European Economic Review 172 (2025) 104929

Contents lists available at ScienceDirect

European Economic Review

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

Motivated political reasoning: On the emergence of belief-value constellations☆

Kai Barron [a], Anna Becker [b], Steffen Huck [c,a,*]

[a] WZB Berlin, Reichpietschufer 50, 10785, Berlin, Germany
[b] Stockholm University, 106 91 Stockholm, Sweden
[c] University College London, Drayton House, 30 Gordon St, WC1H 0AN, London, United Kingdom

ARTICLE INFO

JEL classification:
C90
D72
D74
D83
P16

Keywords:
Motivated beliefs
Values
Polarization
Experiment
Reasoning

ABSTRACT

We study the relationship between moral values ("ought" statements) and factual beliefs ("is" statements). We show that thinking about values affects the beliefs people hold. This effect is mediated by prior political leanings, thereby contributing to the polarization of factual beliefs. We document these findings in a pre-registered online experiment with a nationally representative sample of over 1,800 individuals in the US. We also show that participants do not distort their beliefs in response to financial incentives to do so, suggesting that deep values exert a stronger motivational force than financial incentives.

1. Introduction

Why did Republicans hold different beliefs about the dangers of COVID-19 compared to Democrats, leading them to take fewer precautions and thereby increasing their risk of contracting the illness? Both Allcott et al. (2020) and Clinton et al. (2021) document this startling divide in beliefs and behavior between political camps during the pandemic. This example is illustrative of the phenomenon of partisan "bubbles" comprising *disagreement about facts* along the political spectrum. This phenomenon is, however, much wider and can be traced back to, at least, the early 2000s in the United States (see, for example, Gaines et al., 2007, on polarized beliefs about the Iraq war). A common explanation for the emergence and persistence of these patterns is *politically motivated reasoning* (see e.g. Lord et al., 1979; Taber and Lodge, 2006; Kahan, 2016). Citizens exhibit a tendency to interrogate arguments and information that conflict with their prior partisan attitudes more vigorously, while uncritically accepting attitudinally congruent arguments.

Politically motivated reasoning has been discussed as a reason for partisan disagreement on key policy issues such as redistribution (Alesina et al., 2018) and immigration (Haaland and Roth, 2023; Alesina et al., 2022) and, more generally, as a cause of political polarization (Alesina et al., 2020; Ortoleva and Snowberg, 2015). Much of this extant literature is devoted to documenting

☆ The authors would like to thank Ciril Bosch-Rosa, Christoph Drobner, Nathan Hancart, Anna Kerkhof, Nina McMurry, Sebastian Schneider and Alice Solda for helpful comments and discussions. This research was approved by the UCL Research Ethics Committee (17181/001). The study was pre-registered on OSF (https://osf.io/8jydh/). Barron gratefully acknowledges financial support from the German Research Foundation (DFG), via CRC TRR 190 (project number 280092119). Open Access funding was provided by the WZB Berlin Social Science Center through the DEAL consortium.

* Corresponding author at: WZB Berlin, Reichpietschufer 50, 10785, Berlin, Germany.
E-mail addresses: kai.barron@wzb.eu (K. Barron), anna.becker@su.se (A. Becker), steffen.huck@wzb.eu (S. Huck).

https://doi.org/10.1016/j.euroecorev.2024.104929
Received 19 May 2023; Received in revised form 2 December 2024; Accepted 6 December 2024

Available online 21 December 2024

that beliefs are systematically polarized along political contour lines. However, research on the precise underlying mechanisms that generate these politically polarized beliefs is still scarce. One notable exception to this is Thaler (2024), who presents an innovative experimental study examining how political beliefs may influence the way that individuals draw inference about the veracity of a news source. He shows that individuals tend to trust news more when it aligns with the views of their political party. This reveals one important potential mechanism behind politically polarized beliefs, namely motivated trust in different news sources.

In this study, we examine another important potential mechanism by asking whether the act of *thinking about core values* causally influences the beliefs an individual holds. We conceptualize values as desires about the social world, that is, statements about how the world ought to be (moral positions). Unlike beliefs, which relate to the actual state of the world (how the world is), values cannot be objectively true or false—they can only be endorsed or opposed, with varying degrees of intensity. As previously shown by Alesina et al. (2020) and Enke (2020a), moral values and beliefs are highly correlated. Our study replicates this finding and provides evidence that the salience of values can causally affect factual beliefs. This occurs without the arrival of new information; rather, we examine whether simply encouraging an individual to think about their values can shift their beliefs. One potential channel for this effect is that making a value salient may distort how an individual retrieves information from their memory to reconstruct their factual beliefs. This distinguishes our study from the belief updating literature which examines distortions in information processing (e.g., Thaler, 2024, who uses a clever design where new information is uninformative for a Bayesian, but allows for motivated reasoning by non-Bayesians). Furthermore, we show that individuals distort their beliefs to be more in line with the value system predominant in their political camp. This suggests that political leanings mediate the relationship between values and beliefs that we uncover.[1]

We study the relationship between values and beliefs in the context of six different contentious policy domains: migration, animal welfare, gender equality, abortion, prostitution, and same-sex marriage. In each of these domains, we can test whether individuals adjust their beliefs to conform to the values of their political party. In addition, our experimental design allows us to analyze whether any shift in beliefs translates into a shift in actions. To do this, we give subjects the opportunity to donate money to charities operating in each of the relevant policy domains. Previous studies have shown that politically biased beliefs can have important consequences for individual behavior, for example, in public health (Allcott et al., 2020) or in financial decision-making (Meeuwis et al., 2022). Here, we can examine whether a causal shift in beliefs due to thinking about values also results in a change in donation behavior.

Our study provides direct evidence on several specific motivated reasoning channels shaping the relationship between values and beliefs. Thus, it also contributes to the broader motivated reasoning literature. In our context, motivated reasoning may operate in two ways. First, since an individual may view her values as an integral part of her identity, motivated reasoning may lead her to shift her factual beliefs so that they support her values. This way, she can maintain a self-image of being someone who holds values that are aligned with facts.[2] Second, when individuals anticipate that holding a certain belief will obligate them to take a costly action (in our context, a donation decision), they may shift their beliefs in the opposite direction to reduce the moral obligation to take that costly action.[3] We examine the role of the second channel by exogenously varying whether individuals anticipate a donation decision. To examine the role of persuasion in political belief formation, we test whether participants shift their beliefs when this can help them to persuade another person to make a specific charitable donation. In doing this, we add to recent findings that demonstrate that individuals may shift their own beliefs in order to be more convincing when attempting to persuade another person to believe something (see, e.g., Schwardmann and van der Weele, 2019; Solda et al., 2020; Schwardmann et al., 2022).

To explore these questions, we designed and conducted a preregistered online experiment in January 2020 that surveyed a nationally representative sample of 1863 individuals from the US population.[4] Our experiment employs a between-subject design and comprises four main treatments. These treatments are designed to test hypotheses centered around the following questions: (i) Are there systematic correlations between values and factual beliefs? (ii) Do individuals adjust their factual beliefs when a moral value in the same domain becomes more salient? (iii) Do individuals shift their stated values and factual beliefs to align with their own material self-interest? (iv) Do individuals alter their stated values and beliefs in an attempt to persuade others?

Our first two treatments address questions (i) and (ii) – whether there are correlations between values and beliefs and whether values do indeed causally shape beliefs. First, in treatment ValueNotSalient, we elicit subjective beliefs about factual statements from the six domains mentioned above. This treatment serves as a control for the second treatment, ValueSalient, where we additionally elicit subjects' agreement with value statements pertaining to the six domains prior to the belief elicitation.

The purpose of this elicitation of values is that it encourages subjects to think about them. This serves to raise the salience of these values when subjects subsequently report their associated beliefs. Of course, subjects may also be passively aware of their values in treatment ValueNotSalient, but their direct elicitation in ValueSalient should serve as a priming device that heightens value salience, bringing values to the forefront of the individual's mind. The underlying idea behind the comparison of the ValueSalient and

[1] The importance of conformity with one's preferred political party has previously been shown to be an important factor for politically motivated reasoning. For example, in Druckman et al. (2013) arguments in favor of, or against, a motion are shown to have a stronger effect on partisanship when the arguments are explicitly linked to party stances.

[2] For example, an individual who is pro-life might find it easier to believe the results of a study purporting to show that a fetus experiences pain and consciousness than someone who is pro-choice.

[3] For example, when an individual anticipates being asked to donate to a charity that supports animal welfare, they might convince themselves that the benefits of the donation are low, assuming that much of the money may not reach the animals.

[4] To provide some context, the experiment was, thus, designed and implemented prior to events such as the widespread awareness of COVID-19 (February 2020), the death of George Floyd (May 2020), the claims that the United States presidential election was rigged (November 2020) and the attack on the Capitol (January 2021).

Table 1
Overview of experimental design.

	Treatment				
	ValueNotSalient	ValueSalient	ConvinceSelf	ConvinceOther	BeingConvinced
Screen 1	Elicitation of **beliefs**	Elicitation of **values** & **beliefs**	Elicitation of **values** & **beliefs** & Option to **donate**	Elicitation of **values** & **beliefs** & Learn about BeingConvinced participant option to **donate**	Elicitation of **values** & **beliefs**
	Proceed to next screen as surprise	Proceed to next screen as surprise	End of experiment	End of experiment	Proceed to next screen as surprise
Screen 2	Option to **donate**	Option to **donate**			Option to **donate** (with info. from ConvinceOther participant)
Obs.	**375**	**385**	**377**	**363**	**363**

ValueNotSalient treatments is to examine whether individuals who are confronted with a highly salient value question are motivated to shift their factual beliefs to align them with their values.

It is important to note, however, that individuals may already be reporting beliefs that are distorted to align with their values in the ValueNotSalient treatment. This implies that what we are able to identify causally is the *additional* shift in beliefs due to exogenously making values more salient. We test this by comparing the distribution of beliefs reported in these two treatments. Since party preferences may mediate the direction of this effect, we examine this comparison when conditioning on party preferences. As outlined in our pre-analysis plan, we also report the results for the comparison for the full sample without conditioning.

