score).

General Discussion

This thesis presented an empirical investigation at the intersection of *moral psychology*—people's perceptions of good and bad, right and wrong—and *belief formation*—how people evaluate evidence and update their beliefs. The investigation comprised two main parts. In Part I, across 6 studies I investigated (i) people's beliefs about their own moral goodness relative to the average person—I asked whether and to what extent such beliefs are irrational—and (ii) people's beliefs about the moral goodness of their political in-party relative to their political out-party. Using economic games, I subsequently tested whether people's beliefs regarding (i) and (ii) predicted behavioural outcomes. Finally, I tested whether people's motivation to "do the morally right thing" underpins their prosocial behaviour in economic games. In Part II, also comprising 6 studies, I investigated several factors purported to influence political belief formation. Specifically, I investigated (i) whether belief updating was biased by people's prior beliefs or by their desired political outcomes (or both), and (ii) the extent to which cognitive sophistication facilitates biased information processing such that people who are more sophisticated are more likely to form factual beliefs favourable to their political identities. In the following sections, I present (i) a brief summary of the main results, (ii) a critical discussion of these results and suggested directions for future research, and (iii) my overall conclusions.

Summary of Findings

Part I: Perceived Moral Superiority and Moral Behaviour

In Part 1.1, I adapted a method of measuring self-enhancement that accounts for uncertainty in social perception and thereby enables isolation of residual or "irrational" self-enhancement. My central finding was that such irrational self-enhancement was largest in the *moral* domain (vs. the nonmoral domains of agency and sociability). However, the magnitude of irrational moral self-enhancement was only trivially and non-significantly correlated with self-esteem, opposite to the predictions of influential prior theory.

In Part 1.2, I tested whether the magnitude of irrational moral self-enhancement ("superiority") identified in Part 1.1 predicted behaviours commonly considered moral; freely helping others, and reciprocating trust. My central finding was robust support for the null hypothesis: strength

of self-perceived moral superiority did not meaningfully predict giving in either the Dictator Game or Trust Game. However, I found some evidence that perceptions of one's own moral goodness—not superiority over others *per se*—predicted giving behaviour (albeit weakly).

In Part 1.3, I extended the investigation of perceived moral superiority to Democratic and Republican partisans in the US. I tested whether the disparity in moral evaluation between the in-party and out-party—referred to here as *moral polarization*—predicted behavioural hostility toward the out-party. I found that moral polarization *per se* was large—larger than polarization in nonmoral domains of evaluation—but I observed somewhat *un*convincing evidence that it predicted behavioural expressions of out-party hostility. These results strike an optimistic chord and converge with recent work in political science on the limits of partisan prejudice.

Finally, in Part 1.4 I conducted an improved and extended test of the *morality preference hypothesis*. This hypothesis states that prosocial behaviour is motivated by people's preference for "doing the morally right thing". I corrected confounds that I identified in prior work and extended their design to answer several unresolved questions relevant to this hypothesis. I convincingly replicated support for the morality preference hypothesis. However, through my extension of the original design I found evidence contrary to influential psychological theory.

Part II: Desires, Identities and Bias in Political Belief Formation

In Part 2.1, I attempted to tease apart two "biases" in belief updating that are often confounded: referred to here as confirmation bias and desirability bias. I conducted a study capitalizing on the context of the 2016 US presidential election—where many people held preferences (e.g., for Donald Trump) at odds with their belief about who was likely to win (e.g., Hillary Clinton). I recruited groups of people whose preferences and beliefs were congruent or incongruent, and I randomly assigned evidence emphasizing either that one or the other of these candidates were more likely to win; thereby, decoupling the predictions of confirmation bias and desirability bias in the aggregate sample. Subsequently, I found evidence to suggest that people updated their beliefs by a greater magnitude if the new information was desirable (vs. undesirable). In contrast, I found little evidence for the corresponding asymmetry in updating for information that confirmed (vs. disconfirmed) prior beliefs.

In Part 2.2, I conducted a comprehensive test of the hypothesis that cognitive sophistication facilitates identity-protective bias in political belief formation. The logic of this hypothesis is that people with distinctive cognitive resources are expected to bring those resources to bear on information that threatens their political identities; specifically, to resist and disregard it. I tested this hypothesis in the context of two processes relevant to belief formation: (a) *belief updating*—that is, how beliefs change after receipt of evidence—and (b) *reasoning/evidence evaluation*—beliefs about the validity or quality of the evidence itself. I inferred cognitive sophistication from scores on the Cognitive Reflection Test (CRT), a behavioural measure of the propensity to think analytically. Regarding (a), I observed that—contrary to the target hypothesis—people who scored higher on the CRT deviated less from the posterior beliefs of a Bayesian agent; implying *less* bias in belief updating. Regarding (b), I found evidence to suggest that people who scored higher on the CRT conditioned more strongly on their *prior beliefs*—rather than on their political identities *per se*—when evaluating the new information.

Critical Analysis of Findings and Future Directions

In the following critical analysis, to avoid repetition I group my discussion by broad themes that interlink different subparts of the thesis rather than focusing on the individual studies one at a time—because this was already done within each paper/subpart. I also outline directions for future work.

Perceived Moral Superiority and Behavioural Outcomes

Taken together, the results of Part I present a compelling case for a two-fold conclusion. First, that perceived moral superiority—both at the level of the individual (Parts 1.1 and 1.2) and the (US) political party (Part 1.3)—is prevalent, large in magnitude, and larg*er* in magnitude than superiority perceived in *non*moral domains of social perception. Second, however, that these perceptions are not straightforwardly associated with behavioural outcomes, as predicted by theory and past work.

In particular, in Part 1.1 the magnitude of "irrational" moral superiority was convincingly *not* associated with self-esteem (albeit, a self-report measure of self-esteem). This result stands in contrast to an influential hypothesis that assumes that perceptions of superiority are prevalent

because they protect and/or enhance wellbeing (Taylor & Brown, 1988; see also Sedikides & Gregg, 2008). This result was all the more puzzling, given that self-esteem *did* correlate with irrational superiority perceived in the *non*moral domains of perception (agency and sociability). This implies there may be something special about the moral domain in this case. I speculated that this something special may be the asymmetric costs/benefits that accrue to people who *under*perceive—versus *over*perceive, or, perhaps even *accurately* perceive—the morality of unknown others. For example, mistaking another person as trustworthy, when in fact they are not, may be associated with greater fitness costs than the reverse error. Under such conditions, individuals may tolerate a loss in judgment accuracy—systematically underestimating others' morality—for gains made elsewhere (cf. Fetchenhauer & Dunning, 2006). This suggestion is bolstered by the fact that the magnitude of self-perceived moral superiority was driven mainly by variance in people's perceptions of the *average person*, rather than themselves—a pattern in evidence in both Parts 1.1 and 1.2.

This proposition, however, is at odds with evidence that people are fairly *accurate* in estimating the moral behaviour of others (Epley & Dunning, 2000, 2006), and, furthermore, the argument that behavioural biases—for example, the behavioural tendency not to entrust others with secrets or money—do not entail *cognitive* biases—the tendency to *believe* that others are untrustworthy (McKay & Efferson, 2010). This undercuts the rationale for the above suggestion that self-perceived moral superiority persists because of an advantageous and systematic underestimation of the morality of other people.

With the benefit of hindsight, then, here I offer an alternative explanation—based on moral *signaling*. Given the myriad social benefits that accrue to people who are perceived by others to be moral—that is, fair, trustworthy, and so on—individuals are prone to *signal* their moral qualities, in numerous different contexts and by various different routes (e.g., Jordan, Hoffman, Bloom, & Rand, 2016a; Jordan, Hoffman, Nowak, & Rand, 2016b). For example, in one recent experiment, people *used*—and observers *interpreted*—the time taken to decide whether to help someone else as a signal of trustworthiness (cf. Jordan et al., 2016b). Consequently, observers tended to trust people who chose to help a third-party without hesitation more than those who took longer to decide; *even if* the latter helped after deliberating. Furthermore, beyond directly signaling one's own moral quality in this way, recent evidence suggests that people are also willing to denigrate the morality of *others* in order to make themselves look morally better (Pleasant & Barclay, 2018).

The upshot of this research is the following: Faced with judging themselves and the average person on various moral traits—as in Parts 1.1 and 1.2 of this thesis—it may be *expected* that people are simply prone to report an exaggerated estimate of the moral traits they themselves possess, and a less rosy estimate of the average persons'; primarily—and straightforwardly—in order to signal that they are high quality partners for cooperation. Indeed, the subjects in Parts 1.1 and 1.2 faced *no* disincentive to adopt such a reporting strategy. In other words, the prevalence of self-perceived moral superiority may, in part, be conceived of as a form of "cheap talk" (Farrell & Rabin, 1996). Importantly, this conception does not entail that people's *true* beliefs are, in fact, what they report. This allows for the possibility that people have an underlying accurate perception of the morality of others (cf. Epley & Dunning, 2000, 2006) and, at the same time, sidesteps the criticism that behavioural bias (i.e., cheap talk) does not entail cognitive bias (a *truly* biased belief that one is morally superior to others) (cf. McKay & Efferson, 2010).

