Motivated Political Reasoning: The Emergence of Belief-Value Constellations*

Kai Barron†, Anna Becker‡, Steffen Huck§

January 8, 2022

Abstract

We study the causal relationship between moral values ("ought" statements) and factual beliefs ("is" statements) and show that, contrary to predictions of orthodox Bayesian models, values exert an influence on beliefs. This effect is mediated by prior political leanings and, thus, contributes to increasing polarization in beliefs about facts. We study this process of motivated political reasoning in a pre-registered online experiment with a nationally representative sample of 1,500 individuals in the US. Additionally, we show that subjects do not distort their beliefs in response to financial incentives to do so, suggesting that deep values exert a stronger motivational force.

JEL Codes: C90, D72, D74, D83, P16
Keywords: Motivated Beliefs, Values, Polarization, Experiment.

*This research was approved by the UCL Research Ethics Committee (REC) (17181/001). We pre-registered the analysis on OSF (https://osf.io/8jydh/). Barron gratefully acknowledges financial support from the German Science Foundation via CRC TRR 190 (project number 280092119).

†WZB Berlin: kai.barron@wzb.eu
‡University College London: anna.becker.14@ucl.ac.uk
§University College London and WZB Berlin: steffen.huck@wzb.eu

1 Introduction

Why would Republicans hold different beliefs about the dangers of Covid-19 than Democrats and, as a consequence, take fewer precautions and increase their risk of becoming infected? This startling pattern observed in beliefs and behavior during the pandemic is documented in both Allcott et al. (2020) and Clinton et al. (2021). However, the wider phenomenon of partisan "bubbles" comprising *disagreement about facts* along the political spectrum can be traced back in the United States to the early 2000s (see, for example, Gaines et al. 2007, on polarized beliefs about the Iraq war).

One explanation for this phenomenon is provided in a seminal article by Taber and Lodge (2006) that introduces the notion of "motivated skepticism" to explain the maintenance of partisan beliefs as a consequence of biased information processing: citizens exhibit a tendency to interrogate arguments and information that is in conflict with their prior partisan attitudes more vigorously, while uncritically accepting attitudinally congruent arguments. Furthermore, the motive to conform with one's political party (or signal one's conformity with it to others) has been shown to be an important mediator for motivated reasoning on policy issues. For example, in Druckman, Peterson, and Slothuus (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. More recently, in a conceptual piece that also reviews earlier empirical evidence, Alesina, Miano, and Stantcheva (2020) note that Democrats and Republicans appear to view the same factual reality through different lenses, which the authors refer to as a partisan "polarization of reality". Importantly, the authors note that there are several potential explanations for this phenomenon, and that the direction of causality is often unclear (for example, individuals may select into political parties on the basis of their beliefs or other personal characteristics).

We provide a novel contribution to this line of inquiry by exploring how beliefs depend on underlying *values* – examining the correlations and causal relationships, as well as the consequences for actions. We conceptualize values as desires about the social world, that is, statements about how the world ought to be. In contrast to beliefs, which pertain to states of the physical world, values can neither be objectively true nor false – they can only be endorsed or opposed with possibly different strength.

For a rational decision maker, values might causally depend on beliefs but not vice versa. For example, if I believe that animals suffer from pain in a similar fashion to humans, I might

desire a world in which animals are no longer killed and eaten.[1] In contrast, values should (according to the rational model) have no impact on beliefs. Whether I desire a world in which we are all vegetarians should not influence my assessment of how likely it is that animals suffer from pain when they are injured.

Yet, if values are central to an individual's identity, such that maintaining them is important, then motivated reasoning might be employed in protecting these core values. Such motivated reasoning could reverse the causal relationship, allowing values to causally shape the beliefs individuals hold.

We study this 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 doing so, we reveal the role played by party preferences in motivated reasoning and also analyze the consequences for actions, specifically, donations to charities operating in these six domains.

While values have been shown to be strongly associated with party preferences and voting decisions (see, for example, Enke 2020a) the precise causal relationship between values and factual beliefs remains largely unexplored. In our study we can draw on previously measured party preferences and show how they mediate the relationship between values and beliefs.[2]

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 or attractiveness (e.g., Eil and Rao 2011; Mobius et al. 2011; Coutts 2019; Drobner 2021), 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, Stüber, and Veldhuizen 2019; Amasino, Pace, and Van der Weele 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, Tripodi, and Van der Weele 2021), and engaging in confirmatory reasoning that reinforces one's prior beliefs (e.g., Nickerson 1998;

1. Notice that this maneuver does not contradict Hume's (1978) assessment that ought statements cannot be derived purely from a set of facts. Support for the value statement that animals should not be killed requires implicitly drawing on an ought axiom about the avoidance of suffering in general.

2. To the extent that values may drive party preferences, one could capture the full causal system that we have in mind as a directed acyclic graph (DAG) in the spirit of Pearl (2009) with beliefs being (potentially) driven by values and party preferences, and party preferences also influenced by values (i.e., $v \rightarrow p \rightarrow b \leftarrow v$). In the language of DAGs party preferences would then be a mediator between values and beliefs, and values occupy the parent node, influencing beliefs either directly or via party preferences. Note, also, that any rational model would require that causality run in the opposite direction, with beliefs serving as a parent node to values and party preferences.

Rabin and Schrag 1999).

In contrast to much of this literature, which typically considers motivated reasoning in relation to a belief that is closely tied to an individual's self-interest, their personal characteristics or their pre-existing beliefs about a particular topic, 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 and given also the substantial heterogeneity in values between and within societies, this strikes us as an important question. Individuals who differ in their values might still be able to achieve compromises as long as they agree on the facts. But when the facts are in dispute and beliefs about the very nature of the world diverge, bipartisan action will be severely impeded.

In order to explore the role played by moral values in influencing factual beliefs, as well as the joint influence of beliefs and values on decision making, we designed and conducted a pre-registered online experiment in January 2020 that surveyed a nationally representative sample of 1,500 individuals from the US population.[3] Our experiment comprises four main treatments that allow us to test hypotheses organized around the following questions: (i) whether there exist systematic correlations between values and factual beliefs; (ii) whether individuals adapt their factual beliefs when there is an increase in the salience of a moral value in the same domain; (iii) whether individuals shift their stated values and factual beliefs in order to align them with their own material self interest; and (iv) whether individuals adjust their stated values and beliefs in an attempt to persuade others.

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, the second from the animal welfare domain. The statement "All countries benefit from the free movement of labor" pertains to a fact that may either be true 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. This is similarly the case for the two statements "Animals feel less pain than humans" (belief) and "It is wrong to eat animals" (value).

3. 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).

Our study makes use of these two pairs of statements as well as four others created in a similar way. That is, in each case the factual statement and the moral statement come from the same domain of life such that we can employ the notion of belief-value constellations that can be spatially represented as "bubbles" or constellations in two-dimensional belief-value space.[4]

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. This elicitation of values serves to raise the salience of these values when subjects report their associated beliefs thereafter. 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 them to the forefront of the individual's mind.

The underlying idea for treatments ValueSalient and ValueNotSalient is that when faced with a heightened salience of the value question, individuals may shift their factual beliefs in a motivated way in order to align them with their values. We test this by comparing the distribution of beliefs reported in these two treatments. We examine this comparison both unconditional and conditional on party preferences.

In all our treatments there is one final stage after the belief and value elicitations exploring choices that relate to the six domains. 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, in treatment ConvinceSelf, we put the donation decision on the same screen on which we elicit beliefs and values (in other treatments, it comes as a surprise). We can, hence, test whether subjects adjust their reasoning in a self-interested way when holding particular belief-value constellations would point towards taking a costly action (i.e., making a donation to a charity whose work is aligned with that particular constellation). In such a setting, individuals might be less inclined to hold certain beliefs and/or values as it becomes more costly to do so.

Finally, we test whether subjects adjust their stated beliefs and/or values when they have

4. Essentially, the notion of belief-value constellations reflects the idea that the beliefs and values individuals hold might manifest in clusters of associated beliefs and values, with individuals who hold a certain value more likely to hold a certain belief, and vice versa.

an incentive to persuade others. In the fourth treatment, ConvinceOther, we again ask subjects to state their beliefs and values but 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.

We find strong support for the existence of aligned belief-value constellations in all the policy domains considered, thereby answering our first research question in the affirmative and providing crisp evidence for the existence of "bubbles" with within-bubble homogeneity of beliefs and values and between-bubble heterogeneity in both dimensions. Notice, however, that while such correlations are evidence for some kind of partisanship it is not possible to understand the mechanism behind this without further evidence. Such bubbles may arise when beliefs shape values (in a fully rational manner), when there are filter bubbles or echo chambers (Flaxman, Goel, and Rao 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 to explore the latter mechanism.

In the aggregate, the distributions of reported subjective beliefs are almost identical across the two treatments which, on the surface, appears to suggest that we do not observe a shift in subjects' beliefs when making values more salient. However, the picture changes dramatically when we control for individuals' political preferences (as planned in our preregistration). 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 party when made aware of the associated value debate. We, therefore, do find support for the idea that values shape beliefs through motivated reasoning. As a consequence, heightened salience of contentious policy issues in public debate emerges as an explanatory force for the increasing polarization in factual beliefs along political attitude division lines. This is a new result that is subtly different from motivated skepticism or other forms of biased updating previously documented in the political science literature and adds a new dimension to the motivated reasoning literature in economics—an individual's deep values can motivate their reasoning about factual beliefs as easily as monetary incentives or self-image does in other contexts.

We also find the effect of values to be quite large. The magnitude of the shift in beliefs due to value salience is nearly as large as the baseline difference in beliefs between subjects on the left and the right in the control treatment. We conduct several robustness exercises in support of these results, including replacing our main measure of political attitude with various alternative proxy measures that were elicited by us, as well as measures elicited independently by the recruitment platform prior to our experiment.

Our work therefore contributes empirical evidence to an active recent theoretical discussion about how and why partisan individuals increasingly seem to live with polarized mental models of reality (Leeper and Slothuus 2014; Van Bavel and Pereira 2018; Alesina, Miano, and Stantcheva 2020). In particular, we relate closely to Bonomi, Gennaioli, and Tabellini (2021), who 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. Therefore, the main results of our paper can be viewed as providing support for some of the central ideas expressed in their theory (although our experiment was not designed as a test of their theory).[5]

With respect to our third and fourth research questions, we find beliefs and values 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 which 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. 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. Specifically, the author shows that people don't engage in motivated reasoning in forming beliefs about whether the world is a good place for others to live (i.e., about cancer survival rates, infant mortality and others' happiness). 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 not resulting in motivated reasoning.

5. While we focus predominantly on assessing the causal effect of values 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, Druckman, and Cook (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, Stantcheva, and Teso (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.

2 Existence and Formation of Belief-Value Constellations

Our experimental design consists of four pre-registered[6] treatments that were conducted online using the platform Prolific with a nationally representative sample of 1,863 individuals from the US population.[7] 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 objective of our ValueSalient treatment is to examine whether we observe a systematic correlation between values and associated factual beliefs. The experimental design of the ValueSalient 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 the statement. 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. Therefore, the moral value agreement questions serve to raise the salience of these debates for the participants, who might then view later questions in the experiment through the lens of those debates. Second, participants are asked to state their belief that each of the six factual statements is true. Importantly, each of the six factual statements is related to one of the same six public policy debates as the moral value statements.[8] Together, the moral value assessments and factual belief reports allow us to examine whether there is a correlation between individuals' beliefs and values. Third, to examine how these values and beliefs translate into actions, 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 of one of the six relevant public policy discussions.[9] 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, additionally, we were

6. The full pre-registration document can be found at https://osf.io/8jydh/ and is also reproduced in Appendix C.

7. Table B1 in the appendix shows that our sample is strongly balanced between all the treatments which are described in the following.

8. The moral value statements are evaluated on a 5-point Likert scale from "Strongly Agree" to "Strongly Disagree", while the factual beliefs are also assessed on a 5-point Likert scale from "Very Unlikely" to "Very Likely". The public policy debates that we consider are migration, animal welfare, gender equality, abortion, prostitution and gay rights. A complete overview of the moral value and factual statements can be found in Table 1 in Appendix C.

9. 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 2 in Appendix C.

able to match our data to previously elicited political attitude variables collected by Prolific independently from our experiment.

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 this shift in salience affects the factual beliefs.

2.2 The Existence of Belief-Value Constellations

The first question we seek to answer with these two treatments is whether there is an alignment between the moral values, factual beliefs and political attitudes that individuals hold. This would indicate the presence of "belief-value constellations". It is important to note that such belief-value constellations are not a natural implication of standard economic theory, where individuals process information and update their beliefs about factual statements in a dispassionate way according to Bayes' rule. However, there are several potential reasons why individuals might form aligned beliefs and values, including: i) the *avoidance of cognitive dissonance* from holding incoherent values and beliefs, and ii) the use of value and belief statements to justify self-interested actions (i.e., *motivated reasoning*).

The presence of such belief-value constellations would imply that it is important to take an individual's moral values into consideration when trying to understand belief formation regarding factual statements (which is not typically done in the literature). Our first set of hypotheses address this question of whether belief-value constellations are observed systematically in the population.[10]

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.

10. 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 Appendix C for further details.