In all our treatments, there is one final stage after the belief and value elicitations which explores how choices are influenced. Specifically, we give subjects the opportunity to donate money to charities that operate in each of the six domains. This final stage plays a key role for our next two treatments that address questions (iii) and (iv) – whether subjects are willing to distort their values and beliefs in order to convince themselves or to convince others. To test whether self-interest plays a role in shifting beliefs, we introduce our ConvinceSelf treatment. This treatment is identical to the ValueSalient treatment, except that the donation decision is placed on the same screen where we elicit beliefs and values (in other treatments, it comes as a surprise). We can therefore test whether subjects adjust their reasoning in a self-interested way when holding a particular belief-value constellation would point towards taking a costly action (i.e., making a donation to a charity whose work is aligned with that particular constellation).

Finally, we test whether subjects adjust their stated beliefs and/or values when they have an incentive to persuade others. In the fourth treatment, ConvinceOther, we again ask subjects to state their beliefs and values. However, here, rather than making the donation decision themselves as in the third treatment, they are informed that another participant will have the opportunity to donate after being shown their belief and value responses.

Table 1 provides an overview of the treatments and indicates the number of participants per treatment. We discuss the BeingConvinced treatment in more detail at the end of Section 3.1. This treatment is an auxiliary treatment that we implemented to avoid deception. It is not central to our hypotheses or results because participants in this group mostly served as receivers of information from the ConvinceOther treatment.

Turning to our results, we find support for the existence of aligned belief-value constellations in all the policy domains considered, thereby answering our first research question in the affirmative. Note, however, that while such correlations are evidence of partisanship, it is not possible to understand the mechanism behind this without further evidence. Such bubbles may arise when beliefs shape values, when there are filter bubbles or echo chambers (Flaxman et al., 2016; Enke, 2020b), or when values provide a sufficiently strong force for motivated reasoning. It is the comparison of our first two treatments that helps us to explore the latter mechanism.

On aggregate, the distributions of reported subjective beliefs are almost identical across the two treatments. At surface level, this appears to suggest that we do not observe a shift in subjects' beliefs when values are more salient. However, the picture changes dramatically when we control for individuals' political preferences. Specifically, we find that subjects on both the political right and the political left shift their beliefs to align them with the average beliefs held by those in their preferred political camp when

values are made more salient. We, thus, do find support for the idea that thinking about values shapes beliefs through motivated reasoning. This suggests that the heightened salience of contentious policy issues in public debates may be a key explanation for increasing polarization in factual beliefs along political attitude division lines.

With respect to our third and fourth research questions, we find that beliefs and values are unaffected by the addition of monetary incentives to persuade oneself or the anticipation of the opportunity to persuade another person. If anything, this lack of malleability of beliefs and values to other factors appears to suggest that our subjects care about responding honestly to our belief and value questions. This should lend credibility to the internal and external validity of our first two sets of results. These results are also consistent with a growing body of research documenting the limits of motivated reasoning.[5]

Related Literature: In studying the role played by values in the (motivated) formation of beliefs, we contribute to a growing literature on motivated cognition and wishful thinking. This literature has considered a wide array of factors that may generate motivated beliefs, including: maintaining a positive image of one's own intelligence, attractiveness, or performance (e.g., Eil and Rao, 2011; Möbius et al., 2022; Coutts, 2019; Drobner, 2022; Huffman et al., 2022), judging what is fair or morally appropriate in a self-interested fashion (e.g., Messick and Sentis, 1979; Babcock et al., 1995; Konow, 2000; Barron et al., 2024; Amasino et al., 2021), distorting one's own beliefs in order to be more persuasive to others (e.g., Schwardmann and van der Weele, 2019; Solda et al., 2020; Schwardmann et al., 2022), and engaging in confirmatory reasoning that reinforces one's prior beliefs (e.g., Nickerson, 1998; Rabin and Schrag, 1999).

This previous literature typically considers motivated reasoning in relation to a belief that is closely tied to an individual's self-interest, personal characteristics, or pre-existing beliefs about a particular topic. In contrast, here we examine whether deeper values may exert an influence over related factual beliefs. Given the extent to which many contentious political debates are driven by values, along with the substantial heterogeneity in values between and within societies, this strikes us as an important question. Belot and Briscese (2022) provide evidence suggesting that polarization in beliefs can be decreased when individuals are made aware that they share common values such as human rights or behavioral etiquette rules. Hence, while values might be a divisive factor when they shape belief formation related to contentious issues, they also appear to have the potential to reduce gaps between political camps. It is, thus, important to better understand the precise role they play in motivated political reasoning. Our results suggest that when individuals are prompted to think about contentious values, they adjust their factual beliefs to align their beliefs with those of individuals who share their values and political affiliation.

In relation to the *persuasion of others*, in contrast to findings documented in the literature, in our setting, we do not observe evidence that individuals adjust their beliefs to try to be persuasive. One potential explanation for this could be that individuals do not believe that persuasion operates via the transmission of simple statements of beliefs and, rather, requires a richer communication space. For example, individuals may believe that in order to be persuasive, they need to transmit *arguments* (Schwardmann et al., 2022), *explanations* (Graeber et al., 2024), or *narratives* (Barron and Fries, 2023). Since this richer communication space is ruled out in our design, individuals may not believe that they will be able to persuade others using the limited communication space available and, therefore, do not shift their own beliefs.

Our work also contributes empirical evidence to the recent theoretical discussion about how and why partisan individuals increasingly entertain polarized mental models of reality (Leeper and Slothuus, 2014; van Bavel and Pereira, 2018; Alesina et al., 2020). One important factor appears to be group identity. As discussed by Sherman and Cohen (2006), individuals tend to select policy-relevant information in a way that allows them to maintain beliefs that are consistent with the position of the group they identify with. The beliefs generated in this way then allow members of identity groups to express behavior that signals group membership and hence strengthens the ties with their own community. Given that liberals and conservatives have been shown to have moral systems based on different psychological foundations (Graham et al., 2009), values are likely to be a strong force underlying motivated reasoning. Bonomi et al. (2021) formulate a related idea and discuss a theory of identity politics where increasing the salience of a certain policy conflict leads individuals to identify more strongly with their cultural or economic group, and then to distort their beliefs towards the stereotypical belief of the group they identify with. Our results can also be interpreted in the light of this theoretical framework.[6] Similarly, our work also relates to the large literature investigating how individuals are influenced by social norms, which often results in a desire to conform with the behavior of one's in-group (see, e.g., Bernheim,

[5] In particular, several of the studies that examine whether belief updating is distorted by monetary incentives associated with different states of the world fail to find any influence of motivated reasoning (see, for example, Gotthard-Real, 2017; Coutts, 2019; Barron, 2021). Furthermore, Thaler (2020) convincingly shows an absence of positivity-motivated reasoning in domains where self-image is not present. A second strand of literature examines scenarios where individuals engage in motivated reasoning to justify or excuse their self-serving behavior in pursuit of monetary gain. This strand of work has revealed mixed results, with Di Tella et al. (2015) and Bicchieri et al. (2023) documenting evidence that individuals do distort their beliefs to justify self-serving behavior, while the evidence reported in Ging-Jehli et al. (2020) and Barron et al. (2024) is more mixed. Nevertheless, while these papers explore the motivated reasoning of individuals seeking monetary gains, the underlying motivation in these studies is linked to beliefs about their character. These motivated beliefs allow individuals to perceive themselves as less of a 'bad person'. Together, these results indicate that motivated reasoning operates in certain domains, with internal psychological factors such as self-image and deep values serving as a source for motivated reasoning, but external factors such as monetary rewards and others' well-being often do not result in motivated reasoning.

[6] While we focus predominantly on assessing the causal effect of value salience on beliefs, we also contribute to a broader literature that examines the influence of partisanship on information processing. For example, in the domains of energy policy and climate change respectively, Bolsen et al. (2014) and Druckman and McGrath (2019) examine the role played by partisan differences in information processing due to selectively trusting different *sources* of information. Kahan (2013) explores the role of different thinking styles in generating ideological polarization and Alesina et al. (2018) show that when individuals are provided with pessimistic information about mobility, left-wing individuals become more pessimistic about mobility and increase their demand for redistribution, but right-wing individuals do not. In our paper, individuals are not provided with any new information to process—they must form their beliefs based on the information already stored in their memory. We only vary the presence of a reason for motivated reasoning, such as the salience of a policy conflict.

1994; Goette et al., 2006; Bicchieri et al., 2022, 2023). While this body of work typically focuses on a pull towards conformity in behavior, we focus on the particular mechanism through which beliefs are drawn into tight constellations with members of one's political in-group.

The remainder of this paper is organized as follows. In Section 2, we describe the experimental setup and results pertaining to questions (i) and (ii) – whether there are correlations between values and beliefs and whether values do indeed causally shape beliefs. Section 3 proceeds with the description of the setup and results for questions (iii) and (iv) – whether subjects are willing to distort their values and beliefs in order to convince themselves or to convince others. Section 4 concludes.

2. Existence and formation of belief-value constellations

Our experimental design consists of four pre-registered[7] treatments that were conducted online using the platform Prolific with a nationally representative sample of 1863 individuals from the US population.[8] In our main analyses, we essentially follow our pre-registration plan and explicitly flag any minor deviations. In this section we focus on describing and analyzing the first two treatments, ValueSalient and ValueNotSalient, which allow us to ask: *(i) Do individuals display belief-value constellations? (in the sense of observing a correlation between beliefs and values), and (ii) Do individuals adjust their beliefs to be more coherent with their values?*

2.1. The ValueSalient and ValueNotSalient treatment conditions

The first objective of our ValueSalient treatment is to examine whether we observe a systematic correlation between values and associated factual beliefs. The treatment consists of three parts. First, participants are presented with a sequence of six (randomly ordered) moral value statements and are asked to report the degree to which they agree or disagree with each statement. Table 2 provides an overview of the six different debates together with the moral and factual statements presented to participants. Each of these moral value statements corresponds to a particular contentious topic of debate in public policy, such as gender equality, abortion, or same-sex marriage. The items serve to raise the salience of these value debates such that participants might view later questions through the lens of those debates. It is important to note that we focus on values that lack consensus, which implies that all our topics are inherently political. Hence, raising the salience of a particular value debate may also raise the salience of a political issue. Second, participants are confronted with six factual statements pertaining to the same six domains, and we elicit their beliefs about the veracity of these factual statements.

Both values and beliefs are measured in 5-point Likert scales. For values, responses range from "strongly agree" to "strongly disagree", and for factual beliefs, from "Very Unlikely" to "Very Likely".[9] The debates that we consider relate to migration, animal welfare, gender equality, abortion, prostitution, and gay rights. Together, the moral value assessments and factual belief reports allow us to examine whether there is a correlation between individuals' beliefs and values.