A particular strength of the cheap talk explanation for the prevalence of self-perceived moral superiority is that it can account for the behavioural results reported in Part 1.2, as well. Recall that I found compelling support for the null hypothesis of no association between self-perceived moral superiority and either (a) freely helping others or (b) reciprocating trust. In those cases, I used economic games—the Dictator Game and Trust Game—to measure behaviour, where there were real financial stakes involved. These are exactly the kinds of situations (i.e., financially incentivized) where cheap talk is revealed to be just that (cf. Farrell & Rabin, 1996). In other words, cheap talk provides a poor guide as to how people will behave when there is money—or some other quantity of importance—at stake. Given this, the moral signaling/cheap talk hypothesis would *predict* the null result that I observed in the studies reported in Part 1.2.

What about the results of Part 1.3? There, recall that I found *un*convincing evidence of an association between *moral polarization*—the tendency for partisans to view co-partisans' moral character positively and opposing partisans' negatively—and behavioural expressions of out-party hostility. Here, too, the cheap talk hypothesis may help explain the empirical results. As referenced in the general introduction of this thesis, a growing body of work in political science suggests that *partisan cheerleading*—expressing support for one's political party in responses to survey questions—is likely to inflate estimates of bias in political belief

formation and polarization (Bullock et al., 2015; Khanna & Sood, 2018; Prior et al., 2015; Schaffner & Luks, 2018). Partisan cheerleading, then, is essentially a politics-specific version of "cheap talk". Following this logic, in the context of moral polarization, partisans may be *expected* to morally champion the in-party and denigrate the out-party. Furthermore, given that I used economic games in Part 1.3 to measure out-party hostility, the expectation of the cheap talk hypothesis in this case would likely be a null, or, at least, a small association between moral polarization (self-report) and out-party hostility (financial stakes). In particular, because "putting money on the line" reveals people's self-reported expressions to be only weakly indicative of their behaviour with stakes. Indeed, this is exactly what I observed in those studies: moral polarization *per se* was an order of magnitude larger than the relationship between moral polarization and expressions of out-party hostility.

Finally, the results of the study reported in Part 1.4 are broadly consistent with the moral signaling/cheap talk hypothesis advanced as an explanation for the results in Parts 1.1 and 1.2. In the former study, I used moral frames to manipulate the behaviour of people in economic games. There, I found that simply framing one or the other behaviour as morally appropriate—by labelling the focal choice *fair* or *generous*, or the counterpart choice *unfair* or *ungenerous*—profoundly affected people's choices. Specifically, while approximately 70% of people chose one option when that choice was framed as morally appropriate, this dropped to 40% when the alternative choice was framed as morally appropriate; a swing of 30%, and a reverse in the majority decision. Given this large swing toward the behaviour framed as morally appropriate, this result is entirely consistent with the assumption that people are behaving so as to signal their moral quality; that is, fitting with the moral signaling/cheap talk explanation outlined above. Of course, a particular weakness of this explanation is that people in the economic games used in Part 1.4 were not directly observed by *other* people when making their decisions. As a result, their patterns of decision-making may thus reflect *self*-signaling, that is, signaling one's moral quality to *oneself* (Grossman, 2010; Mijović-Prelec & Prelec, 2010), as much as they reflect social signaling—that is, signaling aimed at other people.

To determine whether social signaling (vs. merely self-signaling) can account for the results of Part 1.4, I propose an experiment that adapts the design of the aforementioned moral signaling studies (Jordan et al., 2016a, 2016b). It will help to describe that design here. In those studies, people completed a *two*-step task, where each step was comprised of a different economic game. In the first step, some people (the "choosers") played a Helping Game, where they were

given money and had the choice of whether to donate some of their endowment to a third-party who received nothing. In the second step, the choosers played a Trust Game in the role of trustee—where a *new* person (the trustor) decided how much money to transfer to them (the chooser). The crucial part of the design was that, in the first step, the choosers were randomly assigned to one of two treatments. In one treatment, choosers were informed that the trustor in the subsequent game (i.e., in the Trust Game) would be able to *condition* their transfer decision on the choice made by the chooser in the first (i.e., Helping) game. This provided strong incentive for the choosers to help—and help without hesitation—in the first game; that is, so they appeared trustworthy to the trustor in the second game. In other words, so as to *signal* their trustworthiness. In the other treatment, in contrast, the choosers' decision to help (or not) was *not* observable by the trustor in the second game—*removing* the social signaling incentive to help in the first game. Therefore, any difference in the decision-to-help rate among choosers between treatments constitutes evidence that the social signaling incentive had a causal effect on their moral decision-making in the first game.

In my proposed experiment, I will integrate this design with the design of the study reported in Part 1.4. Specifically, people will complete a two-step task as above. In the first step, people—*choosers*—will play the Trade-Off Game (TOG), where they must decide between one of two choices about how to distribute money between themselves and two helpless third-parties (I refer back to Part 1.4 for greater detail about the TOG). In one treatment, one choice will be framed as morally appropriate—while, in the other treatment, the alternative choice will be framed as morally appropriate. Then, in a second step, the choosers will play a Trust Game in the role of trustee, with a new person (the trustor) who will decide how much money to transfer them; exactly as in Jordan et al. (2016b). As before, too, I will randomly assign the choosers in the TOG to one of two signaling treatments. In one treatment, I will inform the choosers that their decision in the TOG is observable by their to-be partner (trustor) in the subsequent Trust Game. In the other treatment, in contrast, choosers' TOG decisions will *not* be observable by the trustors. An example of the proposed design is displayed in Figure 1.

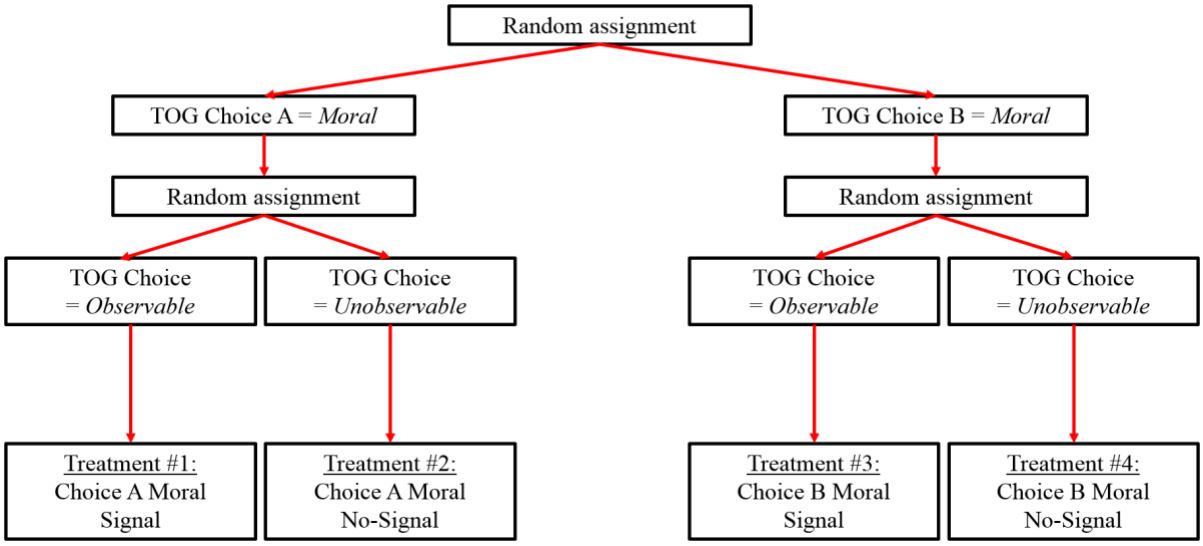


Figure 1. Design for proposed moral signaling experiment. *First, people (i.e., choosers) are randomly assigned to either one or the other moral framing treatment (either Choice A or Choice B is framed as the morally appropriate one). Subsequently, but before making their choices, these choosers are again randomly assigned; this time, to receive either a trustor who will observe their TOG decision vs. will not observe their TOG decision for the purposes of the Trust Game to be played afterwards. See in-text for further detail. TOG = Trade-Off Game.*

As can be seen in Figure 1, the design results in four treatment groups[40]. First, consider the number of people switching from choice A to B in the *no-signal* treatments (#2 and #4). This difference indicates the effect of the moral frame on choices in a *private* context, that is, where choices are *not* observable. Based on the results of the study reported in Part 1.4, I would predict a roughly 30% swing from one choice to the other—since treatments #2 and #4 essentially replicate the design of that study. The test of the social signaling hypothesis, therefore, concerns the number of people switching from choice A to B in the *signal* treatments (#1 and #3). In particular, if this difference is greater than the difference in the no-signal treatments—a statistically significant interaction between *moral frame* and *signal* factors—one

[40] In the actual experiment, there would also be control treatments to allow a comparison to baseline. This is important, because people's choices may change as a function of the no-signal vs. signal treatment—*irrespective* of the moral framing treatment. Put another way, *absent* any moral frames, people's choices may change by virtue of the signaling treatment itself. A control group allows one to observe and account for this change. I omitted discussion of the control group in-text to simplify exposition of the proposed design.

can infer that people's response to the moral frame is, in part, motivated by a desire to signal to *others* (not just themselves) that they are a moral person. In contrast, a failure to reject the null hypothesis of no interaction (assuming high statistical power) would provide support for the inference that the results in Part 1.4 are due to *self*-signaling. Data collection for this study is already underway (with collaborators).