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:

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

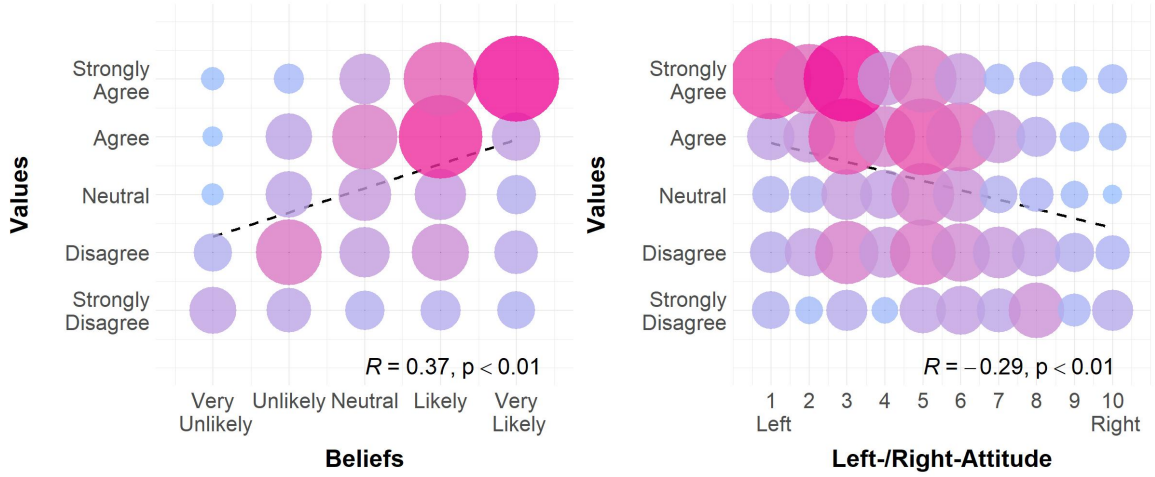
$$Corr(d_{VS}, b_{VS}) \geq 0, Corr(d_{VS}, v_{VS}) \geq 0.$$

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, 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)

In this section, we provide evidence relating to the key hypothesis of whether individuals form beliefs and values in a manner that generates belief-value constellations. Figure 1 summarizes the results pertaining to this hypothesis. 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 Figure 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.

Figure 1: Results for Hypothesis 1a and 1b



(a) Beliefs and Values (Hyp. 1a)

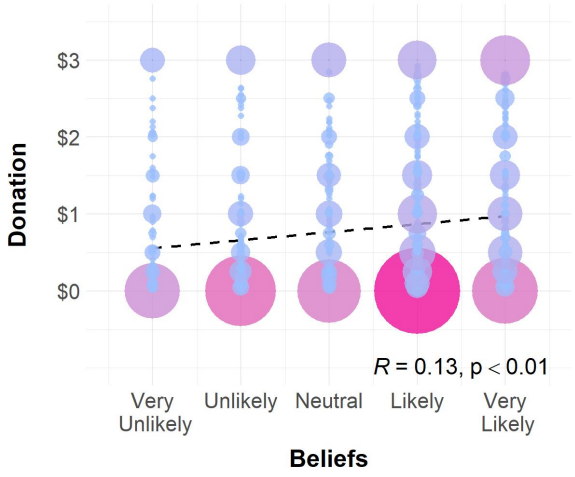
(b) Political Attitudes and Values (Hyp. 1b)

Note: Figure 1(a) shows the results on Hypothesis 1a, i.e. the correlation between values and beliefs in the ValueSalient treatment. Figure 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 shows in color and size of the points. The dotted line represents the result of a linear regression of values on beliefs respectively political attitudes. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

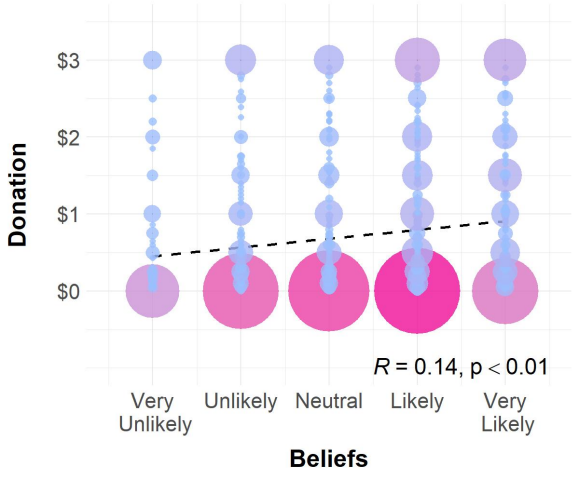
Collectively, these results are completely in line with the pre-registered set of hypotheses, providing strong evidence for the presence of belief-value constellations. This suggests that individuals form beliefs, values and political attitudes in a manner that generates strong associations between the different objects. This is important and striking because economists often conceptualize *values* as being preferences held by individuals, while *beliefs* reflect probability assignments over states of the world and pertain to the (objective) processing of information. These objects are typically treated as being orthogonal by economists. Our results suggest a more nuanced interdependent relationship between these two objects.

In Appendix A.1, we reproduce these results for each of the six topics separately. While most of the topic-specific results are completely in line with the aggregate analysis above, the analysis reveals several interesting findings regarding the heterogeneity in the strength of the associations between the variables across topics. While Figure A.1 shows that all six topics display a positive relationship between beliefs and values at the 1% level, Figure A.2 shows that there is a negative relationship between values and political attitudes for five of the six topics. The prostitution topic in fact displays a statistically significant *positive* relationship between political attitudes and values, with individuals who identify with the political right stating stronger agreement that that prostitution should be illegal than individuals on the political left.

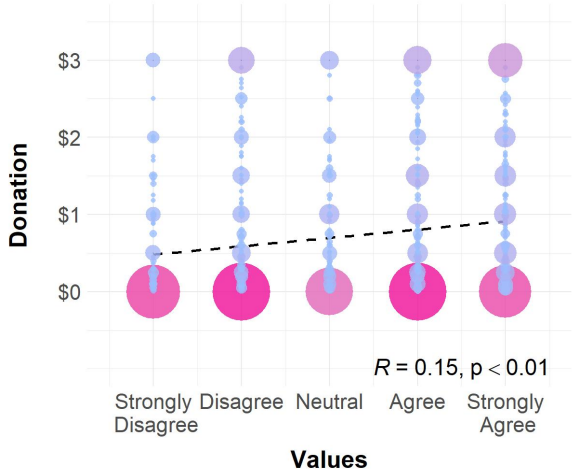
Figure 2: Results for Hypothesis 1c



(a) Donations and Beliefs in VNS



(b) Donations and Beliefs in VS



(c) Donations and Values in VS

Note: The three figures show the results on Hypothesis 1c. Figure 2(a) shows the correlation between beliefs and donations in treatment VALUENOTSALIENT, Figure 2(b) shows the correlation between beliefs and donations in treatment VALUENOTSALIENT, and Figure 2(c) shows the correlation between values and donations in treatment VALUESALIENT. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

Figures A.3, A.4 and A.5 illustrate the relationship between donation decisions and beliefs and values in the VALUENOTSALIENT and VALUESALIENT treatments. While the point estimates of these relationships are positive in all eighteen comparisons, we observe substantial heterogeneity in the strength of the relationship. 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. This is indicative of a shift in contribution choices when seen through the lens of the contentious value debate.

Overall, the results show strong support for Hypothesis 1.

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 the question of why individuals end up holding 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.

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 these polarized factual beliefs. This would provide one potential explanation for observed recent trend 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).[11]

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. This results in an increase in 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.

11. In his theoretical work, Le Yaouanq (2021) 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.

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

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 increasing 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 results in an increase in polarization of factual beliefs.*

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 average factual belief in ValueSalient versus ValueNotSalient is greater for subjects on the left of the political spectrum in comparison to those on the right. To put this another way, the hypothesis states that individuals on the left will increase their beliefs between ValueNotSalient and ValueSalient more than individuals on the right of the political spectrum, on average.[12]

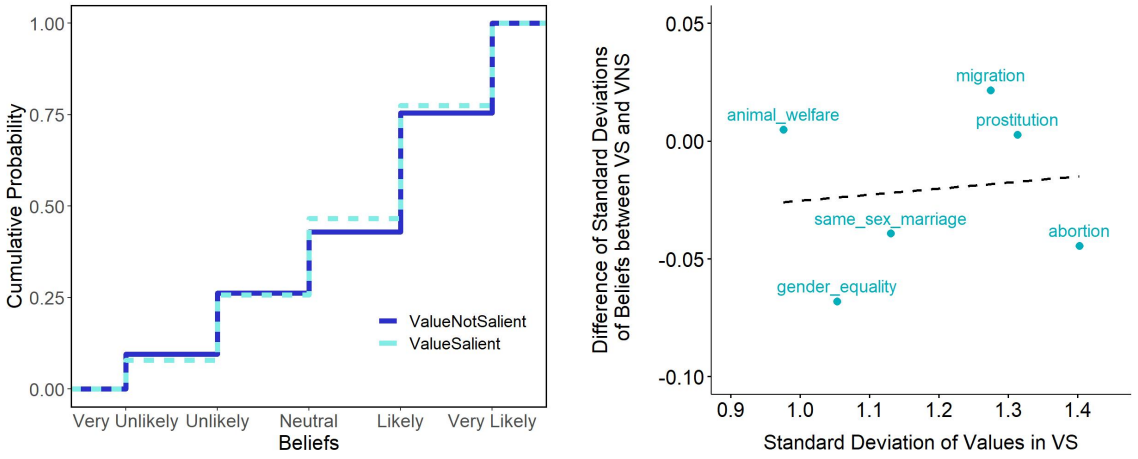
Results (Hypothesis 2)

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

12. Note that this also includes the case where subjects on the right adjust their beliefs downwards which would make the right-hand side negative.

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 dosage of the treatment (i.e., an exogenous shift of values). We therefore view our treatment manipulation as placing a lower bound on the causal relationship between values and beliefs.

Figure 3: Results for Hypotheses 2a and 2b



(a) CDF of Beliefs and Values in VS and VNS

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

Note: Figure 3(a) shows the results for Hypothesis 2a, i.e. the cumulative density function of beliefs in treatments ValueSalient and ValueNotSalient. Figure 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.

Figure 3 provides a summary of the results associated with Hypothesis 2a and 2b. The left panel displays the cumulative distribution of reported beliefs in the ValueNotSalient and ValueSalient treatments. This panel provides suggestive evidence in favor of the polarization of beliefs when pooling across all topics, with the dashed line weakly below the solid line in the left part of the figure, while the dashed line is weakly above the solid line in the right part of the figure. However, this difference is not statistically significant at the 5% level (p-value = 0.09, Kolmogorov-Smirnoff test). 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 of the ValueSalient beliefs relative to the beliefs in ValueNotSalient (Hypothesis 2b). While

we do observe an upward sloping linear relationship, the slope coefficient is not statistically different from 0.

The results for Hypotheses 2a and 2b 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 therefore 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. To do this, we estimate the following regression:

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

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 (i.e. reports a political attitude that is lower than the mean political attitude reported in our sample).

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. Table 1 reports the results from estimating equation 1 in the first column, 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*).

Table 1: Influence of increased salience of values on belief polarization

	(1)	(2)	(3)
ValueSalient	-0.124**	-0.199**	-0.236***
	[0.051]	[0.092]	[0.077]
Pol. Attitude ($\tilde{p}$)	0.269***	0.358***	0.304***
	[0.052]	[0.076]	[0.068]
ValueSalient	**0.199***	**0.294***	**0.315***
× Pol. Attitude	**[0.074]**	**[0.109]**	**[0.096]**
Constant	3.325***	3.202***	3.264***
	[0.037]	[0.063]	[0.055]
Observations	4560	2550	3006
Pol. Attitude ($\tilde{p}$)	Left-Right Scale	Party Affiliation	Last Election
Variable	(Left = 1)	(Democrat = 1)	(Clinton = 1)

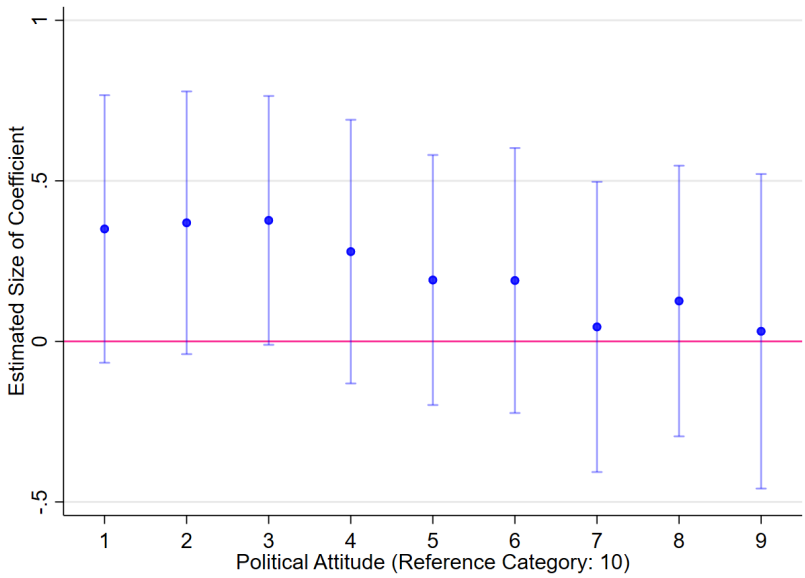
Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: 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. 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 median 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.

In order to test the robustness of this result, we conduct several additional exercises. First, we check whether the results are driven by the specific political attitudes variable that we have chosen to use. To do this, we run two further regression, where we replicate the estimation in the first column of Table 1, 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).[13]

Figure 4: Influence of increased salience on beliefs across the political spectrum (Hyp. 2c)



Note: Figure 4 shows the coefficients α_3 estimated using a regression as the one described in equation (1) above with the difference that $\tilde{p}_{i,j}$ is using the full scale of political attitudes, i.e. we include a dummy for each value on the political left-right scale, using the category '10' (the most right category) as a reference. The vertical lines represent the 95% confidence interval for the estimated coefficents.