In the last step, we provide participants with the opportunity to make six donation decisions to six separate charities (one of which is randomly implemented). Each of the six charities targets a cause that corresponds to one of the six relevant public policy discussions.[10] For each charity, participants are asked to divide $3 between the charity and themselves. In a post-experimental survey, we also collected information on the participants' political attitudes, and we were able to match our data to previously elicited political attitude variables collected by Prolific independently from our experiment.[11] This allows us examine how values and beliefs translate into actions.

The ValueNotSalient treatment is identical to the ValueSalient treatment, with the exception that the first stage in which participants are presented with moral value statements is skipped. This implies that in this treatment the six public policy debates are not made as salient. The exogenous variation in salience between the two treatments allows us to assess how the shift in salience causally affects factual beliefs. Figures A.1 and A.2 in the Online Appendix show the instructions as they were presented to participants in both treatments. Importantly, we ask participants about their beliefs concerning factual statements for which there is no clear scientific consensus. We do this because such beliefs are more likely to foster motivated reasoning as participants do not expect the uncertainty to be resolved. This allows participants to derive anticipatory utility from their beliefs without worrying about learning that they were "wrong". Drobner (2022) demonstrates that motivated reasoning is more prevalent in scenarios where individuals do not expect a resolution of uncertainty.

To measure beliefs about factual statements, we ask subjects, "How likely do you think it is that the following statement is true?"; for moral statements, we ask, "How much do you agree with the following statement?" This reflects that for facts, there is, in principle, an ascertainable truth, while values can only be desirable or undesirable to different degrees.

To fix ideas, let us consider two examples of statement pairs: the first from the migration domain and the second from the animal welfare domain. The statement "All countries benefit from the free movement of labor" pertains to a fact that may be either true

[7] The full pre-registration document can be found at https://osf.io/8jydh/ and is also reproduced in Appendix D in the Supplementary Material.

[8] Table C1 in the Online Appendix shows that our sample is balanced between all the treatments which are described in the following.

[9] In order to keep the experiment as clear, simple, and easy-to-understand as possible, we opted not to incentivize the elicitation of factual beliefs. This decision draws on the evidence discussed in Haaland et al. (2023) and Stantcheva (2023), indicating that the incentivization of beliefs in this type of study does not necessarily lead to improvements in truth-telling and may distract participants from the key questions of interest.

[10] Subjects are provided with information about the aims of the charities and use a slider to indicate how much they would like to donate. Further details about the charities can be found in Table D2 in Online Appendix D.

[11] We provide summary statistics for the main variables from the experiment for all treatments in Table B1 in the Online Appendix.

Table 2
Overview over statements and charities.

	Debate	Moral statement	Factual statement	Donation
		"How much do you agree with the following statement?"	*"How likely do you think it is that the following statement is true?"*	Charity
1	Migration	People should be allowed to migrate freely between countries.	All countries benefit economically from the free movement of labor.	American Immigration Council
2	Animal welfare	It is wrong to eat animals.	Animals feel less pain than humans.	World Animal Protection
3	Gender equality	Gender equality should be an objective of policymaking.	Discrimination against women is the primary reason why women earn less than men.	Equality Now
4	Abortion	Abortion should be legal.	Women who have had an abortion experience more psychological distress than women who have had a miscarriage.	Planned Parenthood
5	Prostitution	Prostitution should be illegal.	Human trafficking is facilitated by liberal prostitution laws.	A21
6	Same-sex marriage	Gay couples should have the same rights as heterosexual couples.	Societies where same-sex marriage is legal are happier than societies where it is illegal.	OutRight

or false. While its veracity might be difficult to ascertain, it is, in principle, ascertainable. In contrast, the statement "People should be allowed to migrate freely between countries" expresses a desire. One may or may not agree with the statement, but there is no truth to be ascertained. The same applies to the statements "Animals feel less pain than humans" (belief) and "It is wrong to eat animals" (value).

2.2. The existence of belief-value constellations

The first question we seek to answer is whether there is alignment between the moral values, factual beliefs, and political attitudes that individuals hold. This would indicate the presence of "belief-value constellations". There are several potential reasons why individuals might have aligned beliefs and values, including: (i) holding values that are shaped by beliefs about facts; (ii) *avoiding cognitive dissonance* from holding incoherent values and beliefs; or (iii) using value and belief statements to justify self-interested actions (i.e., *motivated reasoning*). Our second hypothesis will explore one particular possibility: that beliefs are shaped by values.

Our first set of hypotheses tests whether belief-value constellations are observed systematically in the population.[12]

Hypothesis 1 (*Belief-Value Constellations*)**.** *There is a correlation between the beliefs, values, and political attitudes that individuals hold. The actions individuals take are aligned with their belief-value constellations.*

Let b_t denote the factual beliefs stated by individuals in Treatment $t \in \{VS, VNS\}$, v_t the moral values stated by individuals, d_t their donation decisions and p_t the left–right political stance of individuals.

(a) Moral values are positively correlated with beliefs:

$$Corr(v_{VS}, b_{VS}) \geq 0.$$

(b) Moral values are negatively correlated with political attitudes:

$$Corr(v_{VS}, p_{VS}) \leq 0.$$

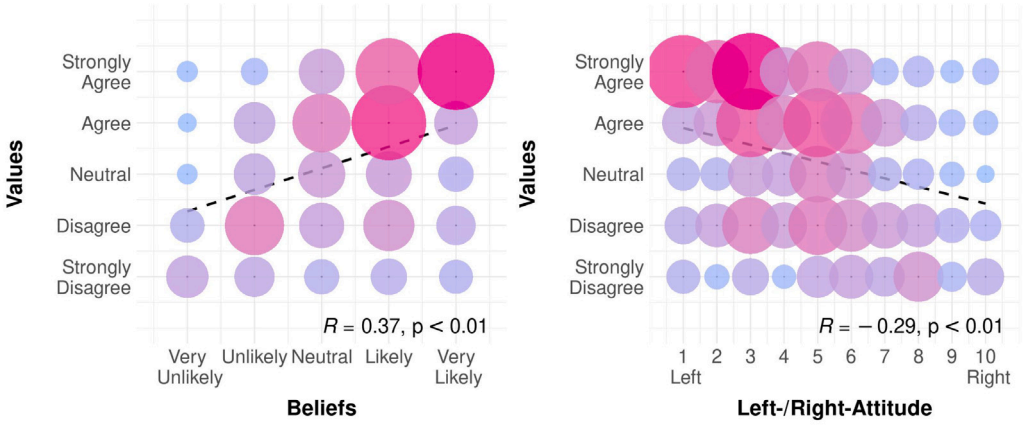
(c) Donations are positively correlated with beliefs and values:

$$Corr(d_{VNS}, b_{VNS}) \geq 0,$$

$$Corr(d_{VS}, b_{VS}) \geq 0, Corr(d_{VS}, v_{VS}) \geq 0.$$

When reading Hypothesis 1, it is important to take note of the way that the variables are encoded. First, the political stance variables, p_t, are constructed to be increasing in the degree to which an individual positions herself on the right of the political spectrum. Second, the moral value, v_t, variables are encoded such that a high value indicates agreement with a value that is typically associated with individuals on the left of the political spectrum. Third, the factual belief variables, b_t, are defined such that if they are true, they would provide empirical support for moral value positions typically held by individuals on the political left. Finally,

[12] In the interest of facilitating a more coherent exposition of the paper and to enhance readability, we have adjusted the formulation of the hypotheses in comparison to the pre-registration document. We encourage the interested reader to refer to the full pre-registration document in Online Appendix D for further details.



(a) Beliefs and Values (Hyp. 1a) (b) Political Attitudes and Values (Hyp. 1b)

Fig. 1. Results for Hypothesis 1a and b.
Note: Fig. 1(a) shows the results for Hypothesis 1a, i.e. the correlation between values and beliefs in the ValueSalient treatment. Fig. 1(b) shows the results for Hypothesis 1b, i.e. the correlation between moral values and political attitudes in the ValueSalient treatment. The data points are weighted by the number of observations, which is reflected in both the color and size of the markers: the larger and redder the marker, the more observations; the smaller and bluer the marker, the fewer observations. The dotted line represents the result of a linear regression of values on beliefs in Fig. 1(a) and a regression of values on political attitudes in Fig. 1(b). The Spearman correlation coefficient, R, and its *p*-value are reported at the bottom right of each graph. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the charitable donation variables, d_t, are constructed such that higher donations are consistent with costly support of a charity aligned with the relevant moral value position.

Results (Hypothesis 1)

Fig. 1 summarizes the results pertaining to Hypothesis 1. First, the top left panel reports the correlation between beliefs and values across all policy domains. This shows a strong positive relationship between beliefs and values that is statistically significant at the 1% level. Second, the top right panel shows the results for the correlation between values and political attitudes. In line with the hypothesis, we observe a negative relationship, with left-leaning political attitudes associated with higher agreement with the moral value statements. Third, the three panels in Fig. 2 show that donations are positively correlated with beliefs in the ValueNotSalient treatment, and are also positively correlated with beliefs and values in the ValueSalient treatment. All three are statistically significant at the 1% level.

Collectively, these results are in line with the pre-registered set of hypotheses, providing evidence for the presence of belief-value constellations. This suggests that individuals form beliefs, values and political attitudes in a manner that generates strong associations between the different objects.

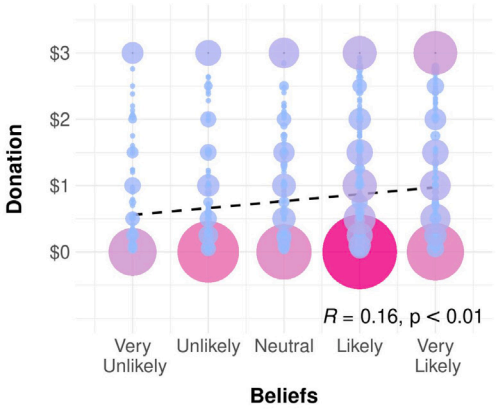
In Online Appendix B.2, we reproduce these results for each of the six topics separately. Unlike the other five topics, the prostitution topic displays a statistically significant *positive* relationship between political attitudes and values, with individuals who identify with the political right stating stronger agreement that prostitution should be illegal than individuals on the political left.[13] Similarly, figures B.6, B.7 and B.8 in the Online Appendix illustrate the relationship between donation decisions and beliefs and values in the ValueNotSalient and ValueSalient treatments for each topic separately. Overall, the relationship between donations and both values and beliefs appears to be weakest for the prostitution-related charity, which received relatively high donation levels across all beliefs and values. Interestingly, for the abortion-related charity, the relationship was very weak in the ValueNotSalient treatment, but very strong when the value debate was made salient in the ValueSalient treatment.

2.3. The formation of belief-value constellations

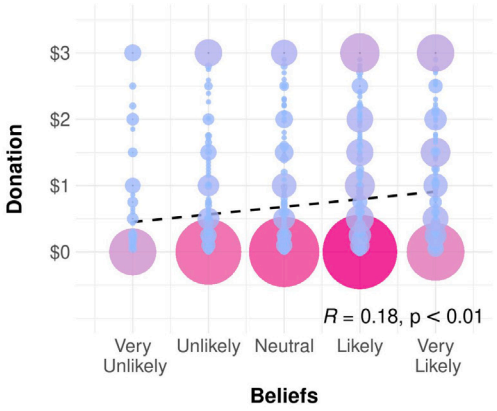
Our second hypothesis asks whether the formation of factual beliefs is influenced by the salience of a particular contentious moral value debate. *Do individuals adjust their factual beliefs when examining them with a related hotly contested moral issue at the forefront of their mind?* If this is the case, it would speak to a neglected mechanism generating tightly clustered beliefs and values.

To test this, we use a between-treatment comparison of the distribution of beliefs observed in the ValueNotSalient and ValueSalient conditions. We can thus assess whether factual beliefs are shifted when we prime individuals to think about these belief statements through the lens of the related value debate.

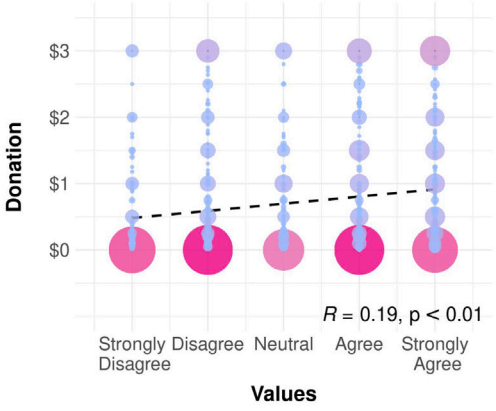
[13] When we designed the experiment, we were aware that the prostitution domain was different to the other domains in the sense that the alignment of aggregate values and political slant was less clear-cut from an ex ante perspective. For this reason, in footnote 35 in Section D.2.1 of our pre-registration document, we noted that our predictions regarding the relationship between values and political slant in the domain of prostitution were more ambiguous.



(a) Donations and Beliefs in VNS



(b) Donations and Beliefs in VS



(c) Donations and Values in VS

Fig. 2. Results for Hypothesis 1c.
Note: Fig. 2(a) shows the correlation between beliefs and donations in treatment ValueNotSalient, Fig. 2(b) shows the correlation between beliefs and donations in treatment ValueSalient, and Fig. 2(c) shows the correlation between values and donations in treatment ValueSalient. The data points are weighted by the number of observations, which is reflected in both the color and size of the markers: the larger and redder the marker, the more observations; the smaller and bluer the marker, the fewer observations (interested readers can also refer to Table B1 in the Online Appendices for additional descriptive statistics). The dotted line represents the result of a linear regression of donations on beliefs in Fig. 2(a) and (b), and a regression of donations of values in Fig. 2(c). The Spearman correlation coefficient, R, and its *p*-value are given at the bottom right of each graph. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This is formalized in Hypothesis 2 below, which posits that: (i) the salience of values affects belief formation, and (ii) this mechanism can result in the polarization of factual beliefs. The rationale for this is that if (i) is true then the heterogeneity in moral values between different political groups would also lead to the formation of polarized factual beliefs. This would provide one potential explanation for recently observed trends of increasingly polarized factual beliefs along ideological lines (see, e.g., Gentzkow, 2016; Enke, 2020a) which has been documented in various domains, such as *climate change* (McCright and Dunlap, 2011) and *COVID-19 beliefs* (Allcott et al., 2020).[14]

Hypothesis 2 (*Construction of Consistent Beliefs*)**.** *Increasing the salience of a contentious moral value debate leads individuals to report factual beliefs that are more strongly aligned with their moral value position. This results in an increase in the polarization of factual beliefs.*

Let F_{b_t} denote the cumulative distribution function (cdf) of factual beliefs b_t in treatment $t \in \{VS, VNS\}$, F_{v_t} the cdf of moral values v_t, and F_{d_t} the cdf of donations d_t. As before, p_t denotes the left–right political stance of individuals.

(a) Raising the salience of a moral value debate influences factual beliefs.

[14] In his theoretical work, Le Yaouanq (2023) links heterogeneity in political attitudes to partisan disagreement about objective facts through people's idiosyncratic preferences regarding the policy implications of scientific findings. Our work seeks to understand the underlying psychological mechanisms in more detail.

The distribution of factual beliefs differs between the ValueNotSalient *and* ValueSalient *treatments:*

$$F_{b_{VNS}} \neq F_{b_{VS}}.$$

(b) *A higher degree of polarization in values in a particular policy domain will result in a stronger effect of increased value salience on the dispersion of factual beliefs.*
Comparing across the six domains indexed by m*, the difference between the variance in beliefs in* ValueSalient *and the variance in beliefs in* ValueNotSalient *is non-decreasing in the variance in values in* ValueSalient*:*

$$\frac{d[Var(b^m_{VS}) - Var(b^m_{VNS})]}{d[Var(v^m_{VS})]} \geq 0.$$

(c) *Raising the salience of a moral value debate results in an increase in the polarization of factual beliefs conditional on political attitudes. Conditional on political attitudes, beliefs in* ValueSalient *are more polarized than beliefs in* ValueNotSalient*:*

$$E(b_{VS}|p_{VS} < E(p_{VS})) - E(b_{VNS}|p_{VNS} < E(p_{VNS}))$$
$$\geq$$
$$E(b_{VS}|p_{VS} > E(p_{VS})) - E(b_{VNS}|p_{VNS} > E(p_{VNS})).$$

Several features of this set of hypotheses are worth highlighting. First, the rationale for part (b) and (c) of the hypothesis is that the raised salience of the relevant value will result in a shift towards more extreme factual beliefs as subjects are drawn towards more coherent belief-value constructions. Second, the inequality in part (c) states that the difference between the factual beliefs of individuals on the left of the political spectrum and those on the right of the political spectrum will increase in ValueSalient versus ValueNotSalient (i.e., the salience of values will increase polarization of factual beliefs, conditional on political attitudes).[15]

Results (Hypothesis 2)

In this section we examine the relationship between values and beliefs more closely by asking whether raising the salience of a particular value leads to a causal shift in an associated factual belief. This comparison represents a fairly conservative test of the existence of a causal relationship between beliefs and values for several reasons. First, our experiment focuses on short-run motivated reasoning and does not consider causal effects of motivated cognition that operate over a longer period of time (e.g., via biased information search or selective memory). Second, our experiment exploits a salience manipulation of values, which represents a fairly weak treatment dosage in relation to an exogenous shift of values. We only encourage participants to *think about* the values they already hold; we do not exogenously shift their values.

Fig. 3 provides a summary of the results associated with Hypothesis 2a and b. The left panel displays the cumulative distribution of reported beliefs in the ValueNotSalient and ValueSalient treatments. We find no statistically significant difference between the two distributions (p-value = 0.09, Kolmogorov-Smirnoff test) and, therefore, do not find support for Hypothesis 2a. Second, the right panel of the figure asks whether there is a heterogeneous effect of increasing the salience of a particular topic. For topics with a high degree of variance in values (i.e. highly polarized issues), we hypothesized that increasing the salience of these values would lead to a larger degree of polarization in the ValueSalient beliefs relative to the beliefs in ValueNotSalient (Hypothesis 2b). Again, we do not find support for our hypothesis, since the slope coefficient is not statistically different from 0 (coefficient: 0.034; p-value: 0.759).

The results for Hypothesis 2a and b suggest that raising the salience of values did not result in a clear shift in the distributions of beliefs across all six issues. Hypothesis 2c posits that even if an increase in value salience does not result in an increase in polarization of the aggregate distribution of beliefs, there may be heterogeneity in the impact of the value salience at the individual level—i.e., the political preferences of an individual could mediate how increasing the salience of their values shifts their beliefs. Essentially, Hypothesis 2c asserts that making a value more salient leads individuals to shift their beliefs even further towards conforming with the average beliefs held by members of their own political party.

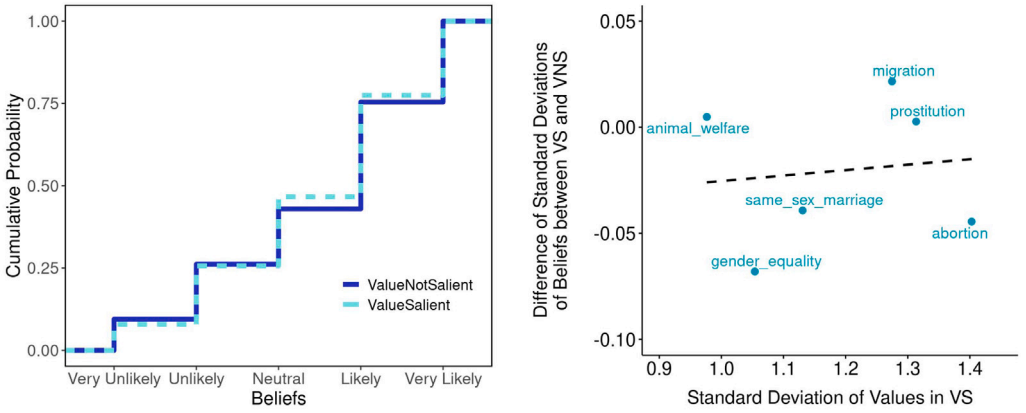
To address this question, we compare the belief movement of individuals on the left of the political attitude spectrum with those on the right of the political attitude spectrum. Using a difference-in-difference style empirical approach, we ask whether the gap between the beliefs of those on the left and the right increases in the ValueSalient treatment relative to in the ValueNotSalient treatment.[16] To do this, we estimate the following regression:

$$b_{i,j} = \alpha_0 + \alpha_1 \cdot \tilde{p}_{i,j} + \alpha_2 \cdot ValSal_{i,j} + \alpha_3 \cdot \tilde{p}_{i,j} \times ValSal_{i,j} + \epsilon_{i,j} \quad (1)$$

where $b_{i,j}$ is the reported belief of individual i for topic j, $ValSal_{i,j}$ is a binary variable that equals 1 if the individual is in the ValueSalient treatment and 0 when the individual is in the ValueNotSalient treatment, and $\tilde{p}_{i,j}$ is an indicator variable that takes a value of 1 if the individual is on the left of the political spectrum, e.g. reports a political attitude that is lower than the mean political attitude reported in our sample. For this purpose, we asked participants the following question after they completed the

[15] This can occur due to individuals on the left of the political spectrum increasing their factual beliefs in ValueSalient versus ValueNotSalient more than those on the right of the political spectrum. Or, it can occur due to individuals on the left increasing their beliefs, while those on the right adjust their beliefs downwards.