Bias in Political Belief Formation

The results of the studies reported in Part II imply several conclusions. I consider these in turn. First, a key result of Part 2.1—in particular, the *lack* of an association between people's prior beliefs and their incorporation of new evidence into their posterior beliefs—stands in contrast to the famous "attitude polarization" result reported by Lord and colleagues (1979). In that study, people provided their prior beliefs before receiving two pieces of evidence—one that was belief-consistent, and one that was belief-*in*consistent. After evaluating the evidence, people reported whether their prior beliefs had become more extreme, less extreme, or stayed the same. As described by those authors, people tended to report that their prior beliefs had become more extreme. Lord et al. (1979) concluded from this that people's belief updating was asymmetric; they incorporated new evidence into their posterior beliefs more if the evidence was consistent (vs. inconsistent) with their priors. A crucial design choice in their study, however, was that instead of asking people to *directly* provide their posterior beliefs—on the same scale as they provided their prior beliefs—people were asked to *judge* whether their prior beliefs had, in fact, become more extreme. In other words, the researchers asked people to *estimate the causal effect of the evidence on their own beliefs*—a meta-belief, of sorts.

This is crucial, for three reasons. First, because in Part 2.1—in contrast to Lord et al. (1979)—I asked subjects to *directly* report both their prior *and* their posterior beliefs—that is, using the same scale. I then inferred belief change as the difference between these values. Second, the measurement strategy used by Lord and colleagues (1979) provides misleading estimates of belief polarization—compared to direct measurement of beliefs in randomized experiments, arguably the gold-standard (Graham & Coppock, 2018). Third, and relatedly, investigations conducted after the publication of Lord et al. (1979) found that "attitude polarization" did not replicate on these more appropriate, direct measures of belief change (Guess & Coppock, 2018; Kuhn & Lao, 1996; Miller et al., 1993; Munro & Ditto, 1997). In other words, where

polarization is defined as *actual* divergence in the posterior beliefs—and not people's meta-perceptions of said divergence—asymmetric updating conditional on prior beliefs is *unobserved*. The findings of my investigation in Part 2.1—the *lack* of an association between people's prior beliefs and their incorporation of new evidence into their posterior beliefs—are thus consistent with these latter studies. On the other hand, my findings are also consistent with the hypothesis that people's prior beliefs and preferences were confounded in Lord et al. (1979); and, thus, that decoupling these factors in my study revealed that asymmetric updating conditional on priors was *not* driving their original result. Rather, perhaps, it was asymmetric updating conditional on *preferences*. Future work could directly compare these two hypotheses by designing an experiment that both (i) manipulates the measurement method—i.e., direct measurement of prior/posterior beliefs vs. meta-perception of belief change—*and*, in addition, (ii) decouples people's prior beliefs and preferences.

The second key result of Part 2.1, as highlighted above, was that of a belief updating asymmetry conditional on people's political preferences. Specifically, after decoupling preferences from prior beliefs (in the aggregate sample), I observed that receipt of politically *desirable* evidence caused a greater magnitude of change in the posterior beliefs than receipt of otherwise-identical *un*desirable evidence. On the basis of this asymmetry, I inferred bias on behalf of people's belief updating. However, whether this asymmetry does, in fact, constitute evidence of biased belief updating depends crucially on the conception of "bias". Where bias is taken to mean an asymmetry in the measurement of people's posterior beliefs given evidence, this inference is valid. But this is unsurprising, since one could argue that it merely restates the phenomenon. As argued at length by Hahn and Harris (2014, see also Shah et al., 2016), a stronger case for bias in belief updating demands demonstration of systematic deviation from some verifiably optimal or normative benchmark. For example, in the case of putative bias in belief updating, evidence of systematic deviation from *Bayesian rationality*.

Bayesian rationality follows from the laws of probability, and offers a normative account of how beliefs ought to update given evidence (Etz & Vandekerckhove, 2018). Most important here is that Bayesian rationality does not *entail* that people's prior beliefs—given the same evidence—will update by the same amount (as measured via self-report scales) (Gerber & Green, 1999; Hahn & Harris, 2014); nor, in fact, that they will converge in the posterior beliefs (Jern et al., 2014). This stems in part from the fact that people may have different models of the world—affecting their interpretation of new evidence—as well as from the *multiplicative*

(nonlinear) nature of Bayes' rule, which lies at the heart of Bayesian inference. With respect to the results of Part 2.1, therefore, the evidence is undiagnostic as to whether people's belief updating was *biased*, where bias is conceived of as deviation from the prevailing normative standard of belief updating, Bayesian rationality (Hahn & Harris, 2014). In particular, because people's patterns of belief updating in Part 2.1 were not obviously inconsistent with Bayesian principles. On the contrary, there is evidence that human belief updating—explicitly compared against this normative benchmark—fairly well approximates Bayesian inference in various contexts (e.g., Barron, 2016; Cao, Kleiman-Weiner, & Banaji, 2018; Coppock, 2016; Coutts, 2018; Gotthard-Real, 2017; Hill, 2017; Hornikx, Harris, & Boekema, 2018; but see Bowers & Davis, 2012 for a critique).

Unfortunately, in Part 2.1 it was not possible to construct a Bayesian estimate of the posterior beliefs in order to provide comparison with the observed (i.e., subject) posterior beliefs. This is because such an estimate requires knowledge not only of the prior beliefs, but also, crucially, the likelihood of the evidence (Hill, 2017). That is, how *diagnostic* the evidence is taken to be one way or the other. This information was not readily available in the study design of Part 2.1. Recognizing this limitation, the relevant studies in Part 2.2 were designed such that I provided people with *objective* likelihoods—in Study 1—and, in addition, in Study 2 I obtained their *subjective* likelihoods. This allowed comparison of the observed patterns of belief updating against the Bayesian predictions.

The benefits of this approach are aptly illustrated in the results of those studies. In particular, recall that in Part 2.2 I tested the hypothesis that people who score higher on the Cognitive Reflection Test (CRT) are more biased in favour of their political identities when forming their beliefs—following the logic of identity-protective cognition (e.g., Kahan, 2016a; Van Bavel & Pereira, 2018). However, in studies 1 and 2 (in Part 2.2) I found that individuals who scored higher on the CRT in fact tended to *converge on the posterior beliefs of a Bayesian agent*—implying the *opposite* of exacerbated bias, and thus contrary to the target hypothesis. In the absence of this comparison with the Bayesian expectation, the raw magnitudes of belief updating in those studies may well have suggested that high CRT individuals were indeed more biased in favour of their political identities; *in line with* the target hypothesis. For example, relative to people who scored low on the CRT, people who scored high tended to update to a greater extent in *raw magnitude terms* towards politically favourable—vs. *un*favourable—evidence (for more detail I refer back to studies 1 and 2 in Part 2.2). Only in the light of the

comparison with a normative (Bayesian) expectation do the latter individuals appear overall less biased.

In studies 3-5 of Part 2.2, I investigated a different outcome variable relevant to political belief formation and identity-protective cognition; that is, people's beliefs about the validity of the new evidence itself—in other words, evidence *evaluation* (as opposed to *updating* a belief upon which that new evidence bears, as was the focus of studies 1 and 2). Accordingly, my findings suggested that people who scored higher on the CRT conditioned more on their *prior beliefs* when evaluating the new evidence, rather than conditioning on their political identities *per se*. Indeed, in the General Discussion section of Part 2.2, I argued at length that the evidence often cited in favour of the hypothesis that cognitively sophisticated people are more biased when evaluating political information is relatively *un*diagnostic. Primarily, because that evidence consists in a treatment by covariate interaction, and thus precludes the inference that political identity—and not some other variable, such as prior beliefs—*causes* the "biased" evidence evaluation (e.g., see Gerber & Green, 2012). By the same token, such designs also preclude the inference that cognitive sophistication causes *more* biased evidence evaluation—for the reason that cognitive sophistication is also not randomly assigned, and, therefore, is also likely subject to numerous confounding variables. Rather than recapitulating those arguments here, however, I will discuss instead what I consider to be a broader, unresolved issue with the phenomenon of "biased" evidence evaluation—often referred to as "biased assimilation" in the literature—that was brought to my awareness by the results of Part 2.2.

Biased assimilation is a classic phenomenon in psychology and political science. It consists in the well-documented observation that people are prone to evaluate new evidence conditional on its congeniality to their prior beliefs, political preferences and/or identity commitments (e.g., Corner et al., 2012; Ditto et al., 2018; Kahan, 2016a; Koehler, 1993; Lord et al., 1979; Taber & Lodge, 2006; Thesis Part 2.1, Supplemental Material; Part 2.2, studies 2-5). In particular, people evaluate congenial evidence more favourably than *otherwise-identical un*congenial evidence. The measurement method that is overwhelmingly used to infer biased assimilation is people's self-reported judgments of new information—indeed, that is the approach I used in studies 2-5 in Part 2.2 of this thesis. Specifically, people's self-reported judgments about how *reliable* the evidence is; how *valid* it is; how *trustworthy* it is. Critically, these judgments are *explicitly* assumed to reflect the "weight" people assigned the new evidence for the purposes

of learning from it and updating their beliefs—that is, directly akin to the "likelihood ratio" in Bayesian inference (cf. Ditto et al., 2018; Kahan, 2016a).

Remarkably, however, this measurement assumption has never been tested. This is remarkable because of the extent and impact of research purporting to show biased assimilation of new evidence. For example, the original Lord et al. (1979) paper has been cited over 4000 times, and a Google Scholar search returns 166,000 hits for the phrase "biased assimilation"[41]. It is also *consequential*, for two reasons. First, because several scholars have recently questioned this measurement assumption (Gerber & Green, 1999; Kim, 2018); suggesting that it is too strong. Specifically, these critics claim it is implausible that classic biased assimilation results perfectly—or even reliably—reflect the weight people assigned the evidence for the purposes of learning from it. Second, there exists a substantial body of research *in tension with* biased assimilation—specifically, research showing "parallel updating" in the mass public. This work reveals that Democrats and Republicans in the US update their beliefs about numerous political matters approximately *in parallel* over time (for a review, see Coppock, 2016, Chapter 1). This implies that, in the aggregate, partisans in the US weigh new evidence roughly *commensurately* when updating their beliefs—in direct contradiction to the phenomenon of biased assimilation.

Given this, I propose a series of experiments that will provide a comprehensive and direct test of the biased assimilation measurement assumption. In these experiments, I will obtain a *direct* measure of the weight people assigned the new evidence—by consulting the extent to which the evidence *changed their beliefs*. In fact, this is the *definition of* the "weight" assigned new evidence. I will derive this weight formally, and in a principled manner via Bayesian inference. Bayes' rule dictates that the weight/diagnosticity of new evidence—the likelihood ratio—is a function of two quantities: the prior odds that a hypothesis is true (i.e., before seeing the evidence) and the posterior odds that it is true (after seeing the evidence). Bayes' rule in odds form is given by,

$$\frac{P(H = TRUE \mid D)}{P(H = FALSE \mid D)} = \frac{P(H = TRUE)}{P(H = FALSE)} \times \frac{P(D \mid H = TRUE)}{P(D \mid H = FALSE)}$$

$$Posterior\ Odds \quad = \quad Prior\ Odds \quad \times \quad Likelihood\ Ratio$$

[41] Both these figures are according to a Google Scholar search conducted 21st November 2018.

Where the *posterior odds* that a hypothesis is true (given data) lies to the left of the equality, followed by the *prior odds* that it is true (in the middle); and, finally, the *likelihood ratio*—the probability of the data given that the hypothesis is true over the probability of the data given that the hypothesis is false. As can be seen from this equation, dividing the posterior odds by the prior odds gives the likelihood ratio,

$$\frac{P(D \mid H = TRUE)}{P(D \mid H = FALSE)} = \frac{\frac{P(H = TRUE \mid D)}{P(H = FALSE \mid D)}}{\frac{P(H = TRUE)}{P(H = FALSE)}}$$

To obtain estimates of individuals' *subjective likelihood ratios*, in my proposed experiments I will measure people's prior beliefs on a 0-1 belief scale before providing them with new evidence that bears on those beliefs. People will then be randomly assigned to one of two treatments[42]. In one treatment, I will subsequently measure their posterior beliefs (on the same 0-1 belief scale as the priors); whereas, in the other treatment, I will ask subjects to evaluate the new evidence on *self-report* scales—the classic biased assimilation method. For subjects in the former treatment, using their prior and posterior beliefs—converted to odds form—I will infer their subjective likelihood ratios via the above equation. That is, I will infer the *actual* weight they assigned the evidence as revealed by the extent to which they updated their beliefs. I will compare this quantity to people's *self-reported* judgments of the new evidence provided in the latter treatment; that is, the classic measure of biased assimilation. Figure 2 illustrates this design.

First and foremost, I expect to replicate the classic biased assimilation result. In other words, I expect people to evaluate the new evidence more favourably if it is congenial to their prior beliefs. In the context of Figure 2, for example, people who believe that gun control laws *reduce* crime (prior belief > 0.5) will evaluate evidence that shows that such laws do indeed reduce crime more positively than otherwise-identical evidence that shows such laws do *not* reduce

[42] As before, in the actual experiment there will be control groups. This is necessary in order to prevent regression to the mean confounding my inference of people's subjective likelihood ratios (e.g., see Yu & Chen, 2015). Specifically, given that people's prior and posterior beliefs are subject to natural variation (i.e., variation *not* due to the evidence treatment), I aim to "partial out" this natural variation and isolate the likelihood ratios as they correspond only to the effect of the evidence treatment. A control group achieves this aim. Here, I omit discussion of the control groups to simplify exposition of the design.

crime. Specifically, for these people, evidence ratings will be more favourable in treatment #2 vs. treatment #4; a statistically significant difference.

For the people in treatments #1 and #3, in contrast, I will conduct the same analysis but on their *subjective likelihood ratios*—inferred from the change in their prior and posterior beliefs, as outlined above. If the pattern of these likelihood ratios diverges from the classic biased assimilation result, I will infer a violation of the measurement assumption underlying biased assimilation. That is, a violation of the assumption that self-reported evaluations of evidence track the weight people assigned the new evidence. The *magnitude* of this violation will be determined by the (i) size and (ii) direction of the difference in subjective likelihood ratios. For example, in the case of gun control advocates (prior beliefs > 0.5), if I fail to reject the null hypothesis of no difference in likelihood ratios between treatments #1 and #3—assuming high statistical power—I will infer that these advocates weighted the congenial and uncongenial evidence *commensurately* for the purposes of updating their beliefs. This would be rather damning for the classic result of biased assimilation; which, as mentioned already, is explicitly assumed to show that people condition the weight of new evidence on their prior beliefs (cf. Ditto et al., 2018; Kahan, 2016a). Of course, it is also possible that the subjective likelihood ratio analysis will closely reproduce the size and direction of the discrepancy in self-reported evidence evaluations—*validating* the key measurement assumption of the classic biased assimilation result. These experiments are in the design/pilot study stage currently (with collaborators).

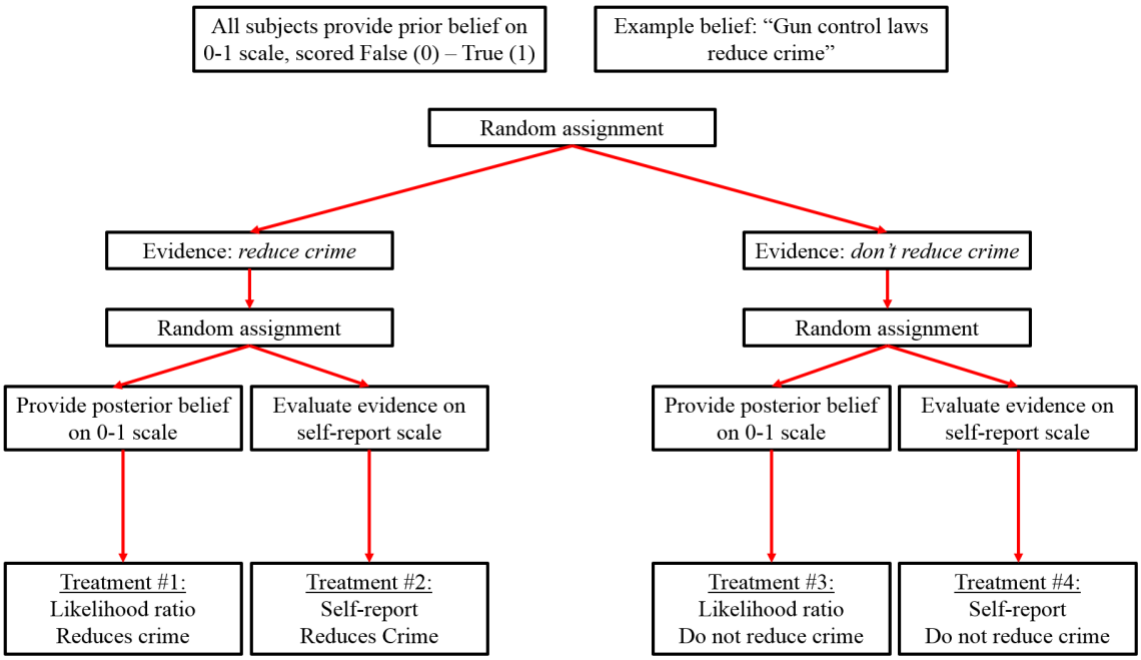


Figure 2. Design for proposed biased assimilation experiment. *First, subjects provide their prior belief on a continuous scale from 0-1 where 0 = False and 1 = True. Subjects are then randomly assigned to receive evidence that is either for or against the belief. They are then randomly assigned again to either (i) provide their posterior belief after seeing the evidence or (ii) evaluate the evidence on self-report scales (i.e., classic biased assimilation measure). This results in four treatment groups. See in-text for more detail.*

Conclusion

On the basis of the research reported in this thesis—and the preceding critical analyses—I draw four main conclusions from my investigation. First, perceptions of moral superiority—both at the level of the individual and the (US) political party—are prevalent, large in magnitude, and larg*er* in magnitude than trait superiority perceived in *non*moral domains of social perception. Second, however, these perceptions do not appear meaningfully associated with behavioural outcomes where there are stakes—for example, money—involved. A plausible explanation for this disconnect is that expressions of moral superiority over the average person, and over one's political rivals, to some extent reflect "cheap talk" and "partisan cheerleading", respectively. Furthermore, these phenomena may be underpinned by the more general motivation to signal

specific things about oneself in responses to survey questions; for example, that one is a moral person, or a loyal partisan group member. I outlined a novel experiment designed to test this hypothesis that is currently in the data collection phase.