Second, instead of collapsing the 10-point political attitudes variable into a binary indicator (as in Table 1), we estimate this regression using the full 10-point scale and report all ten interaction coefficients in Figure 4. This shows that individuals on the left of the political attitude spectrum are shifting their beliefs upwards relative to individuals who reported being completely on the right (i.e., those who reported a 10 and constitute the omitted category). Overall these robustness exercises are highly supportive of the finding that when an issue is viewed as more politically charged (e.g., when a contentious related value discussion becomes more salient), individuals tend to shift their beliefs to conform more with their political in-group. Interestingly, this is not associated with a polarization of factual beliefs at the aggregate level.[14] These results highlight an important distinction between two forms of polarization, namely (i) polarization of the entire unconditional distribution, which in-

13. 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.

14. This can be the case if the individuals from the two parties are moving their beliefs in opposite directions and essentially replacing one another out in the aggregate distribution. This occurs when the movers do not start off close to their party-aligned factual belief extreme, and therefore have to "jump over" one another to conform with their political in-group's beliefs.

volves 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. 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.

Overall, the collective evidence provided by the exercises and robustness checks we implement is supportive of Hypothesis 2c (i.e., *conditional polarization*), but we do not observe strong evidence in favor of Hypotheses 2a and 2b (i.e., *unconditional polarization*).[15]

3 Convincing Yourself and Convincing Others

After having explored how beliefs react to values we now ask whether money exerts a similar influence on beliefs and possibly on values. The third set of our hypotheses below will be 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.[16] In addition, Part B (Convincing Others) studies whether introducing the opportunity to try to convince another participant to take a specific action can lead to a shift in one's own beliefs. Specifically, we ask whether attempts to engage in persuasion lead to a shifts in beliefs.

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

15. In Appendix Section A.2.1, 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.

16. 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).

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. This is in contrast to ValueSalient, where the charitable donation screen arrive as a surprise after the moral value and factual belief reports have been completed. 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.[17]

Second, the ConvinceOther treatment examines how trying to convince others to take an action that is in line with one's own values could lead an individual to further align their factual beliefs with their political agenda or goals, and perhaps to exaggerate these stated beliefs. To do this, treatment ConvinceOther mirrors again 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 reflect on the possibility that their own values and beliefs could exert and influence on 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 in the presence of monetary incentives that could distort them. We, specifically, ask whether the presence of the donation decision on the same screen induces subjects to distance themselves from the charity-aligned value

17. It is also possible that the anticipated donation might operate in the opposite direction inflating the the importance of values and beliefs to convince such that the agent can convince herself that making a high donation is the correct decision. In our pre-registration document we noted this possibility but stated that our prior was that the self-serving bias would dominate.

position, and similarly 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. $F_{v_{VS}} \leq F_{v_{CS}}$.

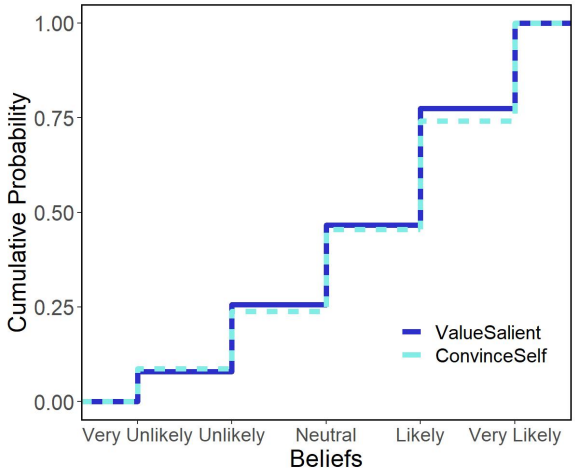
b) Donations in ConvinceSelf are lower than in ValueSalient:

$$E(d_2) \geq E(d_3).$$

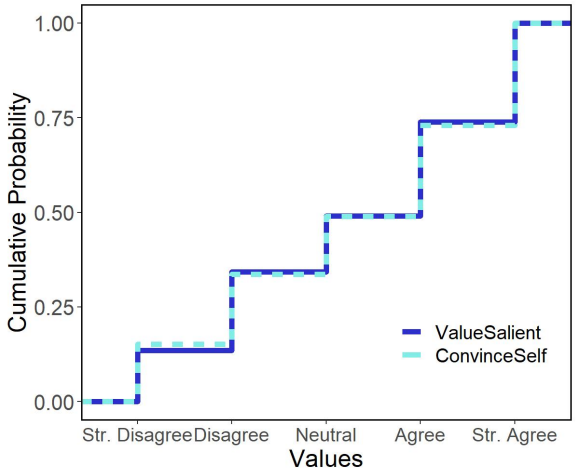
Results (Hypothesis 3A)

Essentially, we find no evidence in support of Hypothesis 3A. Figure 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 behavior between these two treatments, indicating that individuals do not shift their beliefs and values when faced with an imminent donation decision. This immutability of behavior to the anticipated donation decision is in stark contrast to the effects salient values 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 are perhaps not sufficiently high to warrant a shift in beliefs or values to justify it.

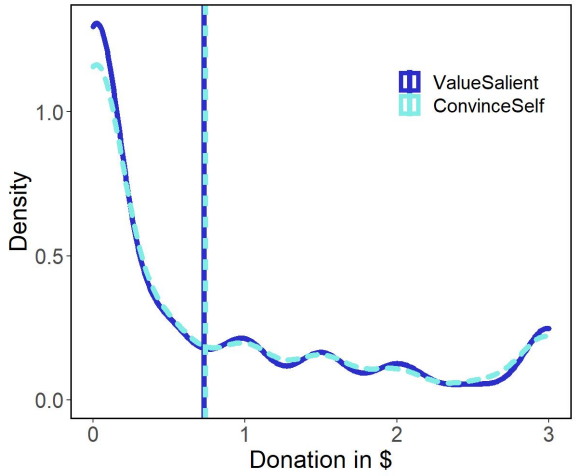
Figure 5: Results for Hypothesis 3A



(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)

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

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), and Schwardmann, Tripodi, and Van der Weele (2021). 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 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:

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

Similarly to Hypothesis 2c above, the inequality here in Hypothesis 3c 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 3c, Table 2 uses the same empirical specification as above and tests for a divergence of beliefs according to political attitudes between the ValueSalient and ConvinceOther treatment. 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. Plausible explanations include: (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.[18]

18. Another potential reason for this is that we are comparing the beliefs in the ConvinceOther with the ValueSalient, where beliefs have already been shifted towards political conformity relative to ValueNotSalient due to the salience of the value debates. This salience shift may crowd out any further shift when individuals wish to convince others.

Table 2: Outcome: Beliefs

	(1)
ConvinceOther	0.089*
	[0.052]
Pol. Attitude ($\tilde{p}$)	0.468***
	[0.052]
ConvinceOther × Pol. Attitude	0.005
	[0.074]
Constant	3.201***
	[0.035]
Observations	4488

Standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: 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.Political Attitude is a dummy equal to one if the individual is below the median on a 1 to 10 scale of political attitudes where 1 is the most left and 10 is the most right attitude.

4 Conclusion

This paper studies the relationship between moral values and factual beliefs based on the results of a pre-registered online experiment that surveyed a nationally representative sample of 1,500 individuals from the US population. First, we ask whether there exist systematic correlations between moral values ("ought" statements) and factual beliefs ("is" statements). This is answered in the affirmative and is consistent with previous research that discusses the societal shift towards an increasingly partisan view of the world (see, e.g., Alesina, Miano, and Stantcheva 2020; Bonomi, Gennaioli, and Tabellini 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 exists a (reverse) causal relationship between values and beliefs where values exert an influence on beliefs (which should not happen in a perfectly rational Bayesian world). We explore this by introducing a treatment that makes a moral value (pertaining to the same domain as the belief question) more salient prior to eliciting beliefs. Strikingly, while there appears to be no effect in the aggregate, a closer inspection shows substantial causal effects of 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.[19]

In our analysis, we proceed exactly as described in our pre-registration document. In addition to examining this reverse causal channel, this includes examining whether there is also 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 shade them through economically motivated reasoning.

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.

Taken together, our results point towards a deep (cognitive) link between values and beliefs. This is in sharp contrast to the notion that they should be treated as disjoint separate objects as described by the standard model. 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, Miano, and Stantcheva 2020) and then adjust their beliefs to conform to the stereotypical belief of the salient identity group (Bonomi, Gennaioli, and Tabellini 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.

19. The behavior observed in our study is consistent with the findings of Bordalo, Tabellini, and Yang (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, Gennaioli, and Tabellini 2021), this perceived increase in the partisan gap could contribute to the shift in beliefs that we observe.

References

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115 (3): 715–753.

———. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19 (1): 9–32.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2020. "The polarization of reality." *AEA Papers and Proceedings* 110:324–28.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso. 2018. "Intergenerational mobility and preferences for redistribution." *American Economic Review* 108 (2): 521–54.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. 2020. "Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic." *Journal of Public Economics* 191:104254.

Amasino, Dianna, Davide Pace, and Joël Van der Weele. 2021. "Fair Shares and Selective Attention." *Working Paper*.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." *American Economic Review* 85 (5): 1337–1343.

Barron, Kai. 2021. "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" *Experimental Economics* 24 (1): 31–58.

Barron, Kai, and Christina Gravert. 2021. "Confidence and career choices: An experiment." *Scandinavian Journal of Economics*.

Barron, Kai, Robert Stüber, and Roel van Veldhuizen. 2019. *Motivated motive selection in the lying-dictator game.* Technical report. WZB Discussion Paper.

Bolsen, Toby, James N Druckman, and Fay Lomax Cook. 2014. "The influence of partisan motivated reasoning on public opinion." *Political Behavior* 36 (2): 235–262.

Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini. 2021. "Identity, Beliefs, and Political Conflict." *Quarterly Journal of Economics* 136 (4): 2371–2411.

Bordalo, Pedro, Marco Tabellini, and David Yang. 2021. "Issue Salience and Political Stereotypes." *NBER Working Paper*.

Clinton, Joshua, Jon Cohen, John Lapinski, and Marc Trussler. 2021. "Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19." *Science Advances* 7 (2): eabd7204.

Costa-Gomes, Miguel A, Steffen Huck, and Georg Weizsäcker. 2014. "Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect." *Games and Economic Behavior* 88:298–309.

Costa-Gomes, Miguel A, and Georg Weizsäcker. 2008. "Stated beliefs and play in normal-form games." *Review of Economic Studies* 75 (3): 729–762.

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics* 22 (2): 369–395.

Drobner, Christoph. 2021. "Motivated Beliefs and Anticipation of Uncertainty Resolution." *American Economic Review: Insights (forthcoming)*.

Druckman, James N, and Mary C McGrath. 2019. "The evidence for motivated reasoning in climate change preference formation." *Nature Climate Change* 9 (2): 111–119.

Druckman, James N, Erik Peterson, and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American Political Science Review* 107 (1): 57–79.

Eil, David, and Justin M Rao. 2011. "The good news-bad news effect: asymmetric processing of objective information about yourself." *American Economic Journal: Microeconomics* 3 (2): 114–38.

Enke, Benjamin. 2020a. "Moral values and voting." *Journal of Political Economy* 128 (10): 3679–3729.

———. 2020b. "What you see is all there is." *Quarterly Journal of Economics* 135 (3): 1363–1398.

Exley, Christine L. 2015. "Excusing Selfishness in Charitable Giving: The Role of Risk." *Review of Economic Studies* 83, no. 2 (October): 587–628.

———. 2020. "Using Charity Performance Metrics as an Excuse Not to Give." *Management Science* 66 (2): 553–563.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80, no. S1 (March): 298–320.

Gaines, Brian J, James H Kuklinski, Paul J Quirk, Buddy Peyton, and Jay Verkuilen. 2007. "Same facts, different interpretations: Partisan motivation and opinion on Iraq." *Journal of Politics* 69 (4): 957–974.

Gentzkow, Matthew. 2016. *Polarization in 2016.* Technical report. Toulouse Network for Information Technology Working Paper.

Gotthard-Real, Alexander. 2017. "Desirability and information processing: An experimental study." *Economics Letters* 152:96–99.

Haaland, Ingar, and Christopher Roth. 2021. "Beliefs about racial discrimination and support for pro-black policies." *Review of Economics and Statistics,* 1–38.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2020. "Designing information provision experiments." *CEBI Working Paper Series, Working Paper 20/20.*

Hume, David. 1978. *Treatise of Human Nature.* 2nd edition. Edited by L.A. Selby-Bigge. London: Oxford University Press.

Ivanov, Asen, Dan Levin, and Muriel Niederle. 2010. "Can relaxation of beliefs rationalize the winner's curse?: An experimental study." *Econometrica* 78 (4): 1435–1452.

Kahan, Dan M. 2013. "Ideology, motivated reasoning, and cognitive reflection: An experimental study." *Judgment and Decision making* 8:407–24.

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American Economic Review* 90 (4): 1072–1091.

Le Yaouanq, Yves. 2021. *Motivated cognition in a model of voting.* Technical report. Working Paper.

Leeper, Thomas J, and Rune Slothuus. 2014. "Political parties, motivated reasoning, and public opinion formation." *Political Psychology* 35:129–156.

McCright, Aaron M., and Riley E. Dunlap. 2011. "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010." *Sociological Quarterly* 52 (2): 155–194.

Messick, David M, and Keith P Sentis. 1979. "Fairness and preference." *Journal of Experimental Social Psychology* 15 (4): 418–434.

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. *Managing self-confidence: Theory and experimental evidence.* Technical report. National Bureau of Economic Research.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology* 2 (2): 175–220.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference.* Cambridge University Press.

Rabin, Matthew, and Joel L Schrag. 1999. "First impressions matter: A model of confirmatory bias." *Quarterly Journal of Economics* 114 (1): 37–82.

Schwardmann, Peter, Egon Tripodi, and Joël J Van der Weele. 2021. "Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions." *American Economic Review (forthcoming)*.

Schwardmann, Peter, and Joël J Van der Weele. 2019. "Deception and self-deception." *Nature Human Behaviour* 3 (10): 1055–1061.

Shayo, Moses. 2020. "Social identity and economic policy." *Annual Review of Economics* 12:355–389.

Solda, Alice, Changxia Ke, Lionel Page, and William Von Hippel. 2020. "Strategically delusional." *Experimental Economics* 23 (3): 604–631.

Taber, Charles S, and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50 (3): 755–769.

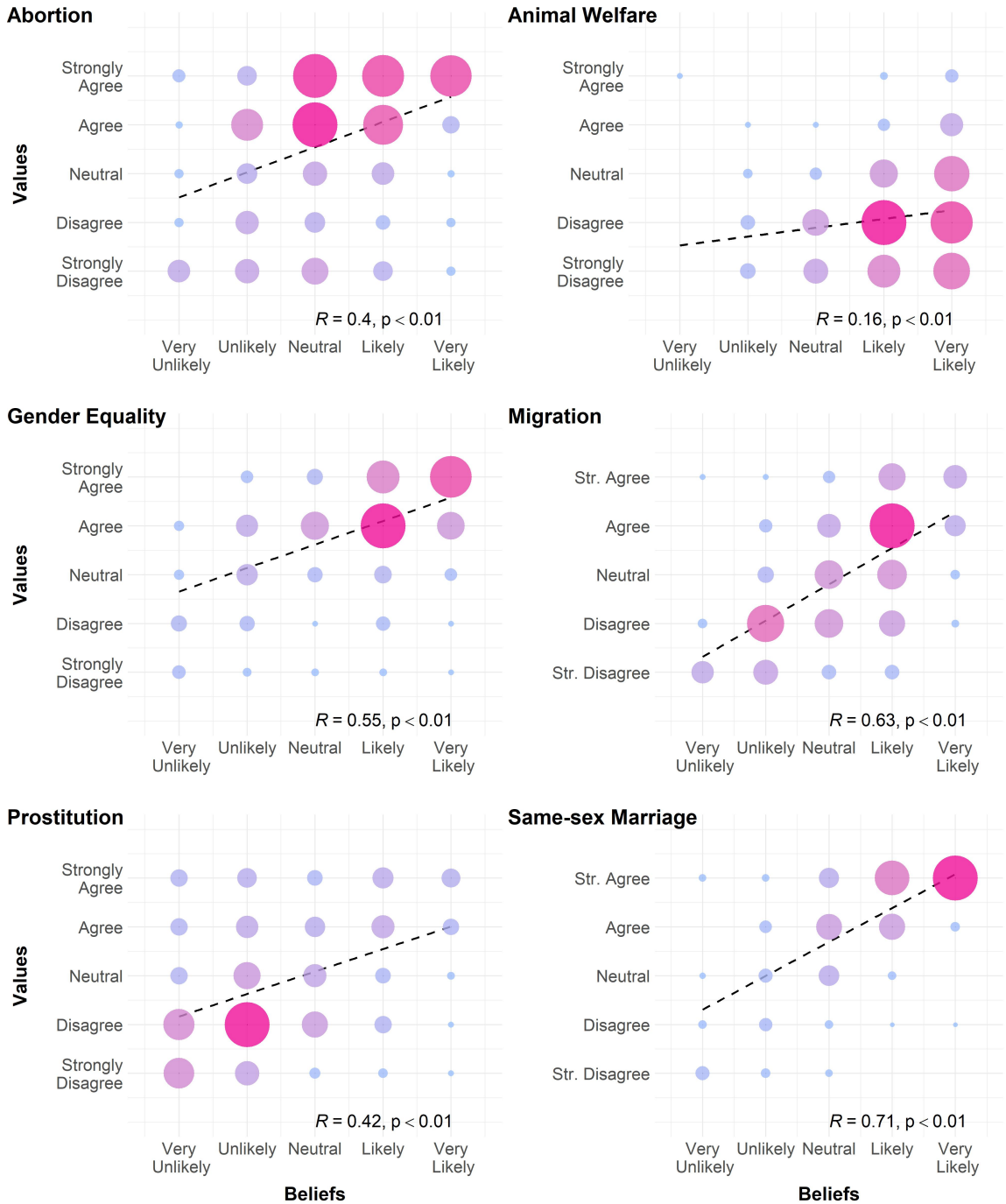
Thaler, Michael. 2020. "Do People Engage in Motivated Reasoning to Think the World Is a Good Place for Others?" *arXiv preprint arXiv:2012.01548*.

Van Bavel, Jay J, and Andrea Pereira. 2018. "The partisan brain: An identity-based model of political belief." *Trends in Cognitive Sciences* 22 (3): 213–224.

A Supplementary Results

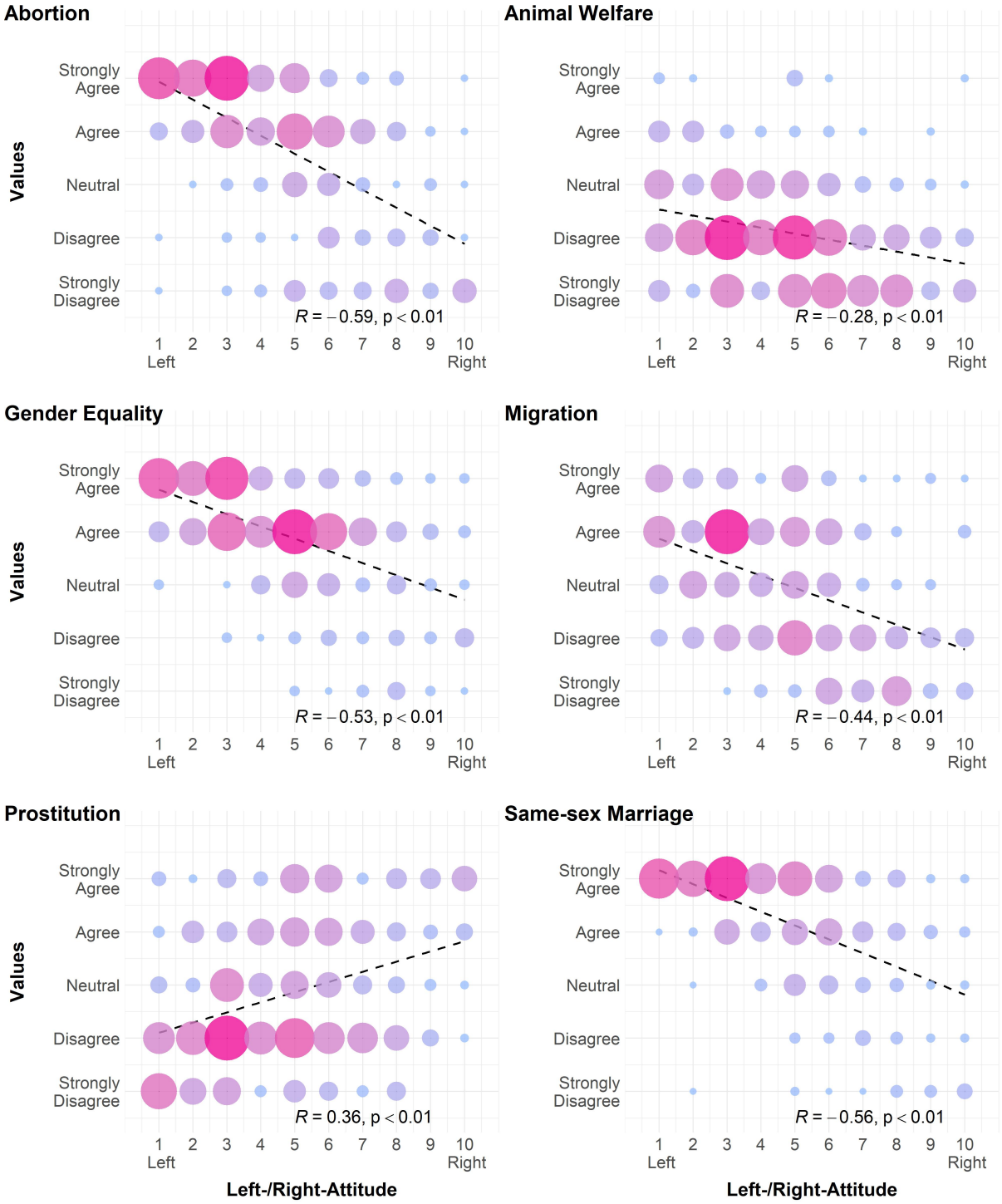
A.1 Supplementary Results for Hypothesis 1

Figure A.1: Relationship between values and beliefs in ValueSalient, by topic



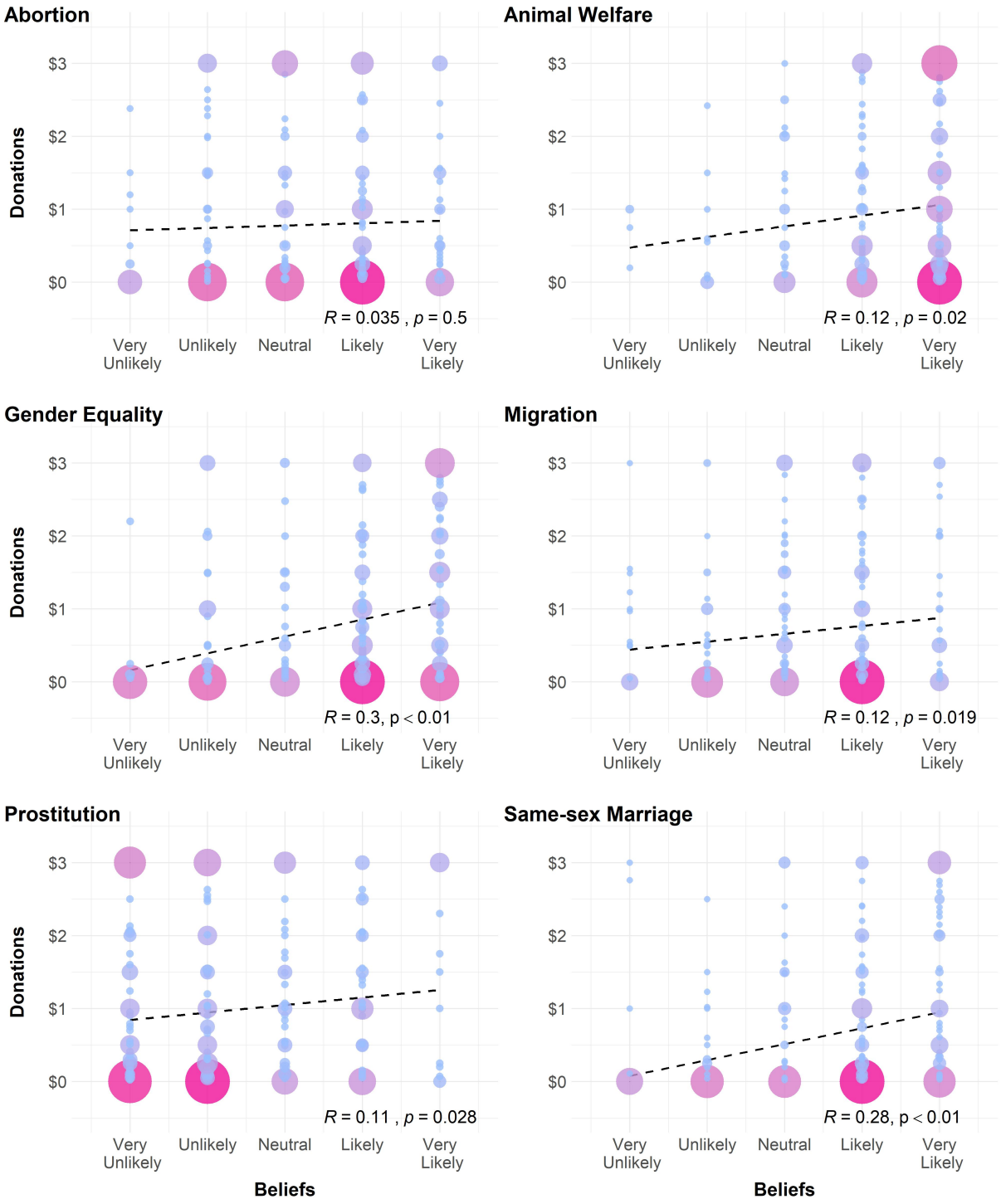
Note: Figure A.1 shows the correlation between values and beliefs in the ValueSalient treatment, separately for each of the six policy debates. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of values on beliefs respectively political attitudes. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

Figure A.2: Relationship between political attitudes and beliefs in ValueSalient, by topic



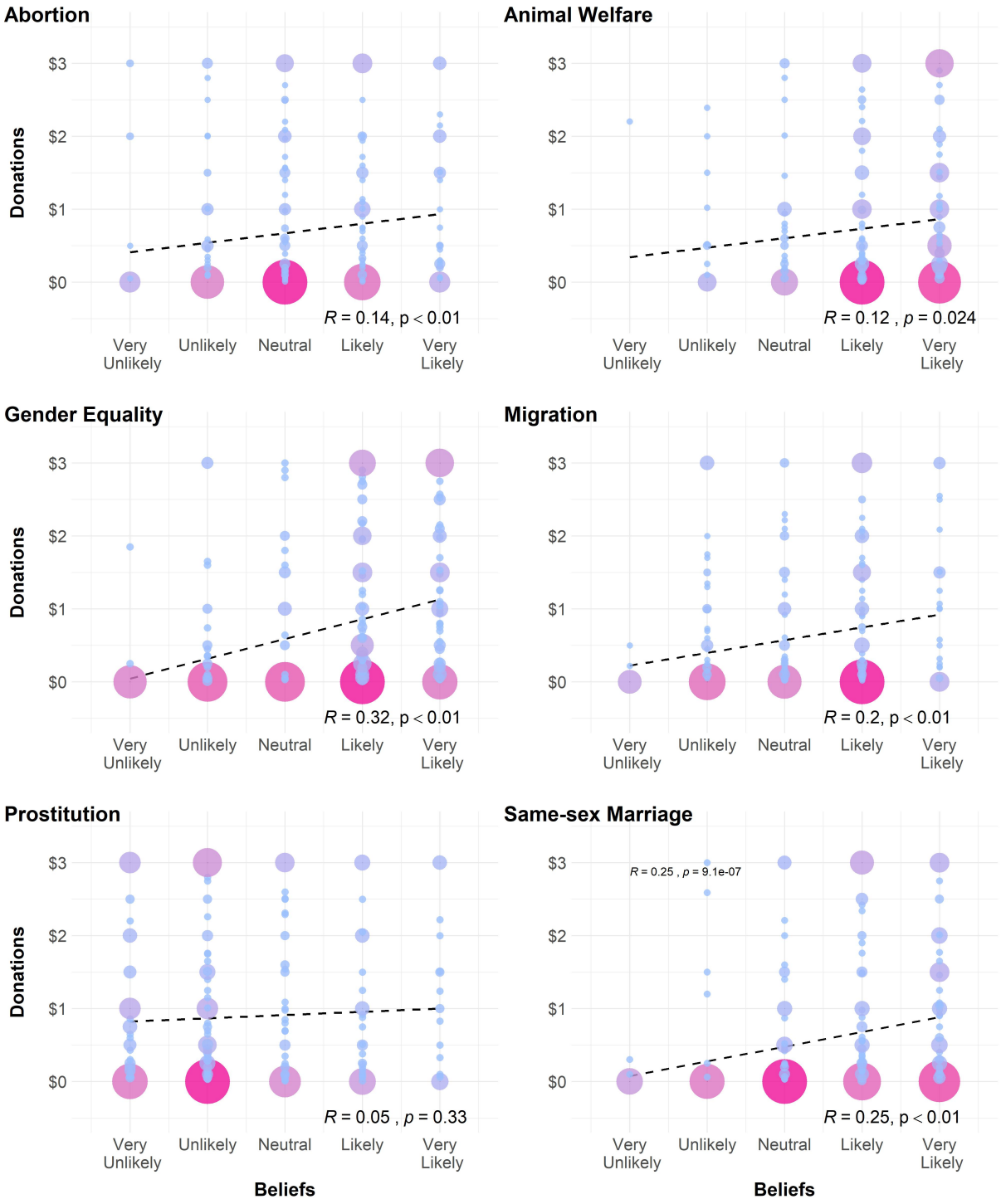
Note: Figure A.2 shows the correlation between moral values and political attitudes in the ValueSalient treatment, separately for each of the six policy debates. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of values on beliefs respectively political attitudes. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

Figure A.3: Relationship between donations and beliefs in ValueNotSalient, by topic



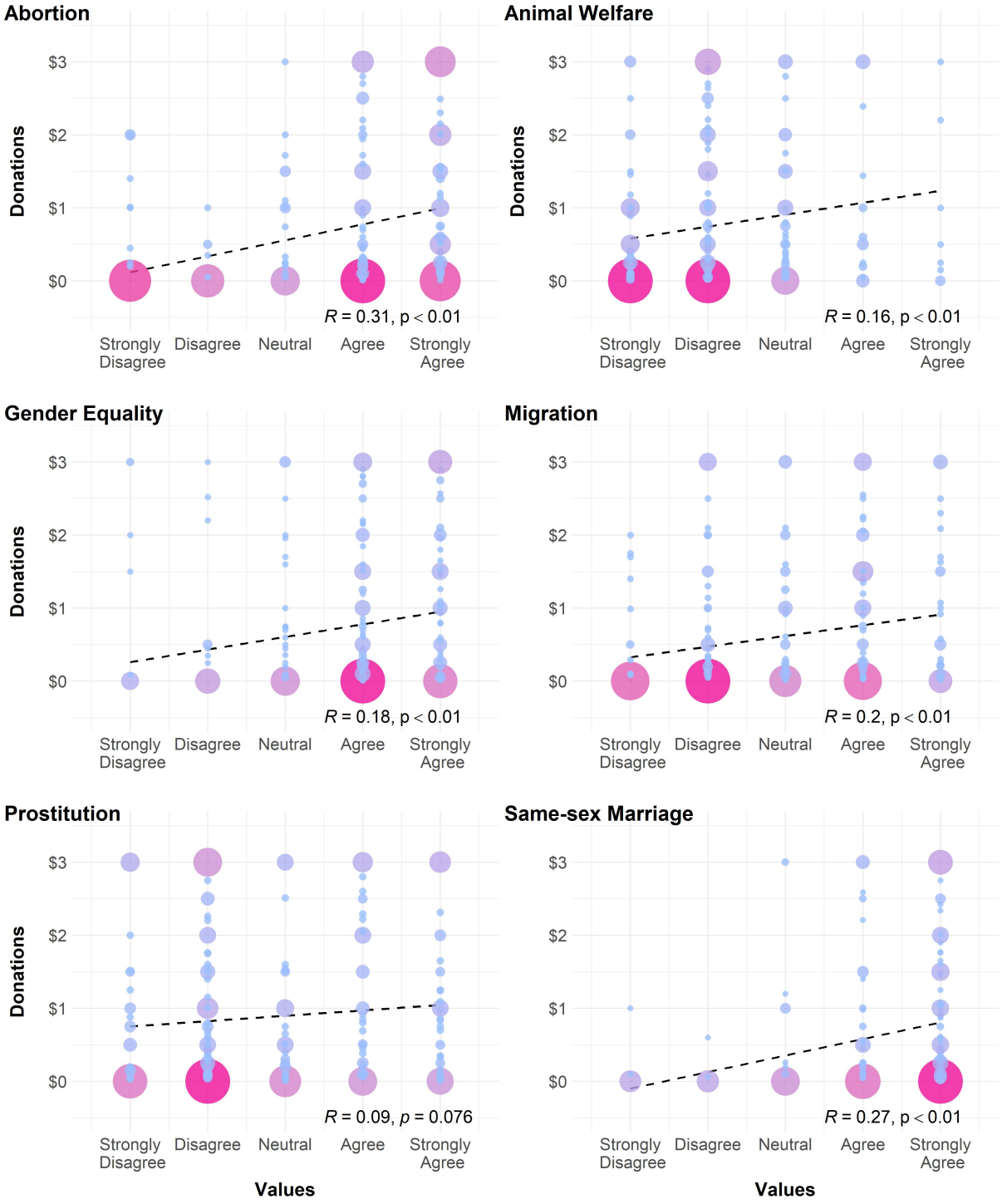
Note: The figure shows the correlation between beliefs and donations in treatment ValueNotSalient separately for each of the six policy debates. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

Figure A.4: Relationship between donations and beliefs in VALUESALIENT, by topic



Note: The figure shows the correlation between beliefs and donations in treatment VALUESALIENT separately for each of the six policy debates. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

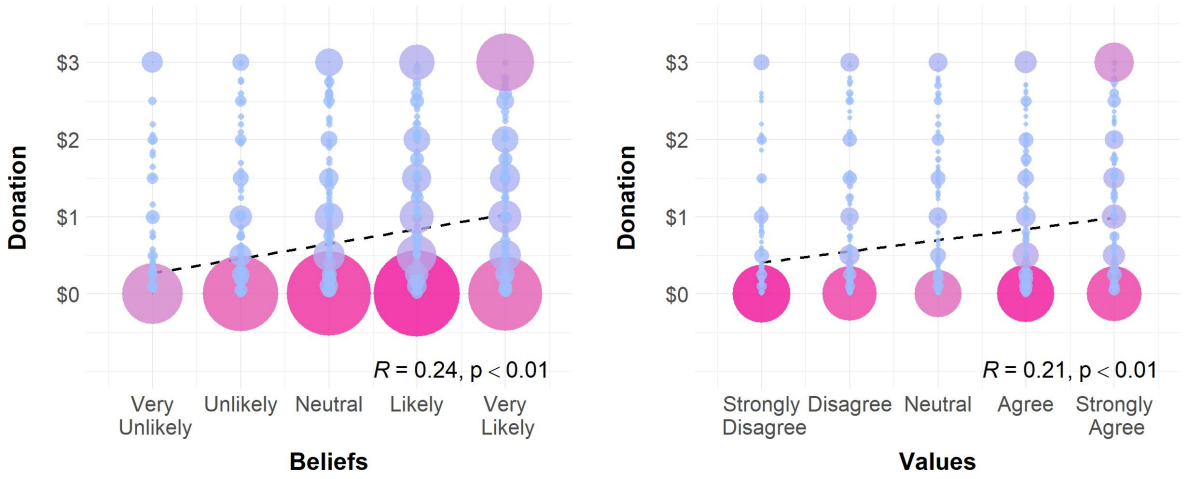
Figure A.5: Relationship between donations and values in VALUESALIENT, by topic



Note: The figure shows the correlation between values and donations in treatment VALUESALIENT separately for each of the six policy debates. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

A.2 Supplementary Results on Beliefs, Values and Donations

Figure A.6: Correlation between Donations and Values/Beliefs in Treatment ConvinceSelf



(a) Donations and Beliefs in CS

(b) Donations and Values in CS

Note: Figure A.6(a) shows the correlation between beliefs and donations in treatment ConvinceSelf, Figure A.6(b) shows the correlation between values and donations in treatment ConvinceSelf. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

Figure A.7: Correlation between Donations and Values/Beliefs in Treatment ConvinceOther

(a) Donations and Beliefs in COR

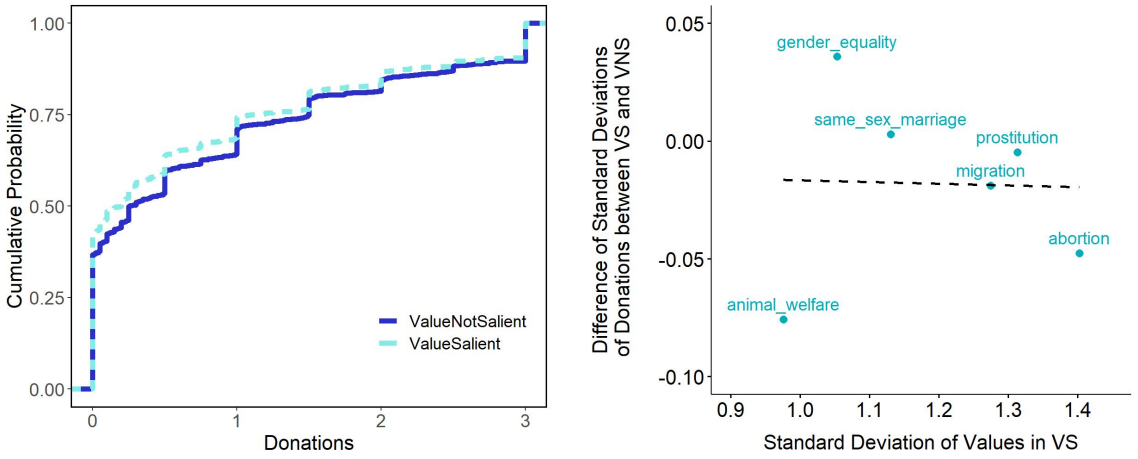
(b) Donations and Values in COR

Note: Figure A.7(a) shows the correlation between beliefs and donations in treatment ConvinceOther, Figure A.7(b) shows the correlation between values and donations in treatment ConvinceOther. The data points are weighted by the number of observations which shows in color and size of the points. The dotted line represents the result of a linear regression of donations on beliefs respectively values. The Pearson correlation coefficient, R, and its p-value are given at the bottom right of each graph.

A.2.1 The Relationship between Hypothesis 2 and Donation Decisions

The results described in this section did not form part of our pre-registration. However, as an additional ex post analysis, we provide documentation of the relationship between the donation decisions observed in the ValueSalient and ValueNotSalient treatment conditions. The general conclusion of the results in this section is that donation decisions were not significantly impacted by the treatment variation. This indicates that although beliefs were shifted by the treatment, this shift did not translate into a change in donation behavior. This result, therefore, contributes to the growing literature that documents a complex relationship between measured beliefs and behavior. While some of the work in this literature documents evidence of beliefs causally affecting behavior in the manner predicted by standard economic models (see, e.g., Costa-Gomes, Huck, and Weizsäcker 2014; Haaland, Roth, and Wohlfart 2020; Barron and Gravert 2021), there is also body of work that show a divergence between predictions and behavior (see, e.g., Costa-Gomes and Weizsäcker 2008; Ivanov, Levin, and Niederle 2010; Haaland and Roth 2021).

Figure A.8: Results for Hypotheses 2a and 2b looking at donations



(a) CDF of Beliefs and Values in VS and VNS

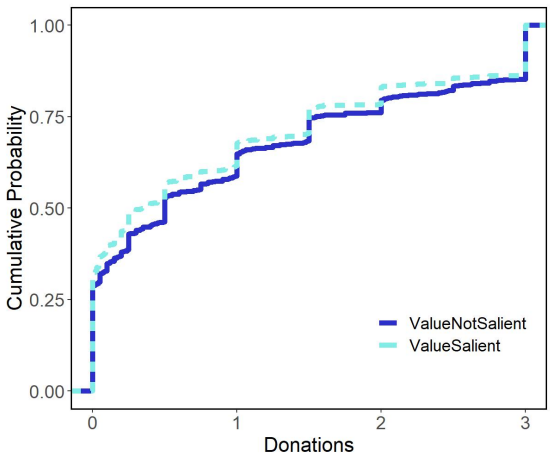
(b) Change in Variance of Beliefs as a Function of Value Spread

Note: Figure 3(a) shows the results for Hypothesis 2a, i.e. the cumulative density function of donations in treatments ValueSalient and ValueNotSalient. Figure 3(b) shows the result for Hypothesis 2b. The y-axis shows the difference of the standard deviations of donations 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.

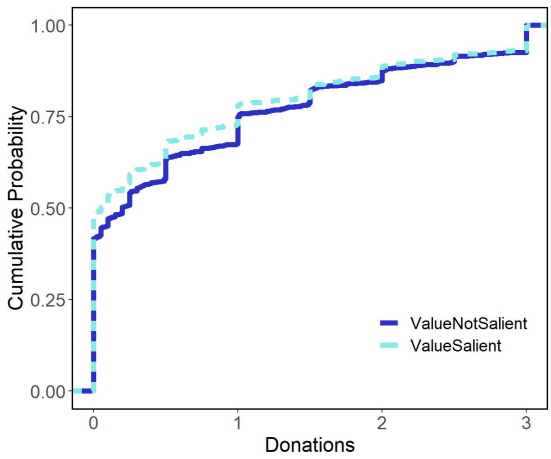
In interpreting these results, it is important to keep in mind that we also observe a strong and robust correlation between donation decisions and both beliefs and values in each of our treatment conditions. There are several reasons why the shift in beliefs might not translate directly into a shift in donation decisions, including the following. First, it may be the

case that deep values are a more important driver of donation decisions than factual beliefs. This would explain the correlations between donations and beliefs and values (since values and beliefs are correlated), but would also be consistent with the fact that the shift in beliefs doesn't translate into a shift in donation decisions. Second, it is plausible that when individuals face their donation decision, the underlying contentious value debate is triggered and becomes salient at the point of making the donation decision. This would potentially negate the treatment differences generates by varying the salience of the value debates introduced by the ValueSalient and ValueNotSalient treatment conditions at the point of making the donation decision.

Figure A.9: Results for Hypotheses 2a and 2b looking at donations



(a) CDF of Beliefs and Values in VS and VNS, subjects on the left of the political spectrum



(b) CDF of Beliefs and Values in VS and VNS, subjects on the right of the political spectrum

Table A1: Influence of increased salience of values on donations

	(1)	(2)	(3)
ValueSalient	-0.071*	-0.019	-0.151**
	[0.041]	[0.072]	[0.062]
Pol. Attitude ($\tilde{p}$)	0.284***	0.512***	0.306***
	[0.042]	[0.060]	[0.055]
ValueSalient	-0.013	-0.114	0.100
× **Pol. Attitude**	[0.059]	[0.086]	[0.077]
Constant	0.671***	0.453***	0.652***
	[0.030]	[0.050]	[0.044]
Observations	4560	2550	3006
Pol. Attitude ($\tilde{p}$) Variable	Left-Right Scale (Left = 1)	Party Affiliation (Democrat = 1)	Last Election (Clinton = 1)

Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: 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. 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 median 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.

A.3 Supplementary Results for Hypothesis 3

Beliefs in Treatment 4a are polarized in comparison to beliefs in Treatment 2:

$$E(b_{4a}|p_{4a} < E(p_{4a})) - E(b_2|p_2 < E(p_2)) \geq E(b_{4a}|p_{4a} > E(p_{4a})) - E(b_2|p_2 > E(p_2)).$$

- Beliefs - Treatment 4a vs. Treatment 2: Using prolific variable 2: Pol. Affiliation (reference group: Republican), same as original test but different measure for political attitude.

Table A2: Outcome: Beliefs

	(1)
Treatment 4a	0.172*
	[0.092]
Democrat	0.652***
	[0.078]
Independent	0.355***
	[0.086]
Treatment 4a × Democrat	-0.063
	[0.108]
Treatment 4a × Independent	-0.157
	[0.119]
Constant	3.003***
	[0.066]
Observations	3654

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

– Beliefs - Treatment 4a vs. Treatment 2: Same as original test but using full left-right scale.

Table A3: Outcome: Beliefs

	(1)
Treatment 4a	0.531***
	[0.199]
Left/Right Attitude=1	0.935***
	[0.151]
Left/Right Attitude=2	0.861***
	[0.154]
Left/Right Attitude=3	0.761***
	[0.144]
Left/Right Attitude=4	0.622***
	[0.153]
Left/Right Attitude=5	0.531***
	[0.144]
Left/Right Attitude=6	0.449***
	[0.148]
Left/Right Attitude=7	0.133
	[0.159]
Left/Right Attitude=8	0.111
	[0.160]
Left/Right Attitude=9	0.133
	[0.183]
Treatment 4a × Left/Right Attitude=1	-0.454**
	[0.226]
Treatment 4a × Left/Right Attitude=2	-0.437*
	[0.226]
Treatment 4a × Left/Right Attitude=3	-0.538**
	[0.221]
Treatment 4a × Left/Right Attitude=4	-0.397*
	[0.233]
Treatment 4a × Left/Right Attitude=5	-0.479**
	[0.219]
Treatment 4a × Left/Right Attitude=6	-0.547**
	[0.227]

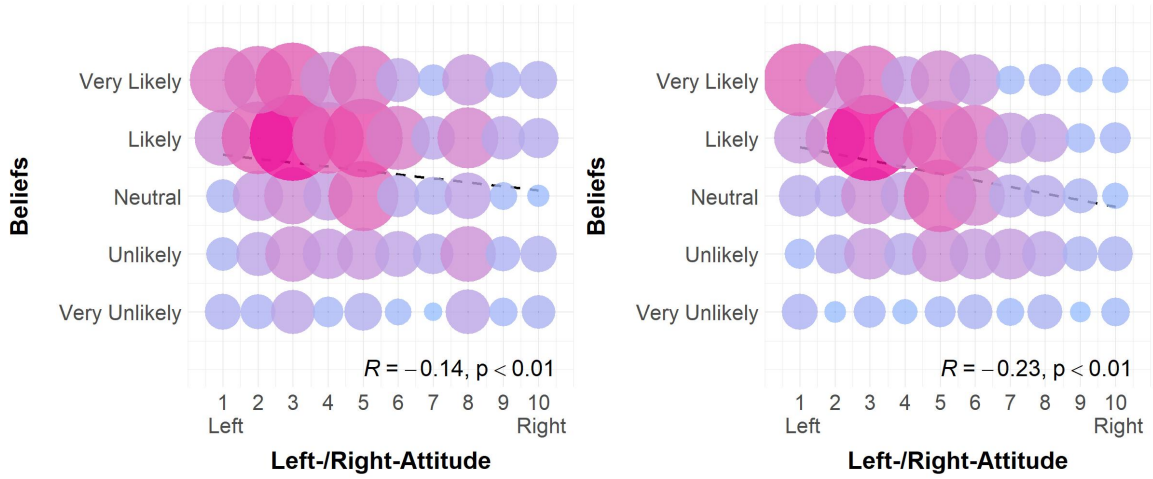
Treatment 4a × Left/Right Attitude=7	-0.148
	[0.248]
Treatment 4a × Left/Right Attitude=8	-0.583**
	[0.235]
Treatment 4a × Left/Right Attitude=9	-0.362
	[0.278]
Constant	2.878***
	[0.130]
Observations	4488

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.4 Further Results on Polarisation

Figure A.10: Results for Hypothesis 2c



(a) Beliefs and Political Attitudes in VNS

(b) Beliefs and Political Attitudes in VS

Additional test: According to Hypothesis 3c, beliefs in Treatment 4a are polarised in comparison to beliefs in Treatment 2 and according to Hypothesis 2c, beliefs in Treatment 2 should be polarised in comparison to beliefs in Treatment 1. Beliefs in Treatment 4a should hence be polarised in comparison to beliefs in Treatment 1.

– Beliefs - Treatment 4a vs Treatment 1, basic test

Table A4: Outcome: Beliefs

	(1)
Treatment 4a	-0.001
	[0.044]
Left (below mean att. - 2)=1	0.290***
	[0.065]
Treatment 4a × Left (below mean att. - 2)=1	0.173*
	[0.089]
Constant	3.398***
	[0.030]
Observations	4428

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

– Beliefs - Treatment 4a vs. Treatment 1: Using prolific variable 2: Pol. Affiliation (reference group: Republican)

Table A5: Outcome: Beliefs

	(1)
Treatment 4a	-0.028
	[0.091]
Democrat	0.358***
	[0.077]
Independent	0.251***
	[0.086]
Treatment 4a × Democrat	0.231**
	[0.109]
Treatment 4a × Independent	-0.053
	[0.120]
Constant	3.202***
	[0.064]
Observations	3606

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

– Beliefs - Treatment 4a vs. Treatment 1: Using full left-right scale

Table A6: Outcome: Beliefs

	(1)
Treatment 4a	0.252
	[0.197]
Left/Right Attitude=1	0.585***
	[0.151]
Left/Right Attitude=2	0.492***
	[0.143]
Left/Right Attitude=3	0.384***
	[0.138]
Left/Right Attitude=4	0.343**
	[0.145]
Left/Right Attitude=5	0.340**
	[0.139]
Left/Right Attitude=6	0.259*
	[0.152]
Left/Right Attitude=7	0.088
	[0.170]
Left/Right Attitude=8	-0.015
	[0.146]
Left/Right Attitude=9	0.102
	[0.172]
Treatment 4a × Left/Right Attitude=1	-0.104
	[0.229]
Treatment 4a × Left/Right Attitude=2	-0.067
	[0.222]
Treatment 4a × Left/Right Attitude=3	-0.161
	[0.220]
Treatment 4a × Left/Right Attitude=4	-0.117
	[0.232]
Treatment 4a × Left/Right Attitude=5	-0.288
	[0.219]
Treatment 4a × Left/Right Attitude=6	-0.358
	[0.234]
Treatment 4a × Left/Right Attitude=7	-0.103

	[0.259]
Treatment 4a × Left/Right Attitude=8	-0.457**
	[0.230]
Treatment 4a × Left/Right Attitude=9	-0.330
	[0.276]
Constant	3.157***
	[0.122]
Observations	4428

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Sample Balance

Table B1: Sample Balance.

	Full	Treatment				
	Sample	(1)	(2)	(3)	(4a)	(4b)
Age	44.09	44.10	43.82	45.04	44.49	42.97
	(15.762)	(15.874)	(15.656)	(15.706)	(15.770)	(15.820)
Female	0.52	0.51	0.51	0.52	0.51	0.54
	(0.500)	(0.501)	(0.500)	(0.500)	(0.501)	(0.499)
Etnicity						
Asian	0.06	0.07	0.06	0.07	0.06	0.06
	(0.245)	(0.254)	(0.232)	(0.249)	(0.244)	(0.244)
Black	0.13	0.16	0.12	0.11	0.14	0.12
	(0.336)	(0.365)	(0.331)	(0.308)	(0.345)	(0.330)
Mixed	0.03	0.02	0.03	0.02	0.02	0.05
	(0.171)	(0.136)	(0.181)	(0.144)	(0.156)	(0.223)
Other	0.02	0.02	0.02	0.02	0.03	0.02
	(0.143)	(0.126)	(0.134)	(0.125)	(0.172)	(0.156)
White	0.76	0.74	0.77	0.79	0.74	0.74
	(0.430)	(0.440)	(0.424)	(0.408)	(0.437)	(0.442)
Employment Status						
Starting new job	0.01	0.02	0.02	0.01	0.01	0.01
within next month	(0.108)	(0.126)	(0.134)	(0.089)	(0.091)	(0.091)
Full-Time	0.47	0.47	0.46	0.49	0.47	0.44
	(0.499)	(0.500)	(0.499)	(0.501)	(0.500)	(0.496)
Not in paid work	0.21	0.20	0.19	0.22	0.23	0.21
	(0.408)	(0.401)	(0.393)	(0.417)	(0.421)	(0.411)
Other	0.04	0.03	0.05	0.05	0.03	0.05
	(0.203)	(0.176)	(0.217)	(0.219)	(0.179)	(0.217)
Part-Time	0.18	0.19	0.20	0.16	0.17	0.19
	(0.386)	(0.394)	(0.402)	(0.366)	(0.372)	(0.393)
Unemployed	0.09	0.09	0.08	0.07	0.09	0.10
(job seeking)	(0.280)	(0.280)	(0.268)	(0.258)	(0.292)	(0.303)
Do not know/	0.00	0.00	0.00	0.01	0.00	0.00
not applicable	(0.046)	(0.052)	(0.000)	(0.073)	(0.000)	(0.052)
Doctorate degree	0.03	0.03	0.03	0.03	0.04	0.04
	(0.181)	(0.161)	(0.174)	(0.169)	(0.206)	(0.193)
Observations	1863	375	385	377	363	363

Continued on next page.