[16] In the pre-registration plan, we did not specifically outline this regression, which allows us to test Hypothesis 2c.



(a) CDF of Beliefs in VS and VNS

(b) Change in Variance of Beliefs as a Function of Value Spread

Fig. 3. Results for Hypothesis 2a and b.
Note: Fig. 3(a) shows the results for Hypothesis 2a, i.e. the cumulative density function of beliefs in treatments ValueSalient and ValueNotSalient. Fig. 3(b) shows the result for Hypothesis 2b. The y-axis shows the difference of the standard deviations of beliefs between treatments ValueSalient and ValueNotSalient and the x-axis shows the standard deviation of values in treatment ValueSalient. The dotted line depicts the result from a linear regression of the difference of the standard deviations on the standard deviation of values. The slope coefficient is 0.034 (p-value: 0.759).

experimental part: "In political matters, people talk of "the left" and "the right". How would you place your views on this scale?". Respondents could choose a number between 1 and 10, where 1 indicates the extreme left and 10 indicates the extreme right.[17]

The coefficient of interest is α_3, corresponding to the interaction term. This essentially compares how individuals on the left and right of the political spectrum change their factual beliefs when exposed to an increase in value salience. A positive coefficient denotes a widening of the gap between the factual beliefs of the left and the right. The first column of Table 3 reports the results from estimating Eq. (1), with ValueSalient × Pol. Attitude denoting the interaction term. The estimates show that the coefficient on the interaction term is positive and statistically significant at the one-percent level, providing evidence that we do indeed observe polarization of the beliefs along political attitude division lines when related contentious values are made salient. It is worth noting that this increase in polarization is on top of the pre-existing difference in factual beliefs reported between individuals on the left and right of the political spectrum in the ValueNotSalient treatment. This is shown by the significant coefficient associated with the Pol. Attitude variable. It is also worth noting that the size of the widening of the gap in factual beliefs between the left and right due to the salience is nearly as large as the baseline difference in factual beliefs between individuals on the left and the right in ValueNotSalient (i.e. the magnitude of the coefficient associated with the variable ValueSalient×*Pol. Attitude* is $\frac{3}{4}$ the size of the coefficient associated with the variable *Pol. Attitude*).

In order to test the robustness of this result, we check whether the results are driven by the specific political attitudes variable that we have chosen to use.[18] To do this, we run two further regressions, where we replicate the estimation in the first column of Table 3, but replace the *Left* indicator variable with a variable that indicates that the individual self-reported being a *Democrat* (Column (2)) and a variable that indicates that the individual voted for Hilary Clinton in 2016 (Column (3)). The results from both of these exercises are highly consistent with our main estimation results in Column (1).[19]

These results highlight an important distinction between two forms of polarization, namely (i) polarization of the entire unconditional distribution, which involves movement towards extreme beliefs, and (ii) polarization conditional on a particular characteristic (e.g., political party) that defines groups in the population. The latter form of polarization involves a reshuffling of the belief distribution and may or may not lead to aggregate or unconditional polarization. As seen in Panel (a) in Fig. 3, we do not observe unconditional polarization as a result of our experiment, i.e., the distribution over beliefs does not change depending on whether individuals are primed with their values or not. However, this does not reveal whether individuals from different points on the political spectrum have adjusted their beliefs as a result of the value priming. Our regression results in Table 3 point towards important differences between individuals, conditional on their political preferences.

[17] As a caveat, it should be noted that this question refers to a general political attitude of the respondent and does not elicit their views on social, economic, or other matters separately.

[18] In the pre-registration plan, we only mentioned using the political left–right scale as a variable. To test the robustness of the results, we also examine party affiliation and voting decisions in the last election. Additionally, the pre-registration plan does not mention the use of control variables which were also added to test for robustness here.

[19] Importantly, both of these variables were collected by Prolific completely separately from our experimental data collection. Therefore, these results also serve to alleviate possible concerns regarding our political attitudes variable being influenced by the treatment condition. However, a caveat to this is that the Prolific variables are only available for a subset of the sample. This is the reason for the differing sample sizes across the three regressions.

Table 3
Influence of increased salience of values on belief polarization.

	(1)	(2)	(3)	(4)	(5)	(6)
ValueSalient	−0.124**	−0.199**	−0.236***	−0.133**	−0.222**	−0.243***
	[0.052]	[0.087]	[0.075]	[0.052]	[0.090]	[0.075]
Pol. attitude ($\tilde{p}$)	0.269***	0.358***	0.304***	0.243***	0.328***	0.284***
	[0.047]	[0.071]	[0.058]	[0.048]	[0.076]	[0.061]
ValueSalient	**0.199***	**0.294***	**0.315***	**0.218***	**0.316***	**0.314***
× Pol. attitude	**[0.067]**	**[0.100]**	**[0.089]**	**[0.067]**	**[0.103]**	**[0.089]**
Constant	3.325***	3.202***	3.264***	3.429***	3.270***	3.247***
	[0.038]	[0.063]	[0.049]	[0.070]	[0.101]	[0.090]
Observations	4560	2550	3006	4548	2544	3000
Incl. controls	No	No	No	Yes	Yes	Yes
Pol. attitude ($\tilde{p}$) Variable	Left-right scale Left = 1	Party affiliation Democrat = 1	Last election Clinton = 1	Left-right scale Left = 1	Party affiliation Democrat = 1	Last election Clinton = 1

Notes: (i) Each of the regressions uses the observations from the two treatments, ValueNotSalient (375 observations) and ValueSalient (385 observations), pooled over all six debates ($760 \times 6 = 4560$). (ii) Smaller sample sizes in columns (2) to (6) result from missing information in the political attitude variables and/or in the control variables. (iii) ValueSalient is a dummy variable equal to one if the individual was assigned to treatment ValueSalient and hence equal to zero if the individual was assigned to treatment ValueNotSalient. (iv) We use three measures of the political attitudes variable. This is indicated in the last two rows of the table. In column (1), Political Attitude is a dummy equal to one if the individual is below the mean on a 1 to 10 scale of political attitudes where 1 is the most left, and 10 is the most right attitude. In column (2), Political Attitude is a dummy equal to one if the individual identifies as a Democrat rather than as a Republican, and in column (3), Political Attitude equals one if the individual indicated that they voted for Clinton in the 2016 elections and zero if they voted for Trump. (v) Columns (4) to (6) show the regressions including controls for age, gender, ethnicity, and education. We present the results for the regression run in column (1) using the subsamples from columns (2) and (3) in Table B2 in the Appendix. (vi) Standard errors clustered at the level of the individual are reported in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To provide a more detailed explanation, in Fig. 4, we disaggregate the beliefs of participants according to their position on the political spectrum.[20] The figure plots the mean beliefs in the two treatments ValueNotSalient and ValueSalient for participants grouped by the political attitude they have indicated on a 10-point scale. Purple stars indicate the means for participants in treatment ValueNotSalient, while pink stars indicate the means for individuals in ValueSalient. We observe that individuals at the extremes of the political spectrum—both on the left and the right—display notable shifts in their mean beliefs when comparing the ValueSalient condition to the ValueNotSalient condition. On the left, the pink stars are above the purple stars; on the right, they are below the purple stars. This asymmetry shows that beliefs move in accordance with political orientation (up on the left; down on the right). In other words, individuals with more extreme political preferences tend to exaggerate their beliefs in a manner that aligns with their group membership when their values are primed.

The figure also suggests that the adjustment is larger on the extreme right than on the extreme left, although we need to take into account that we have fewer observations here. In sum, this suggests that while the unconditional distribution of beliefs does not change with the salience of values, the composition of political preferences of people holding different beliefs is affected. Drawing this distinction between *unconditional polarization* and *conditional polarization* is important as it helps us to understand the mechanisms in play. *Unconditional polarization* can be driven by a variety of mechanisms, such as confirmation bias or other individual cognitive heuristics that favor coherent beliefs and values, while *conditional polarization* points towards social conformity with one's in-group as a driving factor.[21]

3. Convincing yourself and convincing others

After having explored how beliefs react to values, we now ask whether money also exerts a distorting influence on beliefs and possibly on values. Our third set of hypotheses below are divided into two parts, with both parts assessing the malleability of beliefs and values to monetary forces that could pull them in different directions. In Part A (Convincing Yourself), we examine the role of self-serving biases in the context of belief-value constellations by asking whether individuals try to justify selfish behavior by adjusting their beliefs and values to be consistent with taking actions that are in their material self-interest—engaging in a form of motivated reasoning or excuse-driven behavior.[22] Part B (Convincing Others) studies whether introducing the opportunity to try to convince another participant to take an altruistic action can lead to a shift in one's own beliefs. Specifically, we ask whether attempts to engage in persuasion lead to a shift in beliefs.

[20] Fig. 4 was not mentioned in the pre-registration plan. It was added to enhance the exposition of our results for Hypothesis 2c.

[21] In Online Appendix Section B.4.1, we also document the donation behavior in the ValueSalient and ValueNotSalient treatment conditions. In summary, we do not observe evidence of a substantial effect on donation decisions, suggesting that the shift in beliefs is not translating into a change in behavior on this dimension. This result contributes to the growing body of work documenting a complex relationship between beliefs and actions.

[22] Previous work has shown that people develop self-serving biases in order to excuse their selfishness in charitable giving (see, e.g., Exley, 2015 on the role of risk or Exley, 2020 on using charity performance metrics as an excuse).

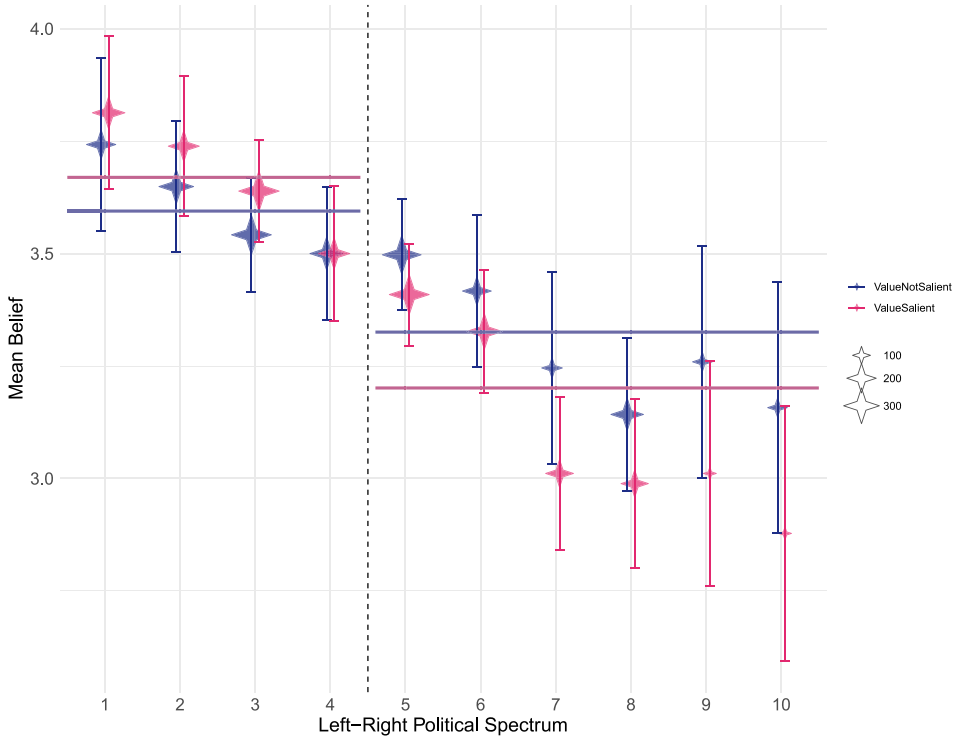


Fig. 4. Mean belief by political orientation and treatment.
Note: The stars indicate the mean of beliefs in treatment ValueNotSalient (purple) and ValueSalient (pink) conditional on the political attitude of respondents. The associated error bars represent 95% confidence intervals around the means. Responses from the Likert scale regarding beliefs were coded as follows: 1 - "Very Unlikely", 2 - "Unlikely", 3 - "Neutral", 4 - "Likely", 5 - "Very Likely". For political attitudes we asked participants the following question after they completed the experimental part: "In political matters, people talk of "the left" and "the right". How would you place your views on this scale?". Respondents could choose a number between 1 and 10 where 1 indicates utmost left and 10 indicates utmost right. The vertical dashed line is the mean of political values in the sample comprising observations from Treatment ValueNotSalient and ValueSalient. The horizontal bars represent the predicted values from our main regression (see Table 3) for the four groups ValueNotSalient - below the mean (purple line on the left), ValueSalient - below the mean (pink line on the left), ValueNotSalient - above the mean (purple line on the right) and ValueSalient - above the mean (pink line on the right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Incentives to convince others and belief polarization.

	(1)	(2)
ConvinceOther	0.089	0.102*
	[0.058]	[0.056]
Pol. attitude ($\tilde{p}$)	0.468***	0.462***
	[0.048]	[0.046]
ConvinceOther × Pol. attitude	**0.005**	**−0.025**
	[0.074]	**[0.072]**
Constant	3.201***	3.379***
	[0.036]	[0.072]
Observations	4488	4476
Incl. controls	No	Yes

Notes: (i) Each regression uses the observations from treatments ValueSalient (385 observations) and ConvinceOther (363 observations), pooled over all six debates. (ii) Smaller sample sizes in column (2) result from missing information in the control variables. (iii) ConvinceOther is a dummy variable equal to one if the individual was assigned to treatment ConvinceOther and hence equal to zero if the individual was assigned to treatment ValueSalient. (iv) Political Attitude is a dummy equal to one if the individual is below the mean on a 1 to 10 scale of political attitudes where 1 is the most left and 10 is the most right attitude. (v) Column (2) show the regressions including controls for age, gender, ethnicity and education. (vi) Standard errors clustered by individual in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.1. The ConvinceSelf and ConvinceOther treatments

We conduct two further treatments. First, the ConvinceSelf treatment speaks to the conjecture that individuals adjust their beliefs and values in a self-serving way. This treatment is very similar to ValueSalient, with only one key difference: in ConvinceSelf, subjects are aware that they will need to make a charitable donation decision when they form and report their moral value and factual belief assessments. The way that this is implemented in the experiment is that they see the donation decision on the same screen as the one where they report their values and beliefs. This is in contrast to ValueSalient, where the charitable donation screen arrives as a surprise after the moral value and factual belief reports have been completed.[23] This difference is important, since subjects' anticipation of the costly charitable donation decision could influence their introspection in forming a personal assessment of the value and factual belief statements. Hypothesis 3a conjectures that individuals bias their (stated) beliefs and values when they take into account the costs of an expected donation decision, with the bias shifting beliefs and values away from those that would justify a higher donation. However, it is also possible that the anticipated donation could have the opposite effect, leading the individual to inflate the importance of her values and beliefs. This self-convincing process may justify a high donation as the correct decision, thereby enhancing her self-image utility from donating.[24]

Second, the ConvinceOther treatment investigates how trying to persuade others to take an action that is aligned with one's own values could lead an individual to further align their factual beliefs with their political agenda or goals, potentially leading to an exaggeration of these stated beliefs. The underlying assumption here is that subjects get utility from higher donations of others if the donation aligns with their own beliefs and values. To do this, treatment ConvinceOther mirrors treatment ValueSalient with just a single exception: before stating their values and beliefs, subjects are informed that *another* participant will have the option to donate to a related charity after being informed about the moral values and factual beliefs that they (the subject in ConvinceOther) reported. So, participants might consider the possibility that their own values and beliefs could influence the donation decision of another subject. In order to avoid deception, we implemented these decisions by others in an auxiliary treatment, BeingConvinced. The first part of the BeingConvinced treatment is identical to ValueSalient, with subjects reporting their values and beliefs on the six relevant topics. The difference arrives prior to subjects making their donation decisions. At this point, subjects in BeingConvinced are informed about the beliefs and values stated by a randomly chosen participant from ConvinceOther.

3.2. Do individuals self-servingly shift their beliefs and values?

To examine this question, we compare behavior in ConvinceSelf, where subjects anticipate their future donation decisions, with behavior in ValueSalient, where subjects report their values and beliefs before they are aware of the future donation decisions. This allows us to study the robustness of elicited beliefs and values to the presence of monetary incentives that could distort them. More specifically, we ask whether the presence of the donation decision on the same screen induces subjects to distance themselves from the charity-aligned value position and to adjust their beliefs away from supporting the charity's goals. This is summarized in the following set of hypotheses.

Hypothesis 3A (*Convincing Yourself*)**.** *As before, let F_{b_t} denote the cumulative distribution function (cdf) of factual beliefs b in Treatment t, and F_{v_t} the cdf of moral values v. Donations in Treatment $t \in \{VS, VNS, CS, CO\}$ are denoted by d_t and p_t denotes the left–right political stance of individuals.*

(a) Individuals shift their beliefs and values to justify taking self-serving actions: In ConvinceSelf individuals shift their beliefs and values downwards in comparison to in ValueSalient in order to justify low future donation decisions. Specifically:

(i) b_{VS} first-order stochastically dominates b_{CS}, i.e. $F_{b_{VS}} \leq F_{b_{CS}}$.
(ii) v_{VS} first-order stochastically dominates v_{CS}, i.e. $F_{v_{VS}} \leq F_{v_{CS}}$.

(b) Donations in ConvinceSelf are lower than in ValueSalient:

$$E(d_{VS}) \geq E(d_{CS}).$$

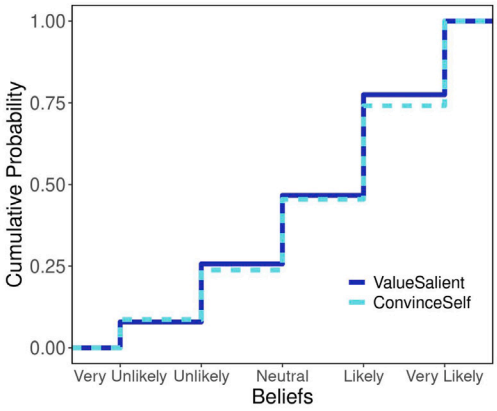
Results (Hypothesis 3A)

Essentially, we find no evidence in support of Hypothesis 3A. Fig. 5 displays the distribution of beliefs (top left panel), values (top right panel) and donations (bottom panel) in the ValueSalient and ConvinceSelf treatments. We observe no significant differences in average behavior between these two treatments, indicating that individuals do not shift their beliefs and values when faced with an imminent donation decision.[25] This immutability of behavior to the anticipated donation decision is in stark contrast to the

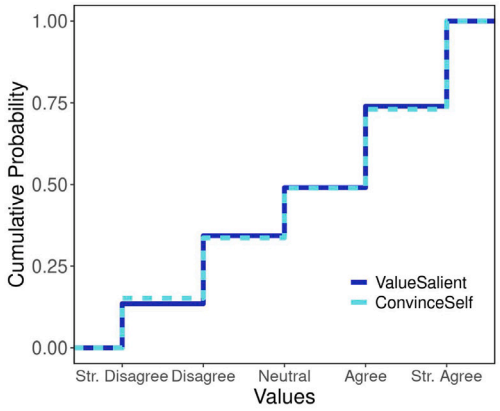
[23] Figures A.1 and A.3 in the Online Appendix show screenshots of the instructions as they were presented to participants in both treatments.

[24] In our pre-registration document we noted this possibility but stated that our prior was that the self-serving bias would dominate.

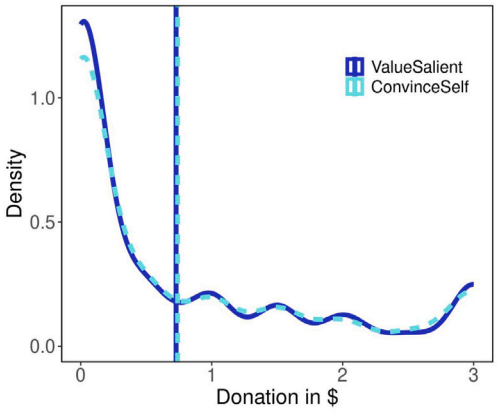
[25] We can also separately examine whether donation decisions are affected on either the extensive or intensive margin. Our analysis reveals no significant average differences on either margin. One possible explanation for the failure to observe an average treatment effect is that some individuals shift their beliefs due to monetary self-interest, while others shift their beliefs to enhance the self-image generated from making a donation, with the two effects offsetting each other. Our data does not allow us to convincingly evaluate this possible explanation empirically, but we do observe a higher correlation between donations and values and beliefs in the ConvinceSelf treatment in comparison to the ValueSalient treatment. This may be interpreted as evidence in favor of this explanation.



(a) CDF of beliefs in VS and CS (Hypothesis 3A.a.i)



(b) CDF of values in VS and CS (Hypothesis 3A.a.ii)



(c) PDF of donations in VS and CS (Hypothesis 3A.b)

Fig. 5. Results for Hypothesis 3A.
Note: The three figures show the results on Hypothesis 3. Fig. 5(a) shows the cumulative density function of beliefs for treatment ValueSalient (dark line) and ConvinceSelf (light dotted line), Fig. 5(b) shows the cumulative density function of values for treatment ValueSalient (dark line) and ConvinceSelf (light dotted line), and Fig. 5(c) probability density function of donations for treatment ValueSalient (dark line) and ConvinceSelf (light dotted line). The vertical lines in Fig. 5(c) depict the mean of donations in the two treatments.

effect of increased value salience documented above. While subjects are engaging in politically motivated reasoning, they do not engage in economically motivated reasoning. Perhaps one reason for this is that individuals place a higher value on their personal identity, which incorporates their beliefs and values, than they do on a small monetary gain that they would obtain by reducing their donation. A second factor worth noting is that a large fraction of subjects donated less than 1 dollar. Thus, the cognitive dissonance costs of donating a low amount may not be sufficiently high to warrant a shift in beliefs or values to justify it.

3.3. Do individuals shift their beliefs and values to convince others?

The second part of Hypothesis 3 asks whether individuals report more polarized factual beliefs when they have the opportunity to try to persuade someone else about the importance of certain value positions. It therefore contributes to the body of existing work that examines the idea that we adjust our own beliefs and attitudes (i.e., convince ourselves) in order to convince others (Babcock et al., 1995; Schwardmann and van der Weele, 2019; Solda et al., 2020; Schwardmann et al., 2022). While this previous work predominantly studies scenarios in which an individual is explicitly mandated to convince others about a particular policy position or that they themselves are of high ability, a key difference in our study is that we focus on examining whether individuals try to persuade others to take an action that is aligned with their own values by stating more extreme beliefs. For example, we ask whether an individual might increase their agreement with the statement that "Animals feel less pain than humans." in order to encourage another person to donate to an animal protection charity.

Hypothesis 3B (*Convincing Others*)**.** *Anticipating the opportunity to persuade another individual about a contentious moral issue shifts one's own factual beliefs towards the in-group party aligned extreme—i.e., factual beliefs in* ConvinceOther *are more polarized than factual*

beliefs in ValueSalient*:*

$$E(b_{CO}|p_{CO} < E(p_{CO})) - E(b_{VS}|p_{VS} < E(p_{VS}))$$
$$\geq$$
$$E(b_{CO}|p_{CO} > E(p_{CO})) - E(b_{VS}|p_{VS} > E(p_{VS})).$$

Similar to Hypothesis 2c above, the inequality in Hypothesis 3B states that the gap in factual beliefs between individuals on the left and the right of the political spectrum widens when there is an anticipated persuasion opportunity.

Results (Hypothesis 3B)

To examine Hypothesis 3b, Table 4 uses the same empirical specification as above and tests for a divergence of beliefs according to political attitudes between the ValueSalient and the ConvinceOther treatment.[26] Essentially, this asks whether individuals shift their beliefs even further towards conforming with their political in-group when they know that their reports will be viewed by others. The results do not support this hypothesis, with the estimated coefficient on the interaction term close to zero. There are several plausible explanations for this finding, including: (i) that individuals do not wish to persuade others, (ii) that individuals are not prepared to adjust their own beliefs to persuade others, and (iii) that they do not believe that others will be easily persuaded in the context of these contentious debates and the limited communication space available (i.e., not being able to send explanations, narratives or justifications for their persuasive claim). Our data does not allow us to convincingly differentiate between these explanations empirically. Therefore, it is important to keep in mind that multiple possible explanations could account for this result.

4. Conclusion

This paper studies whether thinking about moral values can influence beliefs about facts. This is done using a preregistered online experiment that surveyed a nationally representative sample of 1863 individuals from the US population. We ask two broad sets of questions.

First, we examine whether systematic correlations exist between moral values (“ought” statements) and factual beliefs (“is” statements). We find evidence supporting this relationship, consistent with previous research that highlights a societal shift towards increasingly partisan worldviews (see, e.g., Alesina et al., 2020; Bonomi et al., 2021). As we discuss above, there are many mechanisms that might create such a correlation. For example, beliefs might shape values. Our study examines whether there is a (reverse) causal relationship between values and beliefs, whereby thinking about values exerts an influence on beliefs. We explore this by introducing a treatment that makes a moral value more salient prior to eliciting beliefs. Strikingly, while there appears to be no effect on the aggregate distribution, a closer inspection shows substantial causal effects of thinking about values on beliefs—effects that are mediated by prior political leanings. In other words, we find that individuals in our representative sample are engaged in politically motivated reasoning.[27]

Politically motivated reasoning takes place on both sides of the political spectrum: subjects on both the political right and the political left, shift their beliefs to align them with the average party beliefs when values are made salient. This finding contrasts with the popular belief that the flirtation with “alternative facts” is a phenomenon exclusive to populist right-wing movements.

The influence of *thinking about values* may be generated by different potential psychological mechanisms underlying motivated reasoning. One possibility is that thinking about a particular value position strengthens the individual’s desire for this position to be justified by facts; consequently they shift their beliefs (“pure motivated reasoning”). Another possibility is that thinking about a particular value position cues recall of information in the individual’s memory database that is supportive of the value position (“motivated memory”). This relationship between memory and biased belief has been explored in recent work on associative memory (Enke et al., 2024) and motivated memory (Zimmermann, 2020; Amelio and Zimmermann, 2023). Our experiment is not able to cleanly distinguish between these channels through which the effect of thinking about values may operate. We leave this interesting question for future work.

Additionally, we examine whether there is evidence for economically motivated reasoning whereby individuals bias their beliefs and/or values due to the presence of monetary incentives to do so. This is not the case. We believe that this result enhances the credibility of our main findings. Since beliefs and values do not react to (small) monetary incentives, it appears that individuals care about them to the extent that they do not distort them through economically motivated reasoning.

When interpreting our results, several considerations should be kept in mind. In footnote 9, we explained our rationale for not incentivizing the belief elicitation of factual beliefs. Although evidence from Danz et al. (2022), Haaland et al. (2023), and Stantcheva (2023) suggests that this decision is unlikely to cause systematic distortions in reported beliefs, it is still possible that individuals report beliefs that differ from those they hold in mind. Additionally, it is possible that experimenter demand effects might influence the beliefs reported by participants in our experiment. While we consider it unlikely that demand effects systematically account for

[26] In the pre-registration plan, we did not specifically outline this regression, which allows us to test Hypothesis 3B. Additionally, the pre-registration plan does not mention the use of control variables which were added to test for robustness here.

[27] The behavior observed in our study is consistent with the findings of Bordalo et al. (2021), who study the effect of issue salience on beliefs about others’ political attitudes. The authors show that when the salience of a particular policy conflict is raised, this increases the perception of the partisan gap in attitudes. Combined with an identity-induced desire to conform to the stereotypical beliefs of one’s identity group (as in Bonomi et al., 2021), this perceived increase in the partisan gap could contribute to the shift in beliefs that we observe. One caveat is worth keeping in mind when interpreting our results: We focus on a particular kind of belief, namely those without a scientific consensus. For facts where such a scientific consensus does exist, the findings of Drobner (2022) suggest that we might expect less motivated reasoning since individuals will anticipate uncertainty resolution.

the pattern of beliefs reported across treatments (e.g., we observe no effect of the ConvinceOther treatment manipulation where demand effects might be expected to be strongest), we cannot completely rule out this possibility. Therefore, these potential influences should be kept in mind when interpreting the results.

Taken together, our results point towards a deep (cognitive) link between values and beliefs. This tight relationship between values and beliefs is consistent with the conceptual idea of a "polarized reality", where individuals perceive reality through the lens of their economic or social identity (Alesina et al., 2020) and then adjust their beliefs to conform to the stereotypical belief of the salient identity group (Bonomi et al., 2021). More broadly, this recent line of research showing how identity shapes beliefs through the desire for group conformity builds on a longer history of research examining how identity can generate a desire for conformity in *actions* (Akerlof and Kranton, 2000, 2005; Shayo, 2020). With the polarization of social discourse (particularly online) seemingly increasing in society, this body of work points towards identity-induced belief conformity as an important avenue for further research.

Finally, our results also suggest that individuals may engage more readily in politically motivated reasoning than economically motivated reasoning. One caveat to this assertion is that the economic incentives in our experiment are limited. Nevertheless, this points towards a potentially important division of the space of possible motivated reasoning domains. This, too, is an interesting avenue for further research.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.euroecorev.2024.104929.

References

Akerlof, George A., Kranton, Rachel E., 2000. Economics and identity. Q. J. Econ. 115 (3), 715–753.
Akerlof, George A., Kranton, Rachel E., 2005. Identity and the economics of organizations. J. Econ. Perspect. 19 (1), 9–32.
Alesina, Alberto, Miano, Armando, Stantcheva, Stefanie, 2020. The polarization of reality. AEA Papers Proc. 110, 324–328.
Alesina, Alberto, Miano, Armando, Stantcheva, Stefanie, 2022. Immigration and redistribution. Rev. Econ. Stud. 90 (1), 1–39.
Alesina, Alberto, Stantcheva, Stefanie, Teso, Edoardo, 2018. Intergenerational mobility and preferences for redistribution. Amer. Econ. Rev. 108 (2), 521–554.
Allcott, Hunt, Boxell, Levi, Conway, Jacob, Gentzkow, Matthew, Thaler, Michael, Yang, David, 2020. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. J. Public Econ. 191, 104254.
Amasino, Dianna, Pace, Davide, van der Weele, Joël, 2021. Fair Shares and Selective Attention. Working Paper.
Amelio, Andrea, Zimmermann, Florian, 2023. Motivated memory in economics–A review. Games 14 (1), 15.
Babcock, Linda, Loewenstein, George, Issacharoff, Samuel, Camerer, Colin, 1995. Biased judgments of fairness in bargaining. Amer. Econ. Rev. 85 (5), 1337–1343.
Barron, Kai, 2021. Belief updating: Does the 'good-news, bad-news' asymmetry extend to purely financial domains? Exp. Econ. 24 (1), 31–58.
Barron, Kai, Fries, Tilman, 2023. Narrative Persuasion. CESifo Working Paper No. 10206.
Barron, Kai, Stüber, Robert, van Veldhuizen, Roel, 2024. Navigating Moral Trade-Offs. WZB Discussion Paper.
van Bavel, Jay J., Pereira, Andrea, 2018. The partisan brain: An identity-based model of political belief. Trends in Cognitive Sciences 22 (3), 213–224.
Belot, Michele, Briscese, Guglielmo, 2022. Bridging America's Divide on Abortion, Guns and Immigration: An Experimental Study. CEPR Discussion Paper 17444.
Bernheim, Douglas, 1994. A theory of conformity. J. Polit. Econ. 102 (5), 841–877.
Bicchieri, Cristina, Dimant, Eugen, Gächter, Simon, Nosenzo, Daniele, 2022. Social proximity and the erosion of norm compliance. Games Econom. Behav. 132, 59–72.
Bicchieri, Cristina, Dimant, Eugen, Sonderegger, Silvia, 2023. It's not a Lie if you believe the norm does not apply: Conditional norm-following and belief distortion. Games Econom. Behav. 138, 321–354.
Bolsen, Toby, Druckman, James N., Cook, Fay Lomax, 2014. The influence of partisan motivated reasoning on public opinion. Polit. Behav. 36 (2), 235–262.
Bonomi, Giampaolo, Gennaioli, Nicola, Tabellini, Guido, 2021. Identity, beliefs, and political conflict. Q. J. Econ. 136 (4), 2371–2411.
Bordalo, Pedro, Tabellini, Marco, Yang, David, 2021. Issue Salience and Political Stereotypes. NBER Working Paper.
Clinton, Joshua, Cohen, Jon, Lapinski, John, Trussler, Marc, 2021. Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19. Sci. Adv. 7 (2), eabd7204.
Coutts, Alexander, 2019. Good news and bad news are still news: Experimental evidence on belief updating. Exp. Econ. 22 (2), 369–395.
Danz, David, Vesterlund, Lise, Wilson, Alistair J., 2022. Belief elicitation and behavioral incentive compatibility. Amer. Econ. Rev. 112 (9), 2851–2883.
Di Tella, Rafael, Perez-Truglia, Ricardo, Babino, Andres, Sigman, Mariano, 2015. Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. Amer. Econ. Rev. 105 (11), 3416–3442.
Drobner, Christoph, 2022. Motivated beliefs and anticipation of uncertainty resolution. Am. Econ. Rev. 4 (1), 89–105.
Druckman, James N., McGrath, Mary C., 2019. The evidence for motivated reasoning in climate change preference formation. Nature Clim. Change 9 (2), 111–119.
Druckman, James N., Peterson, Erik, Slothuus, Rune, 2013. How elite partisan polarization affects public opinion formation. Am. Polit. Sci. Rev. 107 (1), 57–79.
Eil, David, Rao, Justin M, 2011. The good news-bad news effect: asymmetric processing of objective information about yourself. Am. Econ. J. Microecon. 3 (2), 114–138.
Enke, Benjamin, 2020a. Moral values and voting. J. Polit. Econ. 128 (10), 3679–3729.
Enke, Benjamin, 2020b. What you see is all there is. Q. J. Econ. 135 (3), 1363–1398.
Enke, Benjamin, Schwerter, Frederik, Zimmermann, Florian, 2024. Associative memory, beliefs and market interactions. J. Financ. Econ. 157, 103853.
Exley, Christine L., 2015. Excusing selfishness in charitable giving: The role of risk. Rev. Econ. Stud. 83 (2), 587–628.
Exley, Christine L., 2020. Using charity performance metrics as an excuse not to give. Manage. Sci. 66 (2), 553–563.
Flaxman, Seth, Goel, Sharad, Rao, Justin M., 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. Public Opin. Q. 80 (S1), 298–320.
Gaines, Brian J., Kuklinski, James H., Quirk, Paul J., Peyton, Buddy, Verkuilen, Jay, 2007. Same facts, different interpretations: Partisan motivation and opinion on Iraq. J. Polit. 69 (4), 957–974.
Gentzkow, Matthew, 2016. Polarization in 2016. Technical report, Toulouse Network for Information Technology Working Paper.
Ging-Jehli, Nadja R., Schneider, Florian H., Weber, Roberto A., 2020. On self-serving strategic beliefs. Games Econom. Behav. 122, 341–353.
Goette, Lorenz, Huffman, David, Meier, Stephan, 2006. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. Amer. Econ. Rev. 96 (2), 212–216.
Gotthard-Real, Alexander, 2017. Desirability and information processing: An experimental study. Econom. Lett. 152, 96–99.

Graeber, Thomas, Roth, Christopher, Schesch, Constantin, 2024. Explanations. ECONtribute Discussion Paper.
Graham, Jesse, Haidt, Jonathan, Nosek, Brian A., 2009. Liberals and conservatives rely on different sets of moral foundations. J. Pers. Soc. Psychol. 96 (5), 1029–1046.
Haaland, Ingar, Roth, Christopher, 2023. Beliefs about racial discrimination and support for pro-black policies. Rev. Econ. Stat. 105 (1), 40–53.
Haaland, Ingar, Roth, Christopher, Wohlfart, Johannes, 2023. Designing information provision experiments. J. Econ. Lit. 61 (1), 3–40.
Huffman, David, Raymond, Collin, Shvets, Julia, 2022. Persistent overconfidence and biased memory: Evidence from managers. Amer. Econ. Rev. 112 (10), 3141–3175.
Kahan, Dan M., 2013. Ideology, motivated reasoning, and cognitive reflection: An experimental study. Judgm. Decis. Mak. 8, 407–424.
Kahan, Dan M., 2016. The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In: Emerging Trends in the Social and Behavioral Sciences. John Wiley & Sons, Ltd, pp. 1–16.
Konow, James, 2000. Fair shares: Accountability and cognitive dissonance in allocation decisions. Amer. Econ. Rev. 90 (4), 1072–1091.
Le Yaouanq, Yves, 2023. A model of voting with motivated beliefs. J. Econ. Behav. Organ. 213, 394–408.
Leeper, Thomas J., Slothuus, Rune, 2014. Political parties, motivated reasoning, and public opinion formation. Polit. Psychol. 35, 129–156.
Lord, Charles G., Ross, Lee, Lepper, Mark R., 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.. J. Pers. Soc. Psychol. 37 (11), 2098–2109.
McCright, Aaron M., Dunlap, Riley E., 2011. The politicization of climate change and polarization in the American public's views of global warming, 2001-2010. Sociol. Q. 52 (2), 155–194.
Meeuwis, Maarten, Parker, Jonathan A., Schoar, Antoinette, Simester, Duncan, 2022. Belief disagreement and portfolio choice. J. Finance 77 (6), 3191–3247.
Messick, David M., Sentis, Keith P., 1979. Fairness and preference. J. Exp. Soc. Psychol. 15 (4), 418–434.
Möbius, Markus M., Niederle, Muriel, Niehaus, Paul, Rosenblat, Tanya S., 2022. Managing self-confidence: Theory and experimental evidence. Manage. Sci. 68 (11), 7793–7817.
Nickerson, Raymond S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Rev. Gen. Psychol. 2 (2), 175–220.
Ortoleva, Pietro, Snowberg, Erik, 2015. Overconfidence in political behavior. Amer. Econ. Rev. 105 (2), 504–535.
Rabin, Matthew, Schrag, Joel L., 1999. First impressions matter: A model of confirmatory bias. Q. J. Econ. 114 (1), 37–82.
Schwardmann, Peter, Tripodi, Egon, van der Weele, Joël J., 2022. Self-persuasion: Evidence from field experiments at international debating competitions. Amer. Econ. Rev. 112 (4), 1118–1146.
Schwardmann, Peter, van der Weele, Joël, 2019. Deception and self-deception. Nat. Hum. Behav. 3 (10), 1055–1061.
Shayo, Moses, 2020. Social identity and economic policy. Annu. Rev. Econ. 12, 355–389.
Sherman, David K., Cohen, Geoffrey L., 2006. The psychology of self: defense: Self affirmation theory. Adv. Exp. Soc. Psychol. 183–242.
Solda, Alice, Ke, Changxia, Page, Lionel, von Hippel, William, 2020. Strategically delusional. Exp. Econ. 23 (3), 604–631.
Stantcheva, Stefanie, 2023. How to run surveys: A guide to creating your own identifying variation and revealing the invisible. Annu. Rev. Econ. 15, 205–234.
Taber, Charles S., Lodge, Milton, 2006. Motivated skepticism in the evaluation of political beliefs. Am. J. Polit. Sci. 50 (3), 755–769.
Thaler, Michael, 2020. Do people engage in motivated reasoning to think the world is a good place for others? arXiv preprint arXiv:2012.01548.
Thaler, Michael, 2024. The fake news effect: Experimentally identifying motivated reasoning using trust in news. Am. Econ. J. Microecon. 16 (2), 1–38.
Zimmermann, Florian, 2020. The dynamics of motivated beliefs. Amer. Econ. Rev. 110 (2), 337–363.