My third main conclusion is that the hypothesis that cognitive sophistication facilitates identity-protective *bias* in political belief formation is profoundly underdetermined by current evidence. Primarily, because the designs of oft-cited studies do not permit causal inferences regarding the role of identity *or* cognitive sophistication, and thus face glaring confounds—not least, the prior beliefs of subjects. More generally, too, because such studies rarely attempt to evaluate the observed patterns of results with respect to an optimal or normative benchmark—such as that offered by Bayesian rationality. This segues into my fourth, and final, conclusion: I concur with recent arguments that reasonable inferences of "bias" in human belief formation demand evidence of systematic deviation from well-specified normative standards. I have highlighted one example where such evidence appears absent; that is, in the case of biased assimilation research. I proceeded to outline a novel experimental design that tests the key measurement assumption of classic biased assimilation results—a design situated explicitly with respect to a Bayesian framework, and one currently being piloted by myself and collaborators. It is my desire to continue in this line of work and help advance scientific understanding of when (and why) human belief formation is biased, and when (and why) it is not.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*, e0172792.

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica, 81*, 2269-2302.

Alicke, M. D., & Govorun, O. (2005). *The better-than-average effect*. In M. D. Alicke, D. A. Dunning, & J. Krueger (Eds.) The Self in Social Judgment (pp. 85-106). New York, NY: Psychology Press.

Alicke, M. D., & Sedikides, C. (Eds.). (2011). *Handbook of self-enhancement and self-protection*. New York, NY: Guilford Press.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68*, 804-825.

Alicke, M. D., Vredenburg, D. S., Hiatt, M., & Govorun, O. (2001). The "better than myself effect". *Motivation and Emotion, 25*, 7-22.

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: what counts? *Philosophical Psychology, 26*, 673-701.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*, 211-36.

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality, 40*, 440-450.

Amir, O., & Rand, D. G. (2012). Economic games on the internet: The effect of $1 stakes. *PloS One, 7*, e31461.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423-1440.

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*, 99-131.

Arnold, J. B. (2017). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 3.4.0. https://CRAN.R-project.org/package=ggthemes

Auguie, B, & Antonov, A. (2017). Package 'gridExtra'. Available at https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf

Back, M. D. & Vazire, S. (2012). *Knowing our Personality*. In S. Vazire and T. D. Wilson (Eds.) Handbook of Self Knowledge (pp. 131-156). New York, NY: Guilford.

Bakker, B. N., Lelkes, Y., & Malka, A. (2018). An Expressive Utility Account of Partisan Cue Receptivity: Cognitive Resources in the Service of Identity Expression. Retrieved from https://www.dropbox.com/s/vml5eka1jwp8bhc/bakkeretal.pdf?dl=0

Baron, J. (2008). Actively open-minded thinking. *Thinking and Deciding*, 199-232.

Baron, J., & Jost, J. T. (2018). False Equivalence: Are Liberals and Conservatives in the US Equally "Biased"? Retrieved from https://www.sas.upenn.edu/~baron/papers/dittoresp.pdf

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*, 74-85.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Barranti, M., Carlson, E. N., & Furr, R. M. (2016). Disagreement About Moral Character Is Linked to Interpersonal Costs. *Social Psychological and Personality Science, 7*, 806-817.

Barron, K. (2016). Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? (No. SP II 2016-309). *WZB Discussion Paper*.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1-48.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323-370.

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives, 30*, 141-64.

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field? Evidence from donations. *Experimental Economics, 11*, 268-281.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122-142.

Berinsky, A. J. (2018). Telling the truth about believing the lies? Evidence for the limited prevalence of expressive survey responding. *The Journal of Politics, 80*, 211-224.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*, 351-368.

Bermúdez, J. P. (2018). *The Post-Truth Temperament: What Makes Belief Stray from Evidence? And What Can Bring Them Back Together? America's Post-Truth Phenomenon: When Feelings and Opinions Trump Facts and Evidence*, 87-109.

Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods*, 1-7.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bilewicz, M., & Vollhardt, J. R. (2012). *Evil Transformations: Social-Psychological Processes Underlying Genocide and Mass Killing*. In A. Golec de Zavala & A. Cichocka (Eds.), Social psychology of social problems: The intergroup context (pp. 280-307). New York, NY: Palgrave MacMillan.

Biziou-van-Pol, L., Haenen, J., Novaro, A., Occhipinti-Liberman, A., & Capraro, V. (2015). Does telling white lies signal pro-social preferences? *Judgment and Decision Making, 10*, 538-548.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41*, 540-558.

Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology, 76*, 420-430.

Böhm, R., Rusch, H., & Gürerk, Ö. (2016). What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior, 37*, 29-34.

Bolin, J. L., & Hamilton, L. C. (2018). The News You choose: news media preferences amplify views on climate change. *Environmental Politics, 27*, 455-476.

Bolsen, T., Druckman, J. N., & Cook, F. L. (2015). Citizens', scientists', and policy advisors' beliefs about global warming. *The ANNALS of the American Academy of Political and Social Science, 658*, 271-295.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90*, 166-193.

Boudry, M., Vlerick, M., & McKay, R. (2015). Can evolution get us off the hook? Evaluating the ecological defence of human rationality. *Consciousness and Cognition, 33*, 524-535.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389-414.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*, 7313-7318.

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science, 23*, 27-34.

Brandt, M., Sibley, C. G., & Osborne, D. (2018). What Is Central to Belief System Networks? Retrieved from https://psyarxiv.com/tyr64

Brauer, M., & Curtin, J. J. (2017). Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods, 23*, 389-411.

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An Economic Model of Moral Motivation. *Journal of Public Economics, 87*, 1967-1983.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*, 429-444.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2018). Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. *Journal of Applied Research in Memory and Cognition*.

Brown, J. D. (2012). Understanding the better than average effect motives (still) matter. *Personality and Social Psychology Bulletin, 38*, 209-219.

Brülhart, M., & Usunier, J. C. (2012). Does the trust game measure trust? *Economics Letters, 115*, 20-23.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3-5.

Bullock, J. G. & Gerber, A. S., Hill, S. J. & Huber, G. A. (2015). Partisan Bias in Factual Beliefs about Politics. *Quarterly Journal of Political Science, 10*, 519-578.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550.

Buttel, F. M., & Flinn, W. L. (1978). The politics of environmental concern: The impacts of party identification and political ideology on environmental attitudes. *Environment and Behavior, 10*, 17-36.

Camerer, C. F. *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press (2003).

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M. ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637-644.

Campbell, W. K., Rudich, E. A., & Sedikides, C. (2002). Narcissism, self-esteem, and the positivity of self-views: Two portraits of self-love. *Personality and Social Psychology Bulletin, 28*, 358-368.

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2018). People Make the Same Bayesian Judgment They Criticize in Others. *Psychological Science*.

Capraro, V. (2013). A model of human cooperation in social dilemmas. *PLoS ONE, 8*, e72427.

Capraro, V. (2018). Social versus Moral preferences in the Ultimatum game: A theoretical model and an experiment. Available at ArXiv https://arxiv.org/ftp/arxiv/papers/1804/1804.01044.pdf

Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making, 13*, 99-111.

Capraro, V., & Vanzo, A. (2018). Understanding moral preferences using sentiment analysis. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186134

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior, 35*, 43-63.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156-2160.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130*, 813-838.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53-81.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*, 112-130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*, 1131-1139.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics, 117*, 817-869.

Cheng, I. H., & Hsiaw, A. (2017). Distrust in Experts and the Origins of Disagreement. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2864563

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics, 2*, 2053168015622072.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85*, 808-822.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science, 8*, 160-179.

Coppock, A. (2016). *Positive, small, homogeneous, and durable: Political persuasion in response to information*. PhD Dissertation: Columbia University.

Coppock, A. (2018). Generalizing from survey experiments conducted on mechanical Turk: A replication approach. *Political Science Research and Methods*, 1-16.

Coppock, A. E, & McClellan, O. A. (2017). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. Unpublished Manuscript. Retrieved from https://alexandercoppock.com/papers/CM_lucid.pdf

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). The Generalizability of Heterogeneous Treatment Effect Estimates Across Samples. *Proceedings of the National Academy of Sciences*.

Corner, A., Whitmarsh, L., & Xenias, D. (2012). Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation. *Climatic change*, *114*, 463-478.

Corns, J. (2018). Rethinking the Negativity Bias. *Review of Philosophy and Psychology*, 1-19.

Coutts, A. (2018). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 1-27.

Cragun, J. (2018). Intuition and Reflection as Sources of Individual Differences in Selective Exposure to Attitude-Congruent Political Information. Retrieved from http://jamescragun.com/research/Cragun_reflection_selective_exposure.pdf

Crawford, J. T., & Pilanski, J. M. (2014). Political intolerance, right and left. *Political Psychology, 35*, 841-851.

Crockett, M. (2016). *Deal or no deal? Brexit and the allure of self-expression*. Retrieved from https://www.theguardian.com/science/head-quarters/2016/jul/05/deal-or-no-deal-brexit-and-the-allure-of-self-expression

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*, 1082-1089.

Davis, C. J., Bowers, J. S., & Memon, A. (2011). Social influence in televised election debates: A potential distortion of democracy. *PloS One, 6*, e18154.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics, 127*, 1-56.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). MTurk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233954

Digital, Culture, Media, and Sports Committee. (2018). Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf

Ditto, P. H. (2009). P*assion, reason, and necessity: A quantity-of-processing view of motivated reasoning*. Delusion and self-deception: Affective and motivational influences on belief formation, 23-53.

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*, 568-584.

Ditto, P. H., Clark, C. J., Liu, B. S., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2018b). Partisan bias and its discontents. Unpublished manuscript.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... & Zinger, J. F. (2018a). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. *Perspectives on Psychological Science*.

Douglas, K. M., & Sutton, R. M. (2015). Climate change: Why the conspiracy theories are dangerous. *Bulletin of the Atomic Scientists, 71*, 98-106.

Dowle, M. & Srinivasan, A. (2017). Data.table: Extension of 'data.frame'. R package version 1.10.4-3. https://CRAN.R-project.org/package=data.table

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social contexts? Framing effects in dictator games. *Experimental Economics, 16*, 349-371.

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences, 114*, 9587-9592.

Dunning, D. (2007). Self-image motives and consumer behavior: how sacrosanct self-beliefs sway preferences in the marketplace. *Journal of Consumer Psychology, 17*, 237-249.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090.

Ehret, P. J., Sparks, A. C., & Sherman, D. K. (2017). Support for environmental protection: an integration of ideological-consistency and information-deficit models. *Environmental Politics, 26*, 253-277.

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics, 3*, 114-138.

Engel, C. (2011). Dictator games: A Meta study. *Experimental Economics, 14*, 583-610.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *The American Economic Review, 94*, 857-869.

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction? *Journal of Personality and Social Psychology, 79*, 861-875.

Epley, N., & Dunning, D. (2006). The mixed blessings of self-knowledge in behavioral prediction: Enhanced discrimination but exacerbated bias. *Personality and Social Psychology Bulletin, 32*, 641-655.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science, 58*, 723-733.

Eriksson, K., Strimling, P., Anderson, P. A., & Lindholm, T. (2017). Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *Journal of Experimental Social Psychology, 69*, 59-64.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*, 5-34.

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295-306.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness – Intentions matter. *Games and Economic Behavior, 62*, 287-303.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives, 10*, 103-118.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137-140.

Fehr, E., & Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics, 95*, 1144-1155.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817-868.

Feldman, G., & Albarracín, D. (2017). Norm theory and the action-effect: The role of social norms in regret following action and inaction. *Journal of Experimental Social Psychology, 69*, 111-120.

Festinger, L. (1962). *A theory of cognitive dissonance*. Redwood City, CA: Stanford university press.

Fetchenhauer, D., & Dunning, D. (2006). *Perceptions of prosociality and solidarity in self and others*. In D. Fetchenhauer, A. Flache, B. Buunk, & S. Lindenberg (Eds.), Solidarity and prosocial behavior (pp. 61-74). New York, NY: Springer.

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology, 30*, 263-276.

Fetchenhauer, D., & Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science, 21*, 189-193.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*, 77-83.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software, 8*, 1-27. http://www.jstatsoft.org/v08/i15/

Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics, 16*, 155-169.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*, 25-42.

Galizzi, M. M., & Navarro-Martinez, D. (2017). On the external validity of social preference games: a systematic lab-field study. *Management Science*.

Gangarosa, E. J., Galazka, A. M., Wolfe, C. R., Phillips, L. M., Gangarosa, R. E., Miller, E., & Chen, R. T. (1998). Impact of anti-vaccine movements on pertussis control: The untold story. *The Lancet, 351*, 356–361.

Garrett, N. & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*.

Garrett, K. N., & Bankert, A. (2018). The moral roots of partisan division: How moral conviction heightens affective polarization. *British Journal of Political Science*, 1-20.

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review, 77*, 167-187.

Gebauer, J. E., Wagner, J., Sedikides, C., & Neberich, W. (2013). Agency-communion and self-esteem relations are moderated by culture, religiosity, age, and sex: Evidence for the "self-centrality breeds self-enhancement" principle. *Journal of Personality, 81*, 261-275.

Gerber, A. S., & Green, D. (1999). Misperceptions about perceptual bias. *Annual Review of Political Science, 2,* 189-210.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gershman, S. J. (2018). How to never be wrong. *Psychonomic Bulletin & Review*, 1-16.

Giner-Sorolla, R., Leidner, B., & Castano, E. (2012). *Dehumanization, demonization, and morality shifting: Paths to moral certainty in extremist violence*. In M. A. Hogg & D. L. Blaylock (Eds.), Extremism and the psychology of uncertainty (pp. 165-182). Chichester, UK: Wiley-Blackwell.

Ginges, J., Atran, S., Sachdeva, S. & Medin, D. (2011) Psychology out of the Laboratory: The Challenge of Violent Extremism. *American Psychologist. 66*, 507-519.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*, 153–172.

Goerg, S. J., Rand, D. G., & Walkowitz, G. (2017). Framing effects in the Prisoner's dilemma but not in the dictator game. Available at SSRN https://ssrn.com/abstract=2912982

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

Gotthard-Real, A. (2017). Desirability and information processing: An experimental study. *Economics Letters, 152*, 96-99.

Graham, D. A. (2016, 17 October). Republicans have been 'rigging' elections for years. Retrieved from http://www.theatlantic.com/politics/archive/2016/10/trump-election-rigged/504347/

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029-1046.

Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology, 8*, 125-130.

Graham, M., & Coppock, A. (2018). The Perils of Self-Assessed Attitude Change. Available from https://alexandercoppock.com/papers/GC_perils.pdf

Green, J. & Kapur, S. (2016, 22 June). Nearly half of Sanders supporters won't support Clinton: Poll. Retrieved from http://www.bloomberg.com/politics/articles/2016-06-22/nearly-half-of-sanders-supporters-won-t-support-clinton

Grossman, Z. (2010). Self-signaling versus social-signaling in giving. Available from https://escholarship.org/uc/item/7320x2cp

Hahn, U., & Harris, A. J. (2014). *What does it mean to be biased: Motivated reasoning and rationality?* In Psychology of learning and motivation (Vol. 61, pp. 41-102). Academic Press.

Hahn, U., Merdes, C., & von Sydow, M. (2018). How good is your evidence and how would you know? *Topics in Cognitive Science*.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "outgroup hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science, 19*, 405-411.

Halperin, E. (2008). Group-based hatred in intractable conflict in Israel. *Journal of Conflict Resolution, 52*, 713-736.

Hamilton, L. C. (2008). Who cares about Polar Regions? Results from a survey of US public opinion. *Arctic, Antarctic, and Alpine Research, 40*, 671-678.

Hamilton, L. C. (2011). Education, politics and opinions about climate change evidence for interaction effects. *Climatic Change, 104*, 231-242.

Hamilton, L. C., & Keim, B. D. (2009). Regional variation in perceptions about climate change. *International Journal of Climatology: A Journal of the Royal Meteorological Society, 29*, 2348-2352.

Hamilton, L. C., & Safford, T. G. (2015). Environmental views from the coast: Public concern about local to global marine issues. *Society & Natural Resources, 28*, 57-74.

Hamilton, L. C., & Saito, K. (2015). A four-party view of US environmental concern. *Environmental Politics, 24*, 212-227.

Hamilton, L. C., & Stampone, M. D. (2013). Blowin' in the wind: Short-term weather and belief in anthropogenic climate change. *Weather, Climate, and Society, 5*, 112-119.

Hamilton, L. C., Bell, E., Hartter, J., & Salerno, J. D. (2018). A change in the wind? US public views on renewable energy and climate compared. *Energy, Sustainability and Society, 8*, 11.

Hamilton, L. C., Colocousis, C. R., & Duncan, C. M. (2010). Place effects on environmental views. *Rural Sociology, 75*, 326-347.

Hamilton, L. C., Cutler, M. J., & Schaefer, A. (2012). Public knowledge and concern about polar-region warming. *Polar Geography, 35*, 155-168.

Hamilton, L. C., Hartter, J., & Saito, K. (2015). Trust in scientists on climate change and vaccines. *Sage Open, 5*, 2158244015602752.

Hamilton, L. C., Hartter, J., Lemcke-Stampone, M., Moore, D. W., & Safford, T. G. (2015). Tracking public beliefs about anthropogenic climate change. *PLoS One, 10*, e0138208.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243-1248.

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding Short Form (BIDR-16). *SAGE Open, 5*, 1-9.

Hart, P. S., Nisbet, E. C., & Myers, T. A. (2015). Public attention to science and political news and support for climate change mitigation. *Nature Climate Change, 5*, 541-545.

Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleeson, W. (2016). Morality's centrality to liking, respecting, and understanding others. *Social Psychological and Personality Science, 7*, 648-657.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*, 81-91.

Heck, P. R., & Krueger, J. I. (2015). Self-enhancement diminished. *Journal of Experimental Psychology: General, 144*, 1003-1020.

Heck, P. R., & Krueger, J. I. (2016). Social perception of self-enhancement bias and error. *Social Psychology, 47*, 327-339.

Heck, P. R., & Krueger, J. I. (2017). Social Perception in the Volunteer's Dilemma: Role of Choice, Outcome, and Expectation. *Social Cognition, 35*, 497-519.

Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review, 11*, 4-27.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A. ... & Lesorogol, C. (2006). Costly punishment across human societies. *Science, 312*, 1767-1770.

Hilbig, B. E., Zettler, I., Leist, F., & Heydasch, T. (2013). It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences, 54*, 598-603.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. *The Journal of Politics, 79*, 1403-1418.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221-234.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences, 112*, 10321-10324.

Hornikx, J., Harris, A. J., & Boekema, J. (2018). How many laypeople holding a popular opinion are needed to counter an expert opinion? *Thinking & Reasoning, 24*, 117-128.

Horton, J. J., Rand, D. G., Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*, 399-425.

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a Class of Permutation Tests: The coin Package. *Journal of Statistical Software, 28*, 1-23.

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics, 79*, 269-283.

Huck, S., Kübler, D., & Weibull, J. (2012). Social norms and economic incentives in firms. *Journal of Economic Behavior and Organization, 83*, 173-185.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*, 690-707.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. (2018). The Origins and Consequences of Affective Polarization. *Annual Review of Political Science*, 21.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology social identity perspective on polarization. *Public Opinion Quarterly, 76*, 405-431.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*, 521-537.

JASP Team (2017). JASP (Version 0.8.1.1) [Computer software].

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32*, 865-889.

Jolley, D. & Douglas, K. M. (2014a). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLoS One, 9*, e89177.

Jolley, D. & Douglas, K. M. (2014b). The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology*, *105*, 35–56.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016a). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473-476.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016b). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences, 113*, 8658-8663.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science, 28*, 356-368.

Joslyn, M. R., & Haider-Markel, D. P. (2014). Who knows best? Education, partisanship, and contested facts. *Politics & Policy, 42*, 919-947.

Joslyn, M. R., & Sylvester, S. M. (2017). The Determinants and Consequences of Accurate Beliefs about Childhood Vaccinations. *American Politics Research*, 1532673X17745342.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology, 38*, 167-208.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology, 60*, 307-337.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*, 339-375.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin, 129*, 339-375.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*.

Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology, 36*, 1-43.

Kahan, D. M. (2016a). *The Politically Motivated Reasoning Paradigm, Part I: What politically motivated reasoning is and how to measure it*. In R. Scott & S. Kosslyn (Eds.), Emerging Trends in Social and Behavioral Sciences (pp. 1-16).

Kahan, D. M. (2016b). *The Politically Motivated Reasoning Paradigm, Part 2: Unanswered questions*. In R. Scott & S. Kosslyn (Eds.), Emerging Trends in Social and Behavioral Sciences (pp. 1-15).

Kahan, D. M. (2017). The expressive rationality of inaccurate perceptions. *Behavioral and Brain Sciences*, 40.

Kahan, D. M., & Corbin, J. C. (2016). A note on the perverse effects of actively open-minded thinking on climate-change polarization. *Research & Politics, 3*, 2053168016676705.

Kahan, D. M., & Stanovich, K. (2016). Rationality and belief in human evolution. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2838668

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy, 1*, 54-86.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*, 732-735.

Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review, 97*, 1774-1793.

Kay, A. C., & Ross, L. (2003). The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the Prisoner's Dilemma. *Journal of the Experimental Social Psychology, 39*, 634-643.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217.

Khanna, K., & Sood, G. (2018). Motivated Responding in Studies of Factual Learning. *Political Behavior, 40*, 79-101.

Kim, J. W. (2018). Evidence can change partisan minds: Rethinking the bounds of motivated reasoning. Retrieved from https://jinwookimqssdotcom.files.wordpress.com/2018/10/kim_ws.pdf

Kim, S. (2015). Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients. *Communications for Statistical Applications and Methods, 22*, 665-674.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 608-638.

Kinder, D. R., & Kalmoe, N. P. (2017). *Neither liberal nor conservative: Ideological innocence in the American public*. University of Chicago Press.

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin, 25*, 586-595.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*, 852-884.

Klein, N., & Epley, N. (2016). Maybe holier, but definitely less evil, than you: Bounded self-righteousness in social judgment. *Journal of Personality and Social Psychology, 110*, 660-674.

Klein, N., & Epley, N. (2017). Less evil than you: Bounded self-righteousness in character inferences, emotional reactions, and behavioral extremes. *Personality and Social Psychology Bulletin*.

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2017). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, 0093650217719596.

Koehler, J. J. (1993). The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality. *Organizational Behavior and Human Decision Processes, 56*, 28-55.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality, 46*, 184-194.

Koonz, C. (2003). *The Nazi conscience*. Cambridge, MA: Harvard University Press.

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine, 46*, 575-587.

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience, 32*, 16832-16844.

Krueger, J. I. (2008). From social projection to social behaviour. *European Review of Social Psychology, 18*, 1-35.

Krueger, J. I., & Acevedo, M. (2007). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *The American Journal of Psychology, 120*, 593-618.

Krueger, J. I., & Chen, L. J. (2014). The first cut is the deepest: Effects of social projection and dialectical bootstrapping on judgmental accuracy. *Social Cognition, 32*, 315-336.

Krueger, J. I., & DiDonato, T. E. (2010). Perceptions of morality and competence in (non-) interdependent games. *Acta Psychologica, 134*, 85-93.

Krueger, J. I., Freestone, D., & MacInnis, M. L. (2013). Comparisons in research and reasoning: Toward an integrative theory of social induction. *New Ideas in Psychology, 31*, 73-86.

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence to social norms. *Negotiation and Conflict Management Research, 1*, 31-52.

Krueger, J. I., & Wright, J. C. (2011). *Measurement of self-enhancement (and self-protection)*. In M. D. Alicke & C. Sedikides (Eds.), Handbook of self-enhancement and self-protection (pp. 472-494). New York, NY: Guilford.

Kruger, J. (1999). Lake Wobegon be gone! The" below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*, 221-232.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association, 11*, 495-524.

Kuhn, D., & Lao, J. (1996). Effects of evidence on attitudes: Is polarization the norm? *Psychological Science, 7*, 115-120.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*, 1-26.

Larrick, R. P., & Blount, S. (1997). The claiming effect: Why players are more generous in social dilemmas than in ultimatum games. *Journal of Personality and Social Psychology, 72*, 810-825.

Lazear, E. P., & Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics, 4*, 136-163.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F. ... & Schudson, M. (2018). The science of fake news. *Science, 359*, 1094-1096.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*, 234-249.

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the need to belong scale: Mapping the nomological network. *Journal of Personality Assessment, 95*, 610-624.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology, 113*, 254-261.

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. *The Journal of Politics, 79*, 485-501.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*, 153-174.

Lewandowsky, S., & Oberauer, K. (2016). Motivated Rejection of Science. *Current Directions in Psychological Science, 25*, 217-222.

Lewandowsky, S., Oberauer, K. & Gignac, G. E. (2013). NASA Faked the Moon Landing--Therefore, (Climate) Science Is a Hoax: An Anatomy of the Motivated Rejection of Science. *Psychological Science, 24*, 622–633.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin, 30*, 1175-1185.

Liebrand, W. B. G., Jansen, R. W. T. L., Rijken, V. M., & Suhre, C. J. M. (1986). Might over morality: Social values and the perception of other players in experimental games. *Journal of Experimental Social Psychology, 22*, 203-215.

Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour, 2*, 166-167.

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.

Lüdecke D (2018). SjPlot: Data Visualization for Statistics in Social Science. https://CRAN.R-project.org/package=sjPlot

Malka, A., Krosnick, J. A., & Langer, G. (2009). The association of knowledge with concern about global warming: Trusted information sources shape public thinking. *Risk Analysis: An International Journal, 29*, 633-647.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17*, 11-17.

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2018). Epistemic Spillovers: Learning Others' Political Views Reduces the Ability to Assess and Use Their Expertise in Non-political Domains. Available from SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3162009

Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly, 80*, 351-377.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology, 39*, 257-277.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2004). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization, 52*, 267-275.

McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly, 52*, 155-194.

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior, 31*, 309-319.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology, 21*, 103-122.

Meyer, A., Zhou, E., & Shane, F. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making, 13*, 246-259.

Mijovi-Prelec, D., & Prelec, D. (2010). Self-deception as self-signaling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*, 227-240.

Miller, A. G., McHoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology, 64*, 561-574.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54*, 1053-1090.

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence. National Bureau of Economic Research Working Paper No. 17014. Doi: 10.3386/w17014

Möller, J., & Savyon, K. (2003). Not very smart, thus moral: Dimensional comparisons between academic self-concept and honesty. *Social Psychology of Education, 6*, 95-106.

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: on being both better and worse than we think we are. *Journal of Personality and Social Psychology, 92*, 972-989.

Moritz, A. (2011). *Vaccine-nation: Poisoning the population, one shot at a time*. Ener-chi.com.

Moutsiana, C., Garrett, N., Clarke, R. C., Lotto, R. B., Blakemore, S. J., & Sharot, T. (2013). Human development of the ability to learn from bad news. *Proceedings of the National Academy of Sciences, 110*, 16396-16401.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*, 109-138.

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature*.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021.

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin, 23*, 636-653.

Newman, T. P., Nisbet, E. C., & Nisbet, M. C. (2018). Climate change, cultural cognition, and media effects: Worldviews drive news selectivity, biased processing, and polarized attitudes. *Public Understanding of Science*, 0963662518801170.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*, 1422-1425.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*, 1560-1563.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291-1298.

Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 1-23.

Nyhan, B. (2016). The challenge of false beliefs. Retrieved from https://www.isr.umich.edu/cps/events/Nyhan_20160613.pdf

Ogilvie, D. M. (1987). The undesired self: A neglected variable in personality research. *Journal of Personality and Social Psychology, 52*, 379-385.

O'Mara, E. M., & Gaertner, L. (2017). Does self-enhancement facilitate task performance? *Journal of Experimental Psychology: General, 146*, 442-455.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*, 184-188.

Park, N., Peterson, C., & Seligman, M. E. (2006). Character strengths in fifty-four nations and the fifty US states. *The Journal of Positive Psychology, 1*, 118-129.

Parker, M. T., & Janoff-Bulman, R. (2013). Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love". *Social Justice Research, 26*, 81-96.

Pennycook, G., & Rand, D. G. (2018a). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*.