Table B1 – continued from previous page

	Full	Treatment				
	Sample	(1)	(2)	(3)	(4a)	(4b)
Graduate degree	0.16	0.18	0.14	0.17	0.17	0.13
	(0.366)	(0.386)	(0.350)	(0.378)	(0.372)	(0.339)
High school diploma	0.22	0.23	0.18	0.22	0.21	0.26
	(0.414)	(0.423)	(0.386)	(0.413)	(0.407)	(0.437)
No formal	0.00	0.01	0.00	0.00	0.00	0.00
qualifications	(0.052)	(0.073)	(0.051)	(0.052)	(0.000)	(0.052)
Secondary education	0.02	0.01	0.01	0.02	0.03	0.02
	(0.130)	(0.073)	(0.113)	(0.153)	(0.164)	(0.128)
Technical/	0.17	0.16	0.19	0.15	0.17	0.18
community college	(0.376)	(0.367)	(0.395)	(0.361)	(0.374)	(0.382)
Undergraduate degree	0.40	0.39	0.44	0.40	0.39	0.37
	(0.489)	(0.488)	(0.497)	(0.490)	(0.487)	(0.485)
Income						
Less than $10000	0.06	0.07	0.05	0.07	0.07	0.06
	(0.242)	(0.250)	(0.222)	(0.249)	(0.249)	(0.239)
$10000-$15999	0.06	0.08	0.05	0.06	0.07	0.04
	(0.236)	(0.267)	(0.211)	(0.244)	(0.254)	(0.193)
$16000-$19999	0.03	0.03	0.04	0.02	0.03	0.03
	(0.171)	(0.176)	(0.200)	(0.135)	(0.164)	(0.172)
$20000-$29999	0.11	0.12	0.09	0.11	0.10	0.13
	(0.313)	(0.322)	(0.292)	(0.315)	(0.299)	(0.336)
$30000-$39999	0.10	0.11	0.12	0.11	0.09	0.10
	(0.304)	(0.309)	(0.325)	(0.308)	(0.280)	(0.296)
$40000-$49999	0.10	0.11	0.08	0.09	0.12	0.11
	(0.303)	(0.309)	(0.268)	(0.291)	(0.330)	(0.314)
$50000-$59999	0.10	0.09	0.10	0.10	0.10	0.10
	(0.298)	(0.288)	(0.302)	(0.294)	(0.303)	(0.303)
$60000-$69999	0.07	0.06	0.08	0.09	0.07	0.06
	(0.254)	(0.230)	(0.268)	(0.283)	(0.249)	(0.234)
$70000-$79999	0.08	0.09	0.09	0.08	0.09	0.07
	(0.273)	(0.280)	(0.280)	(0.267)	(0.284)	(0.254)
$80000-$89999	0.05	0.05	0.05	0.04	0.05	0.05
	(0.214)	(0.208)	(0.222)	(0.202)	(0.223)	(0.217)
$90000-$99999	0.05	0.04	0.05	0.04	0.05	0.05
Observations	1863	375	385	377	363	363

Continued on next page.

Table B1 – continued from previous page

	Full	Treatment				
	Sample	(1)	(2)	(3)	(4a)	(4b)
	(0.208)	(0.196)	(0.211)	(0.196)	(0.223)	(0.212)
$100000-$149999	0.11	0.10	0.13	0.12	0.09	0.13
	(0.319)	(0.306)	(0.334)	(0.331)	(0.280)	(0.339)
More than $150000	0.05	0.04	0.05	0.05	0.06	0.04
	(0.218)	(0.196)	(0.227)	(0.224)	(0.244)	(0.193)
Prefer not to say	0.03	0.03	0.02	0.02	0.02	0.04
	(0.162)	(0.176)	(0.151)	(0.144)	(0.138)	(0.193)
Prolific Score	99.58	99.59	99.67	99.60	99.52	99.52
	(1.366)	(1.242)	(1.158)	(1.276)	(1.682)	(1.430)
Left/Right Attitude	4.60	4.76	4.71	4.52	4.42	4.58
	(2.552)	(2.556)	(2.449)	(2.621)	(2.537)	(2.594)
Observations	1863	375	385	377	363	363

Note: The table shows the means and standard deviations (in parenthesis) of demographic variables for the individuals in our sample. The first column provides the information on the whole sample, column 2 (Treatment (1)) looks at treatment ValueNotSalient, column 3 (Treatment (2)) at treatment ValueSalient, column 4 (Treatment (3)) at treatment ConvinceSelf, column 5 (Treatment (4a)) at treatment ConvinceOther, column 6 (Treatment (4b)) at treatment BeingConvinced.

C Preregistration Document

Morals, Beliefs, and Actions

Kai Barron (WZB), Anna Becker (UCL), Steffen Huck (UCL and WZB)

This version: 14/01/2020

PART I: EXPERIMENTAL DESIGN

Setup of Experiment

This experiment will be run as an online survey. The sample for the four main treatments will consist of 1,500 individuals that are representative of the US population in terms of age, sex, and ethnicity. An additional 375 subjects will be recruited later for an auxiliary treatment, Treatment 4b, which builds on subjects' choices in Treatment 4a. All participants will be recruited via the online platform Prolific. Subjects will be paid £3 for participation and have the option to earn a bonus in Treatments 1, 2, 3 and 4b.[20] A strict no-deception policy will be followed. The experiment is programmed using the experimental software o-Tree (Chen, Schonger, and Wickens (2016)).

Experimental Design

At the beginning of the experiment, subjects are randomized into one out of four treatment groups such that each group consists of 375 subjects. The four treatments are described in the following text.

Treatment 1 – "Control"

Part 1

Subjects are asked to state how likely they think a statement that they are presented with is true. The statements have been chosen such that they can be associated with a policy domain and typically refer to facts on which scientific consensus has not been reached yet. Subjects use a five-point Likert scale to indicate their beliefs. They can choose between the following options: "Very Unlikely", "Very Likely", "Neutral", "Likely" and "Very Likely". After a waiting time of 15 seconds subjects can proceed to the next page.

This is repeated six times for six different policy domains. These are migration, animal welfare, gender equality, abortion, prostitution and gay rights. Table 1 in the Appendix provides an overview over all domains and the statements presented to subjects. The order with which the statements on the different domains are shown to participants is randomised at the individual level.

20. The show-up fee is converted into US dollars, since the subjects are recruited from the USA.

Part 2

After subjects have submitted their beliefs on the six different issues, they are informed that they have the option to make six donations to charities. They are also informed that one out of the six decisions they are about to make will be chosen at random to be implemented.

Subjects then see the statement they were earlier confronted with, the belief that they stated and the option to donate to a charity that is active in the respective policy domain. As in Part 1, this process is repeated six times following the same order of domains as in Part 1. Subjects are informed briefly about the aims of the charities and can indicate with a slider how much they would like to donate. They are provided with $3 and can choose to donate any amount between $0 and $3. Subjects will be paid the remaining amount (i.e. what they decided not to donate) at the end of the experiment. They can proceed to the next page at any time after a waiting time of 15 seconds.

Treatment 2 – "Moral Values"

Part 1

Treatment 2 is similar to Treatment 1, with the following exceptions.

Different to Treatment 1, in Treatment 2 subjects will also be asked to state some moral values that are related to the same six policy domains. The question of how much they agree or disagree with a moral statement they are presented with appears above the question regarding the factual statement. Subjects use a five-point Likert scale to indicate their agreement. They can choose between the following options: "Strongly Disagree", "Disagree", "Neutral", "Agree" and "Strongly Agree". The moral statements can be found in Table 1 in the Appendix.

Part 2

Different to Treatment 1, in Treatment 2, subjects are also reminded of the moral values they stated in Part 1. The screen will therefore show the moral statement and the subject's choice above the factual statement and the subject's decision and then offers the subjects to donate to a proposed charity.

Treatment 3 – "Convincing Yourself"

As in Treatment 2, subjects are asked to state moral values, factual beliefs, and then make a charitable donation. The key difference between Treatment 3 and Treatment 2 is that in Treatment 3 the moral statement and the factual statement are presented to subjects on the same screen as the option to donate to charity, which is presented at the bottom of the page. They receive the same information as subjects in Part 2 of Treatments 1 and 2 and

face the same charitable giving decision, but in Treatment 3 all three decisions are made on the same page (i.e. the beliefs, moral value judgments, and charitable donations).

As in the other treatments, this will be repeated six times for the six different policy domains where the order is randomized on the individual level. Subjects use a Likert scale to indicate their values and beliefs and a slider to indicate how much they would like to donate. They are provided with $3 per decision. One of those decisions will be chosen at random to be implemented of which subjects are informed about in advance. They will be paid what they decide not to donate after the experiment.

Treatment 4a – "Convincing Others"

As in Treatment 2, subjects are asked to state their moral values and their factual beliefs. The moral statement and the factual statement are presented to subjects on the same screen underneath each other as in Part 1 of Treatment 2.

Before stating values and beliefs, the subject is informed that another participant will have the option to donate to a related charity. Importantly, the other participant will make their charitable donation decision after being informed about the moral values and factual beliefs that the subject reports (i.e. the moral values and factual beliefs reported by subjects in Treatment 4a will be sent to subjects in Treatment 4b before participants in 4b make their charitable donation decisions).

Subjects in Treatment 4a are presented with the information the other participant in Treatment 4b will be shown about the charity. The donation decision that will be completed by Treatment 4b subjects is the same as in the previous Treatments, i.e. the participant has $3 available of which they keep what they decide not to donate. Both the subjects in Treatment 4a and the other participant in Treatment 4b will be informed in advance that one of the six decisions will be chosen at random to be implemented.

After the main Treatments 1, 2, 3 and 4a have been run, another 375 subjects will be recruited for Treatment 4b:

Treatment 4b – "Being Convinced"

Part 1

This will be identical to Part 1 of Treatment 2.

Part 2

In Treatment 4b, Part 2 will be similar to Treatment 2, with the exception that instead of subjects being reminded of their own decisions regarding the moral and the factual statements,

subjects are now informed about the decisions of a participant from Treatment 4a when they make their charitable donation decision. As before, subjects will be provided with $3 for each decision of which they will be paid what they decide not to donate. They are also informed that one out of their six decisions will be chosen at random to be implemented.

Post-experimental Survey

After the experiment, all subjects will be asked to fill out a survey. The survey covers the following topics:

1. Personal Details
2. Political Attitude
3. Religious Attitude
4. Moral Foundations Questionnaire[21]
5. Questions on Moral Behavior
6. Cognitive Reflection Test

PART II: ANALYSIS PLAN

II.1) Introduction

Standard theories on belief formation typically disregard the desire of individuals to gather, avoid or interpret information in a way that serves non-instrumental purposes. A large recent literature has, however, shown that, for example, self-serving biases, wishful thinking and motivated reasoning are important determinants of belief formation. This project studies individuals' moral values as a potential source for motivated cognition and links it to partisan disagreement about factual statements, i.e. polarisation.[22]

We proceed in three steps. In each step, we test a small set of hypotheses that are interlinked by a common underlying idea. In the first step, we seek to test whether individuals report moral values and factual beliefs that are aligned in the domain of political issues. Potential reasons why individuals might do this include: i) avoiding emotional discomfort, or cognitive dissonance emanating from holding or stating incoherent values and beliefs, and ii) using value and belief statements to justify self-interested actions (e.g. actions that increase the individual's material wealth). From the analysts' perspective, the presence of such belief-value constellations would provide a basis for taking an individual's moral values into consideration in trying to understand belief formation regarding factual statements. A

21. https://moralfoundations.org/questionnaires/, accessed 31/12/2019.
22. See for example Rabin (1995) for a theory on self-serving biases in moral reasoning.

potential motive for this could be a desire to establish something akin to a "moral identity" (Bénabou and Tirole (2011)).

In the second step, we then test whether there is a systematic pattern in the way factual beliefs are formed that may be partially responsible for the creation of these belief-value constellations (i.e. we ask whether factual beliefs are constructed in a way that forms these belief-value constellations). To do this, we compare the distribution of beliefs of individuals that were previously asked about, and hence reminded of, their related values (Treatment 2) with the distribution of beliefs of individuals in the control Treatment 1 (where no moral value statements are reported prior to stating factual beliefs). This allows us to assess the influence of being primed to think about the particular factual belief in question through the lens of the related value debate. Following on from this, we ask whether this mechanism can explain the recent trend of a polarization of beliefs in society that has been demonstrated to run along ideological lines (see, e.g., Gentzkow (2016)). In particular, there appears to be an increased disagreement about objective facts among members of society that is associated with political attitudes.[23] We hypothesize that the heterogeneity in moral values between different political groups may be leading to the formation of these polarized factual beliefs. Hence, we test whether factual beliefs become polarized as a function of political attitudes (e.g. as a result of individuals' desire to adjust their beliefs to their moral values.)[24]

Lastly, we study two potential forces that might increase or decrease the degree of polarization of (stated) beliefs. First, we look at the impact of financial incentives (through motivated reasoning or self-persuasion), and second we will study the role of persuading others, which is particularly relevant in political contexts.

The first channel which considers the role of financial incentives on belief formation is important because there are many reasons why individuals might face costs to hold certain beliefs or values. For example, it may be costly to hold different beliefs and values to those held by individuals in one's peer network.[25] Alternatively, holding particular beliefs and values may be costly when they induce the individual to take a particular costly action. For example, an individual who advocates the merits reducing inequality in society may feel compelled to take actions that reduce their own wealth in order to increase the wealth of a poorer individual. Rather than studying abstract costs (e.g. incoherence with one's peers'

23. The most prominent current example is probably that of climate change where there is a widening gap in the views on the scientific evidence between Republicans and Democrats in the US (see e.g. McCright and Dunlap (2011)).

24. In his theoretical work, Le Yaouanq (2021) 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.

25. On the role of group identity in belief polarization see e.g. Gennaioli and Tabellini (2018).

beliefs), we focus on the latter type of costs that accrue due to taking actions that reduce one's personal payment. In particular, we consider costly donation decisions and hypothesize that this cost leads individuals to bias their (stated) beliefs and values when they expect a related donation decision, with the bias operating in the opposite direction to the beliefs and values consistent with a higher donation.[26]

Second, we look at the potential role of individuals' desire to convince others to take actions that are in line with their own values. This motive of convincing others might lead people to further align their (stated) beliefs with their political agenda or goals, and perhaps to exaggerate these stated beliefs. We study whether subjects adjust their beliefs to be more extreme in order to convince another participant to give more or less to a suggested charity. In this case, we test whether individuals overstate the strength of their beliefs when trying to persuade another person to act in a certain way.

Section II.2) introduces the necessary notation, before we formalize our hypotheses in Section II.3).

C.1 Notation

Let b_t denote the factual beliefs stated by individuals in Treatment t, where $t \in \{1, 2, 3, 4a, 4b\}$, v_t are the moral values stated by individuals in treatment t where $t \in \{2, 3, 4a, 4b\}$ and d_t are the donation decisions of individuals in treatment t where $t \in \{1, 2, 3, 4b\}$. Let F_{b_t} denote the cumulative distribution function (cdf) of factual beliefs in Treatment t where $t \in \{1, 2, 3, 4a, 4b\}$, F_{v_t} the cdf of moral values in Treatment t where $t \in \{2, 3, 4a, 4b\}$, and F_{d_t} the cdf of donations in Treatment t where $t \in \{1, 2, 3, 4b\}$.

Let p_t denote the left-right political stance of individuals in Treatment t which will be elicited for all participants in the post-experimental survey using a Likert scale ranging from 1 to 10 where 1 is left and 10 is right, i.e. p_t is increasing in the degree to which an individual positions herself on the right of the political spectrum. F_{p_t} denotes the respective cdf. Belief and value statements were chosen such that the factual statement being true would provide support for agreement to the value statement. All statements are coded this way for the analysis but not necessarily presented to participants like this (Table 1 in the Appendix shows how statements are presented to subjects during the experiment).

At the same time, we recode all the moral value variables such that they are likely to be increasingly appealing as one moves from the right to the left of the political spectrum.

26. It is also possible that there are individuals that now bias their values and beliefs in the same direction as beliefs and values consistent with a higher donation as they consider them to be more important when relevant to justify a charitable donation. Our prior, however, is that the effect described above dominates.

Similarly, we code the factual belief variables such that if they are true, they support moral value positions typically held by individuals on the political left.[27] Charities are chosen such that if the moral statement is supported it would justify the charities' objectives.[28]

C.2 Hypotheses

C.2.1 Belief-Value Constellations

As stated above, we begin by testing whether individuals report moral values and factual beliefs on political issues that are aligned, whether their moral values are aligned with their political attitudes and whether factual beliefs and moral values are related to decision making in the form of costly charitable donation choices.

Hypothesis 1

a) Moral values are positively correlated with beliefs:

$$Corr(v_2, b_2) \geq 0.$$

b) Moral values are negatively correlated with political attitudes:

$$Corr(v_2, p_2) \leq 0.$$

c) Donations are positively correlated with beliefs and values:

$$Corr(d_1, b_1) \geq 0, Corr(d_2, b_2) \geq 0, Corr(d_2, v_2) \geq 0.$$

Part a) of Hypothesis 1 tests whether moral values and factual beliefs reported by subjects in Treatment 2 are aligned. Recall that in Treatment 2 subjects are not aware of the opportunity to donate by the time they state their values and beliefs.

Part b) tests for a negative correlation between moral values and political attitudes. For example, Enke, Rodriguez-Padilla, and Zimmermann (2016) show that an individuals' moral type is strongly correlated with their political affiliation. Rather than looking at predefined moral types we look at concrete moral convictions with regard to certain policy domains.

27. The policy domain "Prostitution" is an ambiguous case. Individuals from the political left could both support and reject more liberal prostitution rights.

28. Therefore, it is worth pointing out that variables are coded such that a higher value of b_t, v_t, and d_t should be consistent with a lower value of p_t according to the researcher team's priors.

We expect that the further an individual is on the left of the political spectrum (i.e. the lower is p_2) the more likely they are to agree with the moral value statement.

In Part c) of Hypothesis 1, we hypothesize that individuals donate more when their moral values and their beliefs are such that the cause of the charity is justified by them (i.e. that donations are positively correlated with beliefs and moral values consistent with the charity's mandate).

C.2.2 Construction of Beliefs

The following hypothesis tests: (i) whether the formation of factual beliefs is influenced by the values individuals hold, thereby explaining the formation of belief-value constellations, and (ii) whether this influence of values on belief formation may lead to a polarization of factual beliefs across the political spectrum.

Hypothesis 2

a) The distribution of factual beliefs in Treatment 1 is different from the distribution of factual beliefs in Treatment 2:

$$F_{b_1} \neq F_{b_2}.$$

b) Comparing across the six domains indexed by m, the difference between the variance in beliefs in Treatment 2 and the variance in beliefs in Treatment 1 is increasing in the variance in values in Treatment 2:

$$\frac{d[Var(b_2^m) - Var(b_1^m)]}{d[Var(v_2^m)]} \geq 0.$$

c) Beliefs in Treatment 2 are more polarized than beliefs in Treatment 1:

$$E(b_2|p_2 < E(p_2)) - E(b_1|p_1 < E(p_1)) \geq E(b_2|p_2 > E(p_2)) - E(b_1|p_1 > E(p_1)).$$

Part a) of Hypothesis 2 exploits the fact that unlike subjects in Treatment 2, subjects in Treatment 1 are not asked to state their moral values. We test whether reminding subjects of their moral values has an impact on the distribution of factual beliefs which would indicate that moral values are relevant for the construction of beliefs. In Part b) of Hypothesis 2, we go further and test whether values exert a systematic influence on belief formation. In particular, Part b) of Hypothesis 2 posits that when there is more dispersion in the values that subjects hold with regard to a certain policy domain, we expect that there will be a shift towards more extreme beliefs as subjects are drawn towards more coherent belief-value

constructions.

The last part of Hypothesis 2 tests whether a polarization in beliefs can be explained as a result of the posited impact of values on belief formation. The inequality in Part c) states that the difference in the means of beliefs between Treatment 2 and Treatment 1 is greater for subjects below the mean of political attitudes, i.e. relatively on the left of the political spectrum, than for subjects above the mean of political attitudes, i.e. relatively on the right of the political spectrum. To put this another way, the hypothesis states that individuals on the left will increase their beliefs between Treatment 1 and Treatment 2 more than individuals on the right of the political spectrum, on average. Note that this also includes the case where subjects adjust their beliefs downwards (i.e. it is completely consistent with individuals on the right shifting their beliefs downward between Treatment 1 and 2, which would make the right-hand side negative).

If political attitudes are sufficiently widely dispersed, we would expect a positive left-hand side and a negative right-hand side, i.e. what we traditionally refer to as polarization. Otherwise, we expect to see a mild form of polarization where beliefs are adjusted to different extents by those on the left and the right of the political spectrum within our sample.

C.2.3 Convincing Yourself and Convincing Others

The last hypothesis is split in two parts. In Part A, we look the role of self-serving biases that allow subjects to justify selfish behaviour and are expected to lead to a downward[29] bias in beliefs, values and charitable donations. Part B, on the other hand, studies whether introducing the opportunity to convince another participant to take an action that is in line with one's moral values can lead to a greater polarization of beliefs.

Hypothesis 3

A. **Convincing Yourself**

a) When there is an increase in the cost of holding certain beliefs and values individuals may shift their stated beliefs and values in the opposite direction:

i) b_2 first-order stochastically dominates b_3, i.e. $F_{b_2} \leq F_{b_3}$.

ii) v_2 first-order stochastically dominates v_3, i.e. $F_{v_2} \leq F_{v_3}$.

29. Here, "downward" refers to a bias in the direction that is consistent with being self-serving. In terms of the way we have defined our variables, it will also refer to lower values of b_t, v_t, and d_t.

b) Donations in Treatment 3 are lower than in Treatment 2:

$$E(d_2) \geq E(d_3).$$

B. **Convincing Others**

c) Beliefs in Treatment 4a are polarized in comparison to beliefs in Treatment 2:

$$E(b_{4a}|p_{4a} < E(p_{4a}))-E(b_2|p_2 < E(p_2)) \geq E(b_{4a}|p_{4a} > E(p_{4a}))-E(b_2|p_2 > E(p_2)).$$

In Treatments 1, 2 and 3 subjects are given the opportunity to donate to charity. Donating comes at the cost of a foregone bonus payment. 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). In Part a) of Hypothesis 3 we test whether subjects shift their values and beliefs downwards to justify smaller donations in order to receive higher bonuses.

In Part b), we hypothesize that individuals donate less on average in Treatment 3 than in Treatment 2. In both treatments, a higher donation reduces the payment that the subject receives herself (i.e. the material incentives are identical across the two treatments). However, in Treatment 2, individuals only learn about the possibility to donate after they have already stated their values and beliefs which rules out the opportunity to make any adjustments to one's values in order to justify self-interested charitable donation decisions.[30]

The last part of Hypothesis 3 (i.e. Part c) tests whether individuals report more polarized beliefs when they have the opportunity to convince someone to make a donation which is the case in Treatment 4a. As in Hypothesis 2 c), the inequality in Hypothesis 3c) states that the difference in the means of beliefs between Treatment 4a and Treatment 2 is greater for subjects below the mean of political attitudes, i.e. relatively on the left of the political spectrum, than for subjects above the mean of the political spectrum, i.e. relatively on the right of the political spectrum.

As a spillover from Hypothesis 3 and Treatment 4a we can also study the effect of persuasion

30. There is the possibility that there exists a subset of individuals in Treatment 3, who instead of shifting down their reported beliefs and values to justify a lower charitable donation decision, instead shift up their values and beliefs in order to then enhance the signalling value of their donations. These individuals might then also donate more when facing Treatment 3 in comparison to Treatment 2. This effect would operate in the opposite direction to the main effect hypothesised in Hypothesis 3.b). For simplicity, we have stated Hypotheses 3.b) in terms of the average effect for the entire sample, working under the assumption that the main hypothesised effect of adjusting one's beliefs and values downwards in a self-serving fashion will dominate this potential countervailing secondary effect.

on recipients of the persuasion messages by looking at Treatment 4b. More specifically, we can study whether donations to charity increase when an individual sees their own values and beliefs confirmed by another person. We would expect that there will be a polarization of donation decisions when the political attitudes of the sender and the receiver are aligned.[31] In cases where the sender's and the receiver's political attitude are not aligned, we can think of (at least) two opposing possible effects. Individuals might either doubt their own convictions and reassess them or they might want to prove the sender wrong by exaggerating or lowering the donated amount. Ex ante, it is unclear which effect is expected to dominate. This in turn, might also depend on political affiliation. We do not propose any hypothesis here, but will document the data in an exploratory analysis.

31. This could be tested similarly to Part c) of Hypotheses 2 and 3.

References

Bénabou, Roland, and Jean Tirole. 2011. "Identity, morals, and taboos: Beliefs as assets." *The Quarterly Journal of Economics* 126 (2): 805–855.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree – An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

Enke, Benjamin, Ricardo Rodriguez-Padilla, and Florian Zimmermann. 2016. "Moral Universalism and the Structure of Ideology." *Working Paper*.

Exley, Christine L. 2015. "Excusing Selfishness in Charitable Giving: The Role of Risk." *Review of Economic Studies* 83, no. 2 (October): 587–628.

———. 2020. "Using Charity Performance Metrics as an Excuse Not to Give." *Management Science* 66 (2): 553–563.

Gennaioli, Nicola, and Guido Tabellini. 2018. *Identity, Beliefs, and Political Conflict*. Technical report. Working Paper.

Gentzkow, Matthew. 2016. *Polarization in 2016*. Technical report. Toulouse Network for Information Technology Working Paper.

Le Yaouanq, Yves. 2021. *Motivated cognition in a model of voting*. Technical report. Working Paper.

McCright, Aaron M., and Riley E. Dunlap. 2011. "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010." *Sociological Quarterly* 52 (2): 155–194.

Rabin, Matthews. 1995. *Moral Preferences, Moral Constraints, and Self-Serving Biases*. Technical report. Berkeley Department of Economics Working Paper No. 95-241.

Table 1: 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 labour.	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

Table 2: Description of Charities.

	Debate	**Charity**	**Text to introduce charity in experiment**
1	Migration	American Immigration Council	The American Immigration Council envisions an America that values fairness and justice for immigrants and believes that immigrants are part of the national fabric, bringing energy and skills that benefit all Americans. To advance change they engage in litigation, research, legislative and administrative advocacy, and communications.
2	Animal Welfare	World Animal Protection	World Animal Protection works towards a world where animals live free from suffering. They seek to improve the living conditions of animals farmed for food, to protect and save wild animals, animals affected by disasters, and working animals.
3	Gender Equality	Equality Now	Equality Now believes in creating a just world where women and girls have the same rights as men and boys. They use a unique combination of legal advocacy, regional partnership-building and community mobilization to encourage governments to adopt, improve and enforce laws that protect and promote the rights of women and girls.
4	Abortion	Planned Parenthood	The mission of Planned Parenthood is to provide comprehensive reproductive and complementary health care services in settings which preserve and protect the essential privacy and rights of each individual and to advocate public policies which ensure access to such services. They provide information and support to women considering to end a pregnancy in a clinic or using an abortion pill.
5	Prostitution	A21	The mission of A21 is to end human trafficking and slavery. They work closely with law enforcement on the ground to support police operations, identify victims through their hotlines, assist in the prosecution of traffickers, and represent survivors in court proceedings.
6	Same-sex marriage	OutRight	OutRight envisions a world where LGBTIQ (lesbian, gay, bisexual, transgender/transsexual, intersexual and queer) people everywhere enjoy full human rights and fundamental freedoms. They seek to fill research gaps, provide trainings to community members and allies to develop their expertise, and convene key stakeholders to exchange information on best practises related to ending violence based on sexual orientation.