Eliciting Thresholds for Interdependent Behavior*

Moritz Janas‡ Nikos Nikiforakis§ Simon Siegenthaler¶

January 26, 2026

Abstract

Individuals' willingness to act often depends on how many others do, but the structure of such interdependence is hard to disentangle with observational data. We introduce an incentivized method to measure interdependence, grounded in threshold models. We apply it to a stratified U.S. sample of 5,000+ Asian, Black, Hispanic, and White adults to study support for affirmative action. We document substantial heterogeneity in thresholds consistent with preregistered hypotheses from a model. Following changes in federal support for affirmative action, thresholds shift even as perceived benefits and beliefs remain unchanged, indicating that thresholds provide insights not captured by standard behavioral measures.

Keywords: threshold models, conformity, social influence, affirmative action

JEL Code: C83, C90, D63, D70

*For helpful discussions, we thank Alexander Cappelen, Charles Efferson, Ernst Fehr, Urs Fischbacher, Simon Gächter, Sanjeev Goyal, Mark Granovetter, John List, George Lowenstein, Marco Piovesan, Andrea Robbett, Danila Serra, Bertil Tungodden, John Wooders, Leeat Yariv, and Yves Zenou. We thank seminar participants at the University of Arizona, University of Bologna, University of Chicago, University of Cologne, Florida State University, Middlebury College, New York University, Norwegian Business School, University of Texas at Dallas, University of Verona, Yale University, and the University of Zurich. We also thank audiences at Asia Meetings of the Econometric Society (2026), Zurich Workshop in Economics & Psychology (2025), the Social Norms Workshop in Ascona, the Barcelona Summer Forum, the Dynamics of Social Change Workshop at NYU Abu Dhabi, the Norms and Behavioral Change Conference at the University of Pennsylvania, the French Experimental Economic Association meeting, the Maastricht Behavioral Economic Policy Symposium, the European Economic Association meeting (all in 2024), and the Economic Science Association World and Africa meetings (2023). The project was pre-registered at AEARCTR-0010895. IRB approval has been obtained by the NYUAD (HRPP-2022-74) and UT Dallas (IRB-22-582) Institutional Review Boards. NN gratefully acknowledges financial support from Tamkeen under NYU Abu Dhabi Research Institute Award CG005. NN and SS gratefully acknowledge financial support from the National Science Foundation (grant #2242443).

‡Department of Economics, University of Gothenburg (moritz.janas@gu.se)

§Center for Behavioral Institutional Design, NYU Abu Dhabi; Division of Social Science, NYU Abu Dhabi; Faculty of Arts & Science, New York University (nikos.nikiforakis@nyu.edu)

¶Jindal School of Management, University of Texas at Dallas; National Bureau of Economic Research (simon.siegenthaler@utdallas.edu)

1 Introduction

Our willingness to take an action often depends on how many others have already taken it. From deciding whether to adhere to a social norm, to adopting a new technology, investing in a specific stock, participating in a protest, and purchasing a good or service, individuals often condition their own choices on how many others have already made them. Economists have long recognized that such interdependence can arise from the presence of network externalities (Katz and Shapiro, 1985, 1986), social preferences (Rabin, 1993; Fehr and Schmidt, 1999), reputational concerns (Akerlof, 1980; Bernheim, 1994), informational asymmetries (Banerjee, 1992; Bikhchandani et al., 1992), social norms (Akerlof and Kranton, 2000; Bénabou and Tirole, 2006; Bicchieri, 2006), or attitudes toward conformity (Goeree and Yariv, 2015; Baumann and Olszewski, 2021). Irrespective of its origins, interdependence can generate dynamics in which initial changes—such as policy interventions (Carrell et al., 2013; Bursztyn et al., 2020; Banerjee et al., 2024)—are amplified or attenuated, making it difficult to predict individual behavior or aggregate outcomes (Boucher et al., 2024). Understanding interdependent behavior, therefore, is of obvious importance.

While a vast theoretical literature across the social sciences has examined the dynamics of interdependent behavior, empirical research remains limited due to identification problems (Manski, 2000). The reason is that when individual behavior depends on that of others, it is difficult to infer from observational data whether an individual influences her peers, is influenced by them, or whether both respond to common external factors. This "reflection problem" (Manski, 1993) makes it difficult to recover the structure of interdependence from observational data without strong assumptions, many of which are untestable (Durlauf and Ioannides, 2010). Predicting group outcomes, therefore, is challenging without access to richer data sources that provide insight into individual preferences and expectations (Manski, 2000). In this paper, we take a first step by introducing a method that directly measures behavioral interdependence and bypasses the core identification problem. As interdependence is measured at the individual level, the method allows for the study of cross-sectional heterogeneity and opens new avenues for empirical research on social effects.

Our method builds on threshold models, which posit that individuals act when the number of others taking the same action crosses a personal threshold. These models—sometimes referred to as models of social influence (Young, 2009)—have

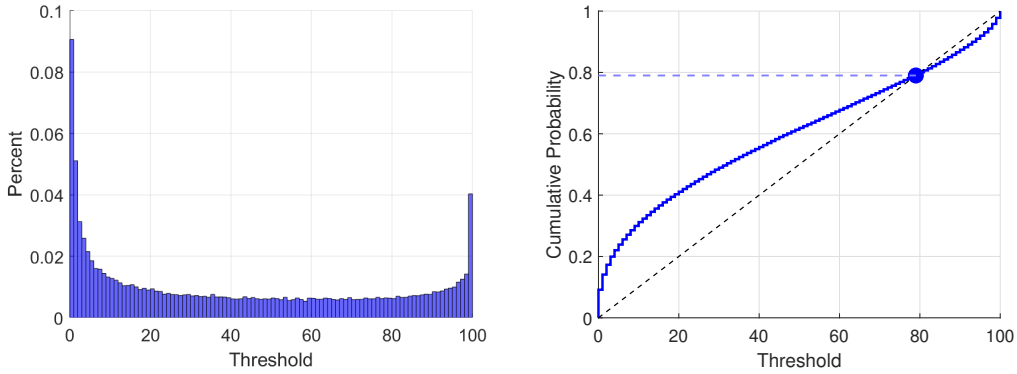
been widely used to model how strategic complementarities in behavior can generate nonlinear dynamics in aggregate outcomes. First introduced by Mark Granovetter (1978) and Thomas Schelling (1978), threshold models have been used extensively by economists, philosophers, political scientists, and sociologists to study phenomena ranging from consumption behavior (Granovetter and Soong, 1986), public good provision (Oliver et al., 1985; Macy, 1991), policy adoption (Roland and Verdier, 1994; Simmons and Elkins, 2004), diffusion of innovation (Jackson and Yariv, 2005; Galeotti and Goyal, 2009; Young, 2009; Centola, 2015), and network dynamics (Jackson and Yariv, 2007; Jackson, 2008; Galeotti et al., 2010; Goyal, 2023), to riots (Granovetter, 1978), racial segregation (Schelling, 1978; Card et al., 2008; Zhang, 2011), revolutions (Kuran, 1995), norm change (Bicchieri, 2016; Efferson et al., 2015, 2020; Andreoni et al., 2021), political polarization (Ehret et al., 2022), and climate action (Scheffer, 2020; Constantino et al., 2022; Berger et al., 2023).[1]

The widespread and enduring use of threshold models reflects their ability to provide a tractable framework for analyzing complex dynamics. In these models, each individual is characterized by a threshold, t_i^a, indicating the share of others who must take action a before individual i does the same. Once the share of others choosing a exceeds t_i^a, individual i also chooses a. A threshold of 0% represents unconditional supporters, who choose a even if no one else does, whereas a threshold of 100% represents unconditional opponents, who never choose a even if everyone else does. Between these extremes are individuals whose decisions depend on the share of others selecting a. Given a probability distribution of thresholds in a population, threshold models use best-response dynamics to generate sharp, testable predictions about individual behavior and aggregate outcomes.