Pennycook, G., & Rand, D. G. (2018b). Cognitive Reflection and the 2016 US Presidential Election. *Personality and Social Psychology Bulletin*.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods, 48*, 341-348.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current Directions in Psychological Science, 24*, 425-432.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34-72.

Pereira, A., & Van Bavel, J. (2018). Identity concerns drive belief in fake news. Retrieved from https://psyarxiv.com/7vc5d/

Pew Research Center (2016, 26 August). *Gun Rights vs. Gun Control*. Retrieved from http://www.people-press.org/2016/08/26/gun-rights-vs-gun-control/#party

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications, 5*:4939.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. New York, NY: Penguin.

Pleasant, A., & Barclay, P. (2018). Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychological Science*, 0956797617752642.

Prior, M., Sood, G., & Khanna, K. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science, 10*, 489-518.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review, 83*, 1281-1302.

Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics, 114*, 37-82.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172-179.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*, 413-425.

Rand, D. G., Brescoll, V. L., Everett, J. A. C., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General, 145*, 389-396.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*, 3677.

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass, 2*, 1313-1344.

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research. https://CRAN.R-project.org/package=psych

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*, 32-47.

Roccas, S., & Brewer, M. B. (2002). Social identity complexity. *Personality and Social Psychology Review, 6*, 88-106.

Rodriguez, C. G., Moskowitz, J. P., Salem, R. M., & Ditto, P. H. (2017). Partisan selective exposure: The role of party, ideology and ideological extremity over time. *Translational Issues in Psychological Science, 3*, 254-271.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*, 27-42.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*, 296-320.

RStudio Team. (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com/

Rusch, H., Böhm, R., & Herrmann, B. (2016). Parochial Altruism: Pitfalls and Prospects. *Frontiers in Psychology, 7*, 1004.

Ryan, T. J. (2014). Reconsidering moral issues in politics. *The Journal of Politics, 76*, 380-397.

Ryan, T. J. (2017). No compromise: Political consequences of moralized attitudes. *American Journal of Political Science, 61,* 409-423.

Samuelson, W., Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty, 1,* 7-59.

Sarathchandra, D., Navin, M. C., Largent, M. A., & McCright, A. M. (2018). A survey instrument for measuring vaccine acceptance. *Preventive Medicine, 109,* 1-7.

Saucier, G., Akers, L. G., Shen-Miller, S., Knežević, G., & Stankov, L. (2009). Patterns of thinking in militant extremism. *Perspectives on Psychological Science, 4,* 256-271.

Schaffner, B. F., & Luks, S. (2018). Misinformation or Expressive Responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly, 82,* 135-147.

Scheitle, C. P. (2018). Politics and the Perceived Boundaries of Science: Activism, Sociology, and Scientific Legitimacy. *Socius, 4,* 2378023118769544.

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology, 65,* 317-338.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science, 3,* 102-116.

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. *Advances in Experimental Social Psychology, 29,* 209-269.

Sedikides, C., Gaertner, L., & CAI, H. (2015). *On the Panculturality of Self-enhancement and Self-protection Motivation: The Case for the Universality of Self-esteem*. In A.J. Elliot (Ed.), Advances in Motivation Science (pp. 185–241).

Sedikides, C., Meek, R., Alicke, M. D., & Taylor, S. (2014). Behind bars but above the bar: Prisoners consider themselves more prosocial than non-prisoners. *British Journal of Social Psychology, 53,* 396-403. Doi: 10.1111/bjso.12060

Shah, P., Harris, A. J. L., Bird, G., Catmur, C., & Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology, 90,* 71-127.

Shao, W., Keim, B. D., Garand, J. C., & Hamilton, L. C. (2014). Weather, climate, and the economy: explaining risk perceptions of global warming, 2001–10. *Weather, Climate, and Society, 6,* 119-134.

Sharot, T. & Garrett, N. (2016). Forming beliefs: why valence matters. *Trends in Cognitive Sciences, 20,* 25-33.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience, 14*, 1475-1479.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General, 141*, 423-428.

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review, 12*, 248–279.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.

Simons, D. J., Shoda, Y., & Lindsay, S. D. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, 1-6.

Simpson, B., Harrell, A., & Willer, R. (2013). Hidden paths from morality to cooperation: Moral judgments promote trust and trustworthiness. *Social forces, 91*, 1529-1548.

Simunovic, D., Mifune, N., & Yamagishi, T. (2013). Preemptive strike: An experimental study of fear-based aggression. *Journal of Experimental Social Psychology, 49*, 1120-1123.

Skitka, L. J., & Mullen, E. (2002). The dark side of moral conviction. *Analyses of Social Issues and Public Policy, 2*, 35-41.

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

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76-105.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision making, 13*, 260-267.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Stephan, W. G., Ybarra, O., & Morrison, K. R. (2011). *Intergroup threat theory*. In T. Nelson (Ed.), Handbook of Prejudice (pp. 43-55). New York, NY: Taylor & Francis Group.

Stern, P. C., Dietz, T., Abel, T. D., Guagnano, G. A., & Kalof, L. (1999). A value-belief-norm theory of support for social movements: The case of environmentalism. *Human Ecology Review, 6*, 81-97.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21,* 736-748.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making, 10,* 479-491.

Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods, 11*, 326-352.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition, 131,* 159-171.

Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science, 26,* 1469-1479.

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How people update beliefs about climate change: Good news and bad news. *Cornell L. Rev., 102,* 1431-1441.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755-769.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior, 31,* 137-155.

Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychology and Personality Science, 8,* 623-631.

Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General, 146,* 1143-1149.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin, 103,* 193-210.

Teper, R., & Inzlicht, M. (2011). Active transgressions and moral elusions: Action framing influences moral behavior. *Social Psychological and Personality Science, 2,* 284-288.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77,* 184-197.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11,* 99-113.

Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology, 44,* 187-194.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275-1289.

Torchiano, M. (2017). Package 'effsize'. Available at https://cran.r-project.org/web/packages/effsize/effsize.pdf

Trippas, D., Kellen, D., Singmann, H., Pennycook, G., Koehler, D. J., Fugelsang, J. A., & Dubé, C. (2018). Characterizing belief bias in syllogistic reasoning: A hierarchical Bayesian meta-analysis of ROC data. *Psychonomic Bulletin & Review*, *25*, 2141-2174.

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning, 21*, 431-445.

Trivers, R. L. (1971). The evolution of reciprocal altruism. T*he Quarterly Review of Biology, 46*, 35-57.

TurkPrime. (2018). *after the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it*. Retrieved from https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134*, 383-403.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An Identity-based model of political belief. *Trends in Cognitive Sciences*.

Van Damme, C., Deschrijver, E., Van Geert, E., & Hoorens, V. (2017). When Praising Yourself Insults Others: Self-Superiority Claims Provoke Aggression. *Personality and Social Psychology Bulletin*.

Van Damme, C., Hoorens, V., & Sedikides, C. (2016). Why self-enhancement provokes dislike: The hubris hypothesis and the aversiveness of explicit self-superiority claims. *Self and Identity, 15*, 173-190.

Van Lange, P. A., & Sedikides, C. (1998). Being more honest but not necessarily more intelligent than others: Generality and explanations for the Muhammad Ali effect. *European Journal of Social Psychology, 28*, 675-680.

Van Lange, P. A., & Sedikides, C. (1998). Being more honest but not necessarily more intelligent than others: Generality and explanations for the Muhammad Ali effect. *European Journal of Social Psychology, 28*, 675-680.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12.

Vazire, S., & Carlson, E. N. (2010). Self-Knowledge of Personality: Do People Know Themselves? *Social and Personality Psychology Compass, 4*, 605-620.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S* (4th Edition). New York, NY: Springer.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48.

Weber, J. M., Kopelman, S., & Messick, D. M. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. *Personality and Social Psychology Review, 8*, 281-307.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*, 806-820.

Weisel, O., & Böhm, R. (2015). "Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. *Journal of Experimental Social Psychology, 60*, 110-120.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One, 11*, e0152719.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*, 1-20.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software, 40*, 1-29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wickham, H. (2018). Scales: Scale Functions for Visualization. R package version 1.0.0. https://CRAN.R-project.org/package=scales

Wickham, H., François, R., Henry, L., & Müller, K. (2018). *Dplyr: A grammar of data manipulation*. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin, 90*, 245-271.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology, 16*, 155-188.

Wojciszke, B., & Bialobrzeska, O. (2014). Agency versus Communion as Predictors of Self-esteem: Searching for the Role of Culture and Self-construal. *Polish Psychological Bulletin, 45*, 469-479.

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A. E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology, 41*, 617-627.

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: does it explain modern human group psychology? *Current Opinion in Psychology, 7*, 39-43.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity. *Advances in Group Processes, 16*, 161-197.

Ybarra, O., Park, H., Stanik, C., & Lee, D. S. (2012). Self-judgment and reputation monitoring as a function of the fundamental dimensions, temporal perspective, and culture. *European Journal of Social Psychology, 42*, 200-209.

Yu, R., & Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology, 5*, 1574.

Zeelenberg, M., van den Bos, K., van Dijk, E., & Pieters, R. (2002). The inaction effect in the psychology of regret. *Journal of Personality and Social Psychology, 82*, 314-327.

Zell, E., & Alicke, M. D. (2011). Age and the Better-Than-Average Effect. *Journal of Applied Social Psychology, 41*, 1175-1188.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 1-5.