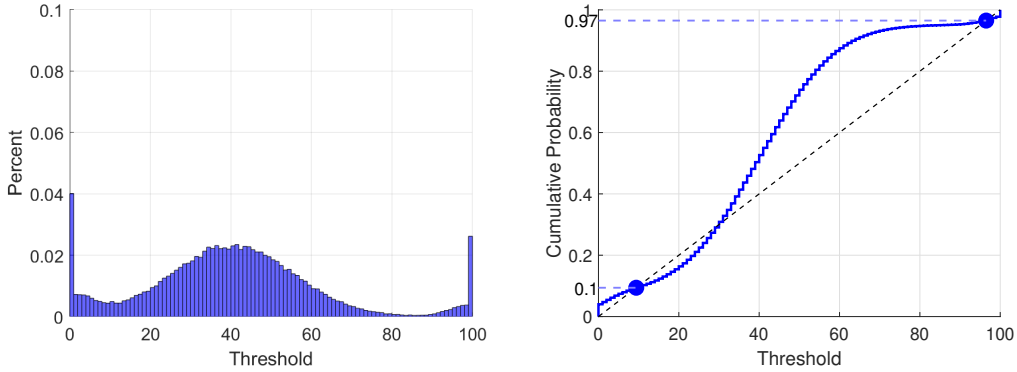
An example helps illustrate how threshold models work and sets the stage for our method. Suppose you are at a faculty meeting where a policy to adopt affirmative action (AA) in hiring is under discussion. Participants are asked to show support for the policy by raising their hands. You evaluate the policy's merits and the reputational consequences from visibly supporting (or opposing) it. Others in the room make

[1]Beyond threshold models, scholars have used models of social learning and social contagion to study interdependent decision-making. While similar in some respects to threshold models, they differ in important ways; see Young (2009) for a detailed discussion. Threshold models are also related to global games—games of incomplete information with strategic complementarities—used to study phenomena such as bank runs, currency attacks, and financial crises (Carlsson and van Damme, 1993).

Figure 1: Examples of Threshold Distributions and Equilibrium Predictions



A. Unique Equilibrium



B. Two Stable Equilibria

Notes: Threshold distributions and resulting equilibria for two stylized examples. The left panels show two threshold distributions with the same mean; the right panels show the corresponding CDFs. The equilibrium share choosing action a is given by the intersection of the CDF with the 45-degree line from above.

similar calculations. How many faculty members will raise their hands depends on the distribution of thresholds. Assume that the distribution of thresholds is the one shown in Figure 1A: 9% of participants have a threshold of 0, meaning they will raise their hand even if no one else does, while 4% have a threshold of 100, meaning they will never raise their hand even if everyone else does. The remaining 87% exhibit interdependent behavior, raising their hands only if enough others do. The stable equilibrium occurs at the point where the cumulative distribution function (CDF) intersects the 45-degree line from above—as long as the CDF lies above the line, the best-response dynamic implies that more individuals will raise their hands, encour-

aging others to follow. In this example, therefore, 78% of participants are predicted to raise their hands in support of the proposed policy.

Apart from illustrating how threshold models reduce complex interdependence to a simple decision rule (individuals will act if enough others do so) the example also serves to highlight the essential role of empirical evidence in understanding behavioral interdependence. Specifically, predicting aggregate outcomes depends critically on knowing the distribution of thresholds rather than only their average. To see this, suppose instead that the distribution corresponds to Figure 1B. Although the distributions in Figures 1A and 1B have the same mean, the clustering of thresholds around 40 now yields two stable equilibria: one with roughly 10% and another with 97% of individuals supporting the AA policy. Without information about the underlying heterogeneity in thresholds, the realized equilibrium and thus the aggregate outcome cannot be determined. This observation raises a fundamental question: what determines individual thresholds?

Despite the widespread use of threshold models in theoretical research, direct empirical evidence on individual thresholds remains scarce. In theoretical work, individuals are assumed to set their thresholds rationally, at the point where the perceived benefit of choosing action a exceeds the perceived cost, which depends on how many others choose a. Existing laboratory studies demonstrate that threshold models can accurately predict aggregate outcomes in controlled settings (Centola et al., 2018; Andreoni et al., 2021; Ehret et al., 2022). However, these studies do not elicit individual thresholds as we do here; instead, they either assume a threshold distribution or induce threshold variation without observing individuals' thresholds. Hence, these studies offer limited insight into the determinants of thresholds, how thresholds shift in response to changing incentives, or what the threshold distribution—commonly assumed to be normal (Bicchieri, 2016; Andreoni et al., 2021)—actually is.[2] These are, at their core, empirical questions. Answering them requires eliciting thresholds at the individual level.

To address these questions, we use our method to elicit individual thresholds for

[2]Several related literatures provide indirect evidence that people behave as if they have thresholds—for example, in global games where actions depend on signals about the state of the world (Heinemann et al., 2004, 2009; Szkup and Trevino, 2020), in dynamic public goods settings where contributions depend on progress toward a collective goal (e.g., List and Lucking-Reiley, 2002; Duffy et al., 2007; Tavoni et al., 2011), or through conditional commitments that trigger cooperation once participation thresholds are met (e.g., Schmidt and Ockenfels, 2021; Oechssler et al., 2022).

supporting affirmative action (AA) in a large online sample of the U.S. population. We focus on AA for two main reasons. Substantively, AA has significant socioeconomic implications (Holzer and Neumark, 2000) and remains a subject of sustained public debate (Bleemer, 2022; Chinoy et al., 2026). Methodologically, AA provides a setting in which individuals' perceived benefits—and hence their incentives to support the policy—are naturally tied to their racial/ethnic/gender (REG) group, generating exogenous variation for studying the determinants of thresholds. We therefore stratify the sample by race/ethnicity (Asian, Black, Hispanic, and White) and gender. Importantly, this structure makes AA an informative test of threshold models: because individuals are likely to hold clear, group-linked preferences over the policy, observing interior thresholds reveals meaningful interdependence rather than indifference or noise.

To guide our empirical analysis, we develop a simple model in which individuals choose their threshold by weighing the benefits of supporting AA against the costs. The model yields testable behavioral predictions, which we preregistered prior to data collection. To test these predictions, the experimental design varies the conditions under which we elicit thresholds, in three dimensions: (*i*) the direction of advocacy relative to the status quo—whether participants are asked to support or oppose AA; (*ii*) the individual's reference group—either the general U.S. population or individuals from the same REG group; (*iii*) the visibility of the individual's decision—whether it is public or private. We also collect information on individual attitudes toward AA, conformity, and anticipated social sanctions.

The main data collection for our study was completed in 2023, at a time when federal support for affirmative action remained in place. Shortly thereafter, a sequence of legal and political developments substantially altered the formal institutional environment surrounding affirmative action. While formal rules can change abruptly, informal institutions—such as norms, conventions, and shared expectations (North, 1990)—often adjust more slowly (Andreoni et al., 2021; Kamm et al., 2021). This creates a setting in which observed choices may lag underlying willingness to act. To examine how individuals' readiness to support affirmative action responds to such institutional shifts, we collected a second wave of data in mid-2025 using a new nationally representative sample of White women, a group with historically mixed views on the policy.

The empirical analysis reveals three central findings. First, interdependent behav-

ior is widespread: between 63.33% and 87.92% of individuals choose interior thresholds across REG groups and experimental conditions. Even in a highly polarizing domain such as affirmative action, a majority of respondents condition their support on others' behavior, underscoring the importance of studying behavioral interdependence directly.

Second, thresholds vary systematically and predictably with incentives and the social environment. Consistent with the model, individuals have lower thresholds for supporting affirmative action when the perceived benefits of doing so are higher, and this relationship reverses when experimental variation reverses the direction of advocacy. Thresholds are also shaped by social context: public visibility, anticipated social sanctions, and the composition of the reference group all affect individuals' willingness to support change in line with our hypotheses. Together, these forces generate substantial and structured heterogeneity in thresholds across race, gender, and political affiliation, highlighting threshold elicitation as a powerful lens for understanding aggregate outcomes.

Third, threshold data reveal dimensions of behavioral change that are difficult to detect using standard measures of preferences or beliefs. Comparing threshold distributions before and after the change in federal support for affirmative action reveals substantial shifts in individuals' readiness to support the policy, even as perceptions of its benefits and beliefs about others' behavior remain remarkably stable. This distinction highlights the value of threshold elicitation for capturing behavioral responsiveness to the social environment that is not visible in choice, preference, or belief data alone.

The next section introduces a theoretical framework that guides our elicitation method and empirical analysis. Section 3 describes our experimental design. Section 4 presents formal tests of behavioral hypotheses. Section 5 compares threshold distributions before and after shifts in federal support for affirmative action. Section 6 discusses how thresholds allow researchers to anticipate when interdependence amplifies or attenuates intervention effects across groups and settings. Section 7 concludes and outlines directions for future research.

2 Theoretical framework

Our method is grounded on the classic threshold model introduced by Schelling (1978) and Granovetter (1978). Consider a society in which each individual is characterized by a threshold t_i^a. The threshold indicates the share of others that must take action a before individual i does the same, with $t_i^a \in [0,1]$ and $a_i \in \{0,1\}$, where 1 indicates support of a. Let r_τ^a denote the adoption rate (the share of supporters) at iteration τ. Individual i chooses $a_i = 1$ if and only if $r_\tau^a \geq t_i^a$. In other words, i will choose a at iteration $\tau + 1$ if and only if the observed adoption rate at iteration τ meets i's threshold. Thresholds are heterogeneous and distributed according to the cumulative distribution function F. The dynamic is governed by $r_{\tau+1}^a = F(r_\tau^a)$ with initial condition, or status quo, $r_0^a = 0$. Starting from the status quo, the adoption rate increases iteratively until it reaches a stable equilibrium $r^{a,*}$ that satisfies $r^{a,*} = F(r^{a,*})$.

To elicit thresholds, we use a two-step method. First, each participant is assigned to a group of n individuals. We commit to donating $\$x$ for each group member to a charity that *opposes* a (i.e., $a_i = 0$ for all i). Thus, absent any change, the status quo entails a total donation of $\$x \times n$ to an organization aligned with $a = 0$.[3] Second, each group member i is asked whether they would like to change their donation to an organization *supporting* a by specifying the share of other group members who must support a before i does so, denoted by $t_i^a \in [0,1]$. To incentivize truthful threshold choices, it is common knowledge that each group member i will switch her donation to $a_i = 1$ if and only if her stated threshold is less than or equal to $r^{a,*}$, the eventual adoption rate in the group.

To obtain testable hypotheses, we consider a simple model that explains why individuals may have different thresholds. The model draws on a theoretical literature that emphasizes the importance of individual benefits and beliefs (Granovetter, 1978; Bicchieri, 2006), social alignment motives such as conformity and coordination (Bernheim, 1994; Durlauf and Ioannides, 2010), and social pressure to conform to others' behavior (Akerlof, 1980; Kuran, 1995). Building on these foundations, we

[3]This design avoids income effects by not giving participants a personal endowment to retain or donate. To evaluate the robustness of our findings, we also elicit thresholds in conditions in which participants are given an endowment and choose whether to keep it or donate to the organization, and another in which they choose between supporting action a and remaining neutral. These variations are discussed in Online Appendix B.

model individual utility as:

$$U_i(a_i, r(a_i), \Delta b_i, \beta, \gamma),$$

where:

- $a_i \in \{0, 1\}$ is individual i's *action*, with $a_i = 1$ indicating support of a.
- $r(a_i) \in [0, 1]$ is the *adoption rate*, defined as the share of others who choose $a = 1$.
- $\Delta b_i \in \mathbb{R}$ represents the *perceived benefit* of choosing $a_i = 1$ rather than $a_i = 0$.
- $\beta \geq 0$ captures *social alignment*, which imposes a cost of choosing differently from others.
- $\gamma \geq 0$ captures *social pressure*, an asymmetric cost associated with deviating from the status quo and taking the action that challenges existing practices.[4]

The individual's net benefit from choosing a is given by

$$\Delta U_i \equiv U_i(1, r(1), \cdot) - U_i(0, r(0), \cdot).$$

The optimal threshold $t_i^{a,*}$ is the value of $r(0)$ at which the individual is indifferent. So we evaluate ΔU_i at $r(0) = t_i^{a,*}$:

$$\Delta U_i\big(t_i^{a,*}, \Delta b_i, \beta, \gamma, \Delta r_i(t_i^{a,*})\big) = 0,$$

where

$$\Delta r_i \equiv r(1) - r(0) \geq 0$$

denotes the belief about the fraction of others who would support a only if individual i chooses $a_i = 1$. Because individuals have a larger marginal influence on others when their threshold is lower, Δr_i is endogenous and depends on $t_i^{a,*}$. Differentiating the indifference condition implicitly yields

$$\frac{dt_i^{a,*}}{dx} = -\frac{\frac{\partial \Delta U_i}{\partial x}}{\frac{\partial \Delta U_i}{\partial r(0)} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \cdot \Delta r_i'(t_i^{a,*})}, \qquad x \in \{\Delta b_i, \beta, \gamma\}.$$

[4]The reputational cost of publicly supporting (or opposing) affirmative action may be asymmetric when inaction is interpreted as a need for more information or inattention rather than as opposition (or support), as is often the case.

The numerator determines the direction of the threshold shift: a positive (negative) numerator implies that an increase in x makes supporting action a more (less) attractive. The denominator governs the magnitude of this shift. Its first term captures how sensitive the utility difference is to changes in the adoption rate. When this sensitivity is high (low), only a small (large) adjustment in the threshold is needed to restore indifference. The second term reflects forward-looking beliefs about how one's own action affects others' adoption. Because $\Delta r_i'(t_i^{a,*}) < 0$, this belief-based component lowers the denominator and therefore amplifies all comparative statics.[5]

We now turn to the comparative statics implied by the model. These results hold for the general specification, while Figure 2 illustrates them using a linear utility function. An increase in perceived benefits Δb_i lowers the optimal threshold, $\frac{dt_i^{a,*}}{d\Delta b_i} < 0$. The reason is straightforward: higher perceived benefits shift ΔU_i upward, making action attractive even when fewer others adopt. The first panel of Figure 2 illustrates this using the linear specification: the two lines show ΔU_i for low and high values of Δb_i. The point where each line crosses the x-axis indicates the threshold at which acting becomes worthwhile. When Δb_i is higher, this crossing occurs at a smaller value of $r(0)$, reflecting a lower threshold.

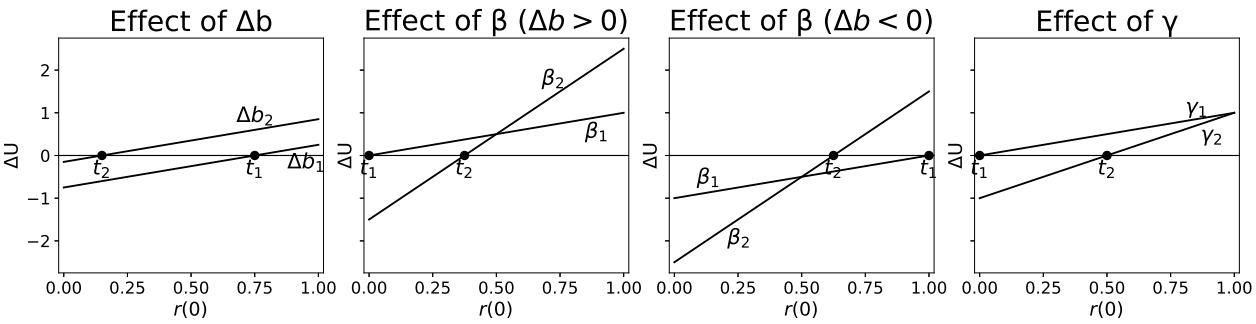
Changes in the social-alignment parameter β pull thresholds toward the midpoint. When individuals are more favorable to a ($t_i^{a,*} < 50\%$), higher β raises their threshold, $\frac{dt_i^{a,*}}{d\beta} > 0$; when they are less favorable ($t_i^{a,*} > 50\%$), higher β lowers it, $\frac{dt_i^{a,*}}{d\beta} < 0$. The intuition is that stronger social alignment penalizes holding a minority position. The second and third panels in Figure 2 illustrate this pattern. Regardless of whether individuals initially favor or oppose change, increasing β rotates the ΔU_i lines toward $r(0) = 50\%$, shifting the threshold toward the midpoint.

The social-pressure parameter γ raises thresholds, $\frac{dt_i^{a,*}}{d\gamma} > 0$. Stronger pressure to maintain the status quo makes supporting a less attractive. Figure 2 illustrates that increasing γ rotates the ΔU_i lines toward $r(0) = 100\%$, shifting the threshold to the right. The effect is particularly pronounced for individuals with low thresholds, as those willing to act early are most discouraged by stronger social pressure.

Forward-looking beliefs amplify these comparative statics: when one's action can

[5]The model exhibits strategic complementarities: as the adoption rate rises, supporting a becomes more attractive, implying $\frac{\partial \Delta U_i}{\partial r(0)} > 0$ and $\frac{\partial \Delta U_i}{\partial \Delta r_i} > 0$. An individual's marginal influence on others necessarily declines as the threshold increases, so $\Delta r_i'(t_i^{a,*}) < 0$. This makes the belief-based term negative, but the denominator remains positive. The sign of each comparative static is therefore fully determined by the numerator. The full derivation is provided in Online Appendix A.

Figure 2: Effects of Model Parameters on Optimal Threshold



Notes: Optimal thresholds for the utility function $U_i(a_i) = b_i(a_i) - \beta_i r(a_i) - \gamma_i r(a_i)\mathbb{1}_{a_i=1}$. The linear specification is used for illustration only; the comparative statics do not rely on linearity. Y-axis (ΔU): the utility difference from choosing $a_i = 1$ rather than $a_i = 0$; the optimal threshold is defined by $\Delta U = 0$. X-axis ($r(0)$): the fraction of others choosing $a = 1$ when $a_i = 0$. Each panel varies one determinant of the threshold and shows the resulting optimal thresholds, t_1 and t_2, corresponding to a lower and higher value of the parameter.

induce others to follow, changes in incentives translate into larger shifts in the optimal threshold.

3 Experimental design

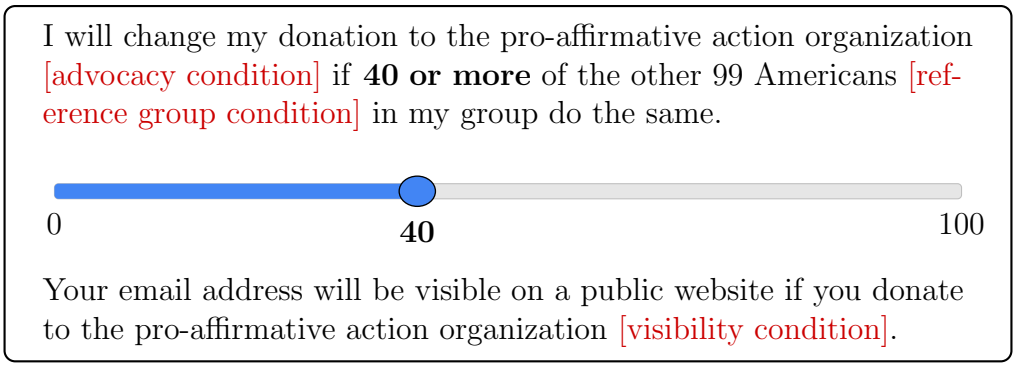
3.1 Decision context

Our method can be used to study situations of interdependence when an individual's action space is binary.[6] We apply it to study support for affirmative action (AA) in the United States. Participants are placed in groups of $n = 100$ people. For each group member, we commit to donating $1 to one of two organizations: the American Association for Access, Equity, and Diversity (pro-AA) or the American Civil Rights Institute (anti-AA). For each group, one organization is randomly assigned as the status quo ($a_i = 0$). Choosing $a_i = 1$ therefore corresponds to advocating *change*, that is, redirecting the donation to the other organization. Participants then indicate the *number* of others who must support a before they would also do so. Formally,

[6]Our method is related to strategy methods in experimental economics (e.g., Brandts and Charness, 2011; Fischbacher et al., 2001; Fischbacher and Gächter, 2010), which elicit conditional choices under different information sets. The crucial difference is that, instead of holding others' behavior fixed, each choice in our design is embedded in a mutually dependent system in which thresholds jointly determine the group outcome. An individual's chosen a both affects and is affected by others through threshold interdependence.

each participant reports a threshold $t_i^a \in \{0, 1, \ldots, 100\}$. Reporting 100 means the participant always sticks with the status quo ($a_i = 0$), while reporting 0 means they switch regardless of what others do ($a_i = 1$). Interior values allow individuals to condition their action ($a_i = 0$ or 1) on the emerging level of support within the group. Participants are informed how the distribution of thresholds in a group determines their own action, the collective dynamics, and the resulting equilibrium adoption rate $r^{a,*}$.

Figure 3: Threshold Elicitation Interface



The threshold elicitation interface is shown in Figure 3. Participants view a sentence describing their choice in a specific group context, incorporating the experimental conditions: advocacy (i.e., the organization they can support), the relevant reference group, and the visibility of their choice (explained in the next section). Using a slider, they then indicate the minimum number of other group members (t_i^a) who must advocate for action a before they would also switch. Separate explanations clarified the meanings of interior, zero, and 100 thresholds and emphasized the distinction between conditioning on others' actions and not doing so.

3.2 Experimental treatments and behavioral hypotheses

The goal of our experiment is to measure the distribution of individual thresholds and to identify how thresholds shift with changes in perceived benefits, social alignment, and social pressure. Table 1 summarizes how our theoretical framework informs our empirical design. Column 1 lists the theoretical components, Column 2 the exogenous variation used to identify their causal effects, Column 3 the corresponding individual-level measures, and Column 4 the model's predicted effects on thresholds.

Table 1: Experimental Design

Threshold Component	**Exogenous Variation**	**Individual Measures**	**Predicted Effect**
Perceived Benefits (Δb_i)	Advocacy & REG Group	Social Benefits Index	$\frac{\partial t_i^{a,*}}{\partial \Delta b_i} < 0$
Social Alignment (β_i)	Reference Group	Conformity Index	$\frac{\partial t_i^{a,*}}{\partial \beta_i} \begin{cases} > 0 & \text{if } t_i^{a,*} < 0.5 \\ < 0 & \text{if } t_i^{a,*} > 0.5 \end{cases}$
Social Pressure (γ_i)	Visibility	Beliefs About Social Sanctions	$\frac{\partial t_i^{a,*}}{\partial \gamma_i} > 0$

Notes: The table summarizes how theoretical components of the model map into empirical tests. REG abbreviates racial/ethnic/gender. Online Appendix B provides the construction and descriptive statistics for the social benefits index, conformity index, and beliefs about social sanctions.

All hypotheses were preregistered (AEARCTR-0010895; see Online Appendix C). To identify the causal effect of perceived benefits of change (Δb_i), we vary the advocacy associated with action a. In half of the groups, action a corresponds to advocating for affirmative action, so we elicit thresholds for changing toward AA. In the other half, action a corresponds to advocating against affirmative action, so we elicit thresholds for changing away from AA. Throughout, t^{AA} denotes thresholds for advocating *for* affirmative action, while t^{NoAA} denotes thresholds for advocating *against* affirmative action. We also exploit natural variation across racial, ethnic, and gender (REG) groups, who differ systematically in their expected personal and social benefits from AA. This provides additional cross-sectional variation in Δb_i, beyond that induced by the randomized direction of advocacy. We further construct a *social benefits index* for each individual to proxy perceived benefits using their agreement on a five-point Likert scale to the following statements: (*i*) AA programs help decrease institutional injustice; (*ii*) AA does more harm than good to minority groups; (*iii*) AA is itself a form of discrimination; (*iv*) AA enhances organizational performance in the long run.

Hypothesis 1 (Perceived benefits, Δb_i)**:** *Greater perceived benefits from AA should reduce t^{AA} and increase t^{NoAA}. Specifically: (i) individuals with a higher social benefits index will have lower t^{AA} and higher t^{NoAA}; (ii) individuals from underrepresented*

REG groups will have lower t^{AA} and higher t^{NoAA} than White men.

To identify the effect of social alignment (β_i), we vary the reference group in which thresholds are elicited. Each participant reports two thresholds: a *population threshold*, based on a group of 100 individuals representing the general U.S. population, and a *REG threshold*, based on a group homogeneous in their own race/ethnicity and gender. One of these two thresholds is randomly selected for payment.[7] We also construct an individual *conformity index* as a proxy for β_i, adapting a widely used measure of conformity (e.g., Hong and Page, 1989; Andreoni et al., 2021). Participants indicate their agreement on a five-point scale with the following statements: (*i*) I resist the attempts of others to influence me; (*ii*) I become frustrated when I am unable to make free and independent decisions; (*iii*) I become angry when my freedom of choice is restricted; (*iv*) It makes me angry when another person is held up as a model for me to follow; (*v*) When someone forces me to do something, I feel like doing the opposite.

Hypothesis 2 (Social alignment, β_i)**:** *Greater preference for social alignment will make interior thresholds more common. Specifically, thresholds will more often be interior (i) when the reference group is narrow (own REG group) rather than broad (U.S. population), and (ii) the higher an individual's conformity index is.*

To identify the effect of social pressure (γ_i), we vary whether participants' choices are revealed publicly. Twenty percent of participants are assigned to a *Private* condition, in which donations remain confidential. The remaining 80 percent are assigned to a *Public* condition, in which the email addresses and donation choices of those who support change ($a_i = 1$) are posted on a publicly accessible website. Participants are informed that the study results and the website may be shared on social media.[8] The public condition is expected to raise thresholds through social pressure, though the lack of in-person interaction likely leads us to underestimate the

[7]Social alignment is expected to be stronger when others are more similar to oneself, consistent with evidence that people place greater weight on conforming to and learning from similar others (e.g., Fatas et al., 2018; Bicchieri et al., 2022; Ehret et al., 2022). Social alignment combines two components: intrinsic conformity (valuing norm-following) and social learning (treating others' behavior as informative). Experiments by Goeree and Yariv (2015) separate these motives by letting subjects choose between a statistically informative private signal and an uninformative social signal. Many subjects nevertheless chose the social signal, showing that both intrinsic and informational motives contribute to conformity.

[8]Participants retain full control over whether their email addresses appear online by choosing to keep the status quo donation. The email addresses correspond to those provided to Ipsos during registration to their panel. Hence, they are addresses they use regularly. All personal information

magnitude of such pressure outside the experimental setting. We also construct an individual proxy for perceived social pressure by measuring participants' *beliefs about social sanctions*. Following standard approaches (e.g., Fehr and Fischbacher, 2004; Bicchieri, 2006), participants first report how likely they would be to confront others who publicly support or oppose AA. They then provide an incentivized guess (with accuracy-based rewards) about how many others in their group reported being likely to confront others.

Hypothesis 3 (Social pressure, γ_i)**:** *Thresholds will be lower when choices are private than when they are public. Thresholds will also be lower for individuals who perceive weaker social sanctions for advocating change.*

3.3 Sample and Procedures

The sample consists of 5,099 U.S. residents (Table 2): 4,086 from the main study in 2023 and 1,013 from a follow-up wave in 2025. The main study included roughly equal numbers of Asian, Black, Hispanic, and White men and women. We stratify by race, ethnicity, and gender (REG) because perceived benefits from affirmative action plausibly differ across these groups, and this natural variation provides additional leverage to test the model's predictions. Stratification also ensures sufficient power to detect differences among underrepresented groups, which a purely random U.S. sample would not.

Respondents were recruited by Ipsos from its online panel using quota sampling. Quotas for race, ethnicity, and gender were set to match our target composition and aligned with the 2021 American Community Survey (ACS), ensuring that each REG group is representative of its source population. Potential respondents entering the survey router were screened for eligibility, and invitations continued until quotas were met. To ensure high attentiveness, we embedded two instructional checks (e.g., "Please select 'Disagree' here") and removed any participant who failed either check. The median completion time was 14.27 minutes.

Participants completed four main parts: demographic questions, the threshold-elicitation task, opinion measures on affirmative action, and psycho-sociological measures (full materials in Online Appendix D). The order of the threshold-elicitation

was removed six months after data collection. A redacted screenshot is provided in Online Appendix D. The website is available at: `www.HowPeopleThinkAbout.org/AffirmativeAction`.

Table 2: Sample

		Education		Age Group				US Region			
REG Group	Total	No College	College	21-24	25-34	35-44	45-65	Mid-west	North-east	South	West
2023 Sample											
Asian, F	488	106	382	33	121	133	201	61	108	125	194
Asian, M	507	141	366	34	122	139	212	63	106	121	216
Black, F	502	295	207	26	130	115	231	82	75	304	41
Black, M	503	336	167	39	126	119	219	85	76	291	51
Hispanic, F	484	278	206	51	143	123	167	46	71	187	180
Hispanic, M	499	323	176	42	137	136	184	46	71	189	193
White, F	502	251	251	26	106	110	260	132	95	180	95
White, M	501	244	257	27	110	113	251	130	96	175	100
Unassigned	100	62	38	8	19	26	47	19	23	33	25
2025 Sample											
White, F	1,013	470	543	98	180	223	512	256	202	365	190
Total	5,099	2,506	2,593	384	1,194	1,237	2,284	920	923	1,970	1,285

Notes: REG refers to racial/ethnic/gender group. Education, region, and age quotas are derived from the 2021 American Community Survey (ACS). The Main Data sample was collected in 2023. The 2025 sample was collected two years later, after the removal of federal support for affirmative action, to examine how thresholds respond to the changed institutional environment and to address additional design questions that emerged after the initial wave. Participants who declined to state their race/ethnicity or identified as non-binary are listed as *Unassigned*. No data were collected for American Indians, Alaska Natives, Native Hawaiians, or other Pacific Islanders (1.3% of the U.S. population).

task and the psycho-sociological measures was randomized. Participants were informed how their threshold choices determined donation amounts and were shown descriptions of the two organizations. The median participant spent 117 seconds on their first threshold choice (25^{th} percentile: 64 s; 75^{th} percentile: 200 s). Participants then stated their beliefs about (*i*) the share of others who would switch and (*ii*) the distribution of others' threshold choices.

Monetary incentives included the standard Ipsos fee ($1 per participant), the donations linked to threshold choices ($1 per participant), and rewards for belief elicitation ($1.20 per participant on average). In the Online Appendix, we show that threshold distributions and their main determinants replicate when the donation-linked incentive is removed.

4 Empirical analysis

The empirical analysis focuses on formal tests of our preregistered hypotheses using the 2023 sample and regression analysis. Before turning to these tests, we provide an overview of threshold distributions across racial, ethnic, and gender (REG) groups and political affiliation. The aim is to highlight differences across groups that are both substantively large and theoretically informative.

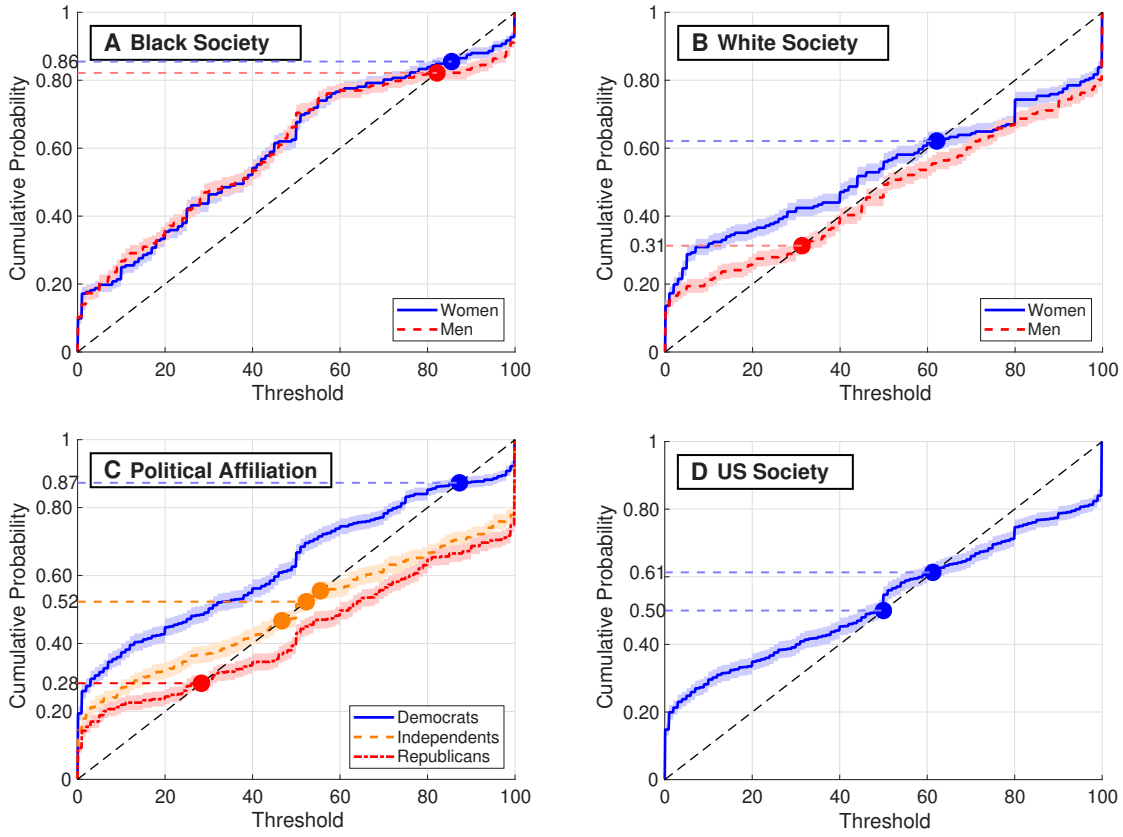
4.1 Overview of threshold distributions

We begin by contrasting Black and White, men and women, as affirmative action is explicitly designed to address racial and gender inequality, and these groups are therefore expected to differ most sharply in the perceived benefits of such policies. Figures 4A–C illustrate the substantial heterogeneity in threshold distributions for supporting affirmative action (AA) across groups. The cumulative threshold distribution for Black women is shifted far to the left relative to that of White men, indicating substantially lower thresholds for collective support. This difference is statistically significant according to a Kolmogorov–Smirnov (K–S) test ($p < 0.001$). The implied societal equilibria—given by the intersections of the cumulative distributions with the 45-degree line—correspond to 86% support for affirmative action among Black women and 31

Gender differences within racial groups are more nuanced. Among White respondents, men exhibit significantly higher thresholds than women (K–S, $p = 0.034$), whereas thresholds do not differ significantly between Black men and Black women (K–S, $p = 0.606$). These patterns are consistent with Hypothesis 1, which posits that threshold differences reflect variation in the perceived benefits of affirmative action across groups rather than uniform conformity pressures.

We next turn to political affiliation, focusing on Democrats and Republicans, for whom affirmative action has long been a salient and polarizing policy issue, and where sharp differences in perceived benefits are therefore expected to translate into distinct threshold distributions. As can be seen in Figures 4D there exists stark contrasts across political affiliation. Democrats display much lower thresholds for AA than Independents (K–S, $p < 0.001$), and Independents in turn have lower thresholds than Republicans (K–S, $p = 0.005$), underscoring the role of ideology in shaping conditional support for collective action.

Figure 4: Selected Threshold Distributions



Notes: Distribution of thresholds for AA (t^{AA}) for Black men and women (top left), White men and women (top right), a representative sample split by political preference (bottom left), and a U.S. representative sample (bottom right). Thresholds for Black and White respondents are based on narrow reference groups. Shaded areas show 90% confidence intervals of the CDFs constructed from 10,000 random samples of size $n = 1{,}000$. Markers indicate the social equilibria. Distributions of thresholds for the other REG groups and conditions are shown in the Online Appendix.

Given the substantial heterogeneity across groups, it is also interesting to examine the threshold distribution in the U.S. as a whole. Using sampling weights, we construct a U.S.-representative distribution of thresholds. As can be seen in Figure 4D, 16% of respondents support affirmative action unconditionally ($t_i^{AA} = 0$), while 15% oppose it even if everyone else supports it ($t_i^{AA} = 100$). The remaining 69% exhibit interior thresholds, indicating that their support depends on others' behavior– thresholds close to 0 or 100 reflect limited sensitivity to others' choices, whereas interior thresholds capture meaningful interdependence. The bimodal distribution-with a substantial interior mass-is at odds with the common assumption of normally distributed thresholds (e.g., Granovetter, 1978; Young, 2009; Bicchieri, 2016; Andreoni et al., 2021).

The high prevalence of interior thresholds is not specific to the U.S.-representative distribution. Across experimental conditions and subsamples, we consistently find that a large majority of respondents condition their support for affirmative action on others' behavior. In particular, across all treatments and REG groups, between 63 and 86% of respondents exhibit interior thresholds, indicating meaningful interdependence in support for affirmative action rather than unconditional support or opposition.[9] As discussed at the start of this paper, behavioral interdependence can arise for different reasons. In the context of our study, a plausible channel is the informational asymmetry concerning the benefits of AA.

4.2 Tests of behavioral hypotheses

The patterns above provide descriptive support for our hypotheses about the determinants of thresholds. We now provide formal tests using regression analysis.

Result 1 (**Perceived benefits,** Δb_i)**:** *In line with Hypothesis 1, greater perceived benefits from affirmative action are associated with lower* t^{AA} *and higher* t^{NoAA}*. In addition, members of underrepresented REG groups have lower* t^{AA} *and higher* t^{NoAA} *than White men, even after accounting for perceived benefits.*

Support: As shown in Table 3, column (1), higher perceived benefits of affirmative action are associated with lower thresholds for supporting AA; a shift of the benefits index from -0.5 to 0.5 corresponds to a 38.6 percentage point lower threshold on

[9]The Online Appendix presents the full set of threshold distributions and summary statistics across all REG groups and experimental conditions.

Table 3: Perceived Benefits and Social Pressure Shift Threshold Levels

Thresholds:	(1) All	(2) All	(3) All	(4) t^{AA}	(5) t^{NoAA}	(6) t^{AA}	(7) t^{NoAA}
Perceived Benefits (Δb_i)							
Benefits index	-38.618***					-35.589***	38.577***
	(3.480)					(3.666)	(3.492)
Advocacy: Anti-AA	-1.423						
	(1.116)						
Benefits index	79.122***						
× Advocacy: Anti-AA	(4.810)						
Social Pressure (γ_i)							
Public		4.575***	5.334**			5.981***	3.584**
		(1.266)	(2.472)			(1.657)	(1.795)
Social Sanctions			20.165***			16.389***	7.342**
			(4.710)			(3.151)	(3.210)
Public			-2.771				
× Social Sanctions			(5.314)				
REG groups							
Asian/Female				-5.787*	12.132***	-5.460*	3.648
				(3.088)	(3.054)	(3.011)	(3.097)
Asian/Male				-2.390	14.005***	-2.008	10.283***
				(2.987)	(3.018)	(2.803)	(2.924)
Black/Female				-14.109***	5.808*	-6.788**	0.233
				(2.958)	(3.058)	(2.951)	(3.054)
Black/Male				-13.953***	5.215*	-7.698***	0.281
				(2.917)	(3.009)	(2.850)	(2.935)
Hispanic/Female				-5.566*	10.484***	-6.930**	0.157
				(3.061)	(3.137)	(2.930)	(3.124)
Hispanic/Male				-5.932**	6.130**	-1.814	2.530
				(2.870)	(2.858)	(2.698)	(2.780)
White/Female				-8.155**	7.271**	-7.942***	4.780
				(3.180)	(3.148)	(3.050)	(2.948)
Democrat						-5.695***	1.696
						(1.703)	(1.872)
Republican						2.168	-2.284
						(2.165)	(2.128)
College						2.917**	7.567***
						(1.481)	(1.549)
Age						0.066	-0.145**
						(0.060)	(0.062)
Constant	48.618***	44.016***	36.178***	51.627***	43.134***	38.551***	40.030***
	(0.812)	(1.116)	(2.171)	(2.234)	(2.190)	(4.160)	(4.014)
Observations	7,972	7,972	7,972	4,070	3,902	3,836	3,624
Subjects	3,986	3,986	3,986	2,035	1,951	1,918	1,812

Notes: OLS regressions on thresholds $t \in \{0, 1, ..., 100\}$ with s.e. clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data include two thresholds per individual (U.S. population and REG reference groups). The benefits index (normalized to -0.5 to 0.5) reflects perceived social benefits of AA policies. Social Sanctions are measured using participants' incentive-compatible expectations about whether others would confront them for speaking in favor of affirmative action (normalized between 0 and 1). Columns 4 and 6 report thresholds for supporting AA; columns 5 and 7 report thresholds for opposing AA. White men are the omitted REG group in columns 4–7. Independents and individuals without a college degree are the omitted categories in columns 6 and 7. In the Online Appendix, this table is replicated separately for interior and non-interior thresholds.

average. The interaction with the advocacy condition shows a complete reversal when change means opposition to affirmative action: the same increase in perceived benefits now raises t^{NoAA} by a comparable amount. The split-sample estimates replicate this pattern: in column (6), perceived benefits reduce t^{AA} by 35.6 points, while in column (7), they increase t^{NoAA} by 38.6 points. These mirrored effects across advocacy condition show that perceived benefits causally shift threshold choices in the predicted direction. These correlations between the benefits index and threshold choices are robust to alternative constructions of the benefits index based on any subset of its constituent items (see Online Appendix).

Members of underrepresented REG groups have lower t^{AA} and higher t^{NoAA} than White men (columns 4 and 5). Adding controls attenuates but does not eliminate these differences (columns 6 and 7), indicating that group identity influences thresholds beyond perceived benefits. A plausible interpretation is that while the benefits index captures perceived social benefits of affirmative action, REG group membership also proxies for private or group-specific benefits.

Result 2 (**Social pressure,** γ_i)**:** *In line with Hypothesis 3, thresholds are lower when choices are private rather than public. Thresholds are also lower among individuals who perceive weaker social sanctions for advocating change.*

Support: The estimates in Table 3 show that individuals become more hesitant to act, both in favor or against AA, when their choices are publicly observable: thresholds increase by 4.575 percentage points in the Public condition (column 2), and this effect remains robust with additional controls and when estimated separately for both advocacy conditions (columns 6 and 7).

Individuals are also more hesitant to act when they expect stronger social sanctions for advocating change. Thresholds rise by 20.2 percentage points with an individual's belief about others' likelihood of confronting someone who speaks in favor of change (column 3). This effect likewise remains robust with additional controls and in separate estimations by status quo (columns 6 and 7). As predicted, perceived sanctions raise thresholds most strongly for individuals who are otherwise inclined toward change (see Online Appendix).

Table 4: Reference Groups and Conformity Shift Thresholds Inward

	(1) $t_i^a \notin \{0,100\}$	(2) Dist. to 0 or 100	(3) $t_i^a \notin \{0,100\}$	(4) Dist. to 0 or 100	(5) $t_i^a \notin \{0,100\}$	(6) Dist. to 0 or 100	(7) $t_i^a \notin \{0,100\}$	(8) Dist. to 0 or 100
Social alignment (β_i)								
REG ref/ce group	0.045***	1.745***					0.045***	1.745***
	(0.005)	(0.232)					(0.005)	(0.232)
Conformity index			0.136***	3.531***			0.117***	2.838**
			(0.029)	(1.191)			(0.029)	(1.198)
REG groups								
Asian/Female					0.056**	1.707	0.051*	1.584
					(0.027)	(1.057)	(0.027)	(1.059)
Asian/Male					0.111***	2.680***	0.112***	2.699***
					(0.026)	(1.033)	(0.026)	(1.034)
Black/Female					0.080***	0.991	0.077***	0.907
					(0.026)	(1.040)	(0.026)	(1.042)
Black/Male					0.101***	1.896*	0.097***	1.805*
					(0.026)	(1.020)	(0.026)	(1.020)
Hispanic/Female					0.119***	1.377	0.111***	1.183
					(0.026)	(1.031)	(0.026)	(1.038)
Hispanic/Male					0.173***	6.655***	0.164***	6.429***
					(0.025)	(1.016)	(0.025)	(1.018)
White/Female					0.033	-1.126	0.030	-1.198
					(0.028)	(1.019)	(0.027)	(1.018)
Constant	0.741***	18.203***	0.705***	17.541***	0.680***	17.303***	0.610***	15.293***
	(0.007)	(0.282)	(0.014)	(0.584)	(0.020)	(0.752)	(0.023)	(0.907)
Observations	7,972	7,972	7,972	7,972	7,972	7,972	7,972	7,972
Subjects	3,986	3,986	3,986	3,986	3,986	3,986	3,986	3,986

Notes: OLS regressions with s.e. clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in odd-numbered models is whether or not a threshold is interior, $0 < t_i^a < 100$. The dependent variable in even-numbered models is the distance from the extreme points, $\min(t_i^a, 100 - t_i^a)$. REG ref/ce group is a dummy variable indicating whether group members share the same gender and race/ethnicity. White men are the omitted category in columns 5-8.

Result 3 (Social alignment, β_i): *Most individuals exhibit interdependent behavior. In line with Hypothesis 2, thresholds are more likely to be interior when the reference group is narrower and when individuals report stronger conformity preferences.*

Support: Table 4 tests Hypothesis 2 using two dependent variables: a binary indicator for whether a threshold is interior (i.e., $t_i^a \notin \{0, 100\}$), and a continuous measure of the distance from the nearest extreme (0 or 100). The results show that individuals are more likely to choose interior thresholds when the reference group is narrow: choosing thresholds within one's REG group rather than the U.S. population increases the probability of an interior threshold by 4.5 percentage points (column 1)

and shifts thresholds farther from the extremes (column 2).

Interior thresholds are also more common among individuals with stronger conformity preferences. We find that a shift in the conformity index from 0 to 1 raises the likelihood of an interior threshold by 13.6 percentage points and increases the distance from the extremes by 3.53 points (columns 3–4). These effects remain robust when REG-group controls are included (columns 7–8). Controlling for the benefits index does not affect the estimates (see Online Appendix).

Columns (5) and (6) show that non-White Americans choose interior thresholds significantly more often than White men. These differences largely persist after controlling for individual conformity preferences (columns 7–8), indicating that the greater threshold interiority of non-White Americans is not solely psychological but also cultural or identity-based. One possible explanation is that White men experience greater cultural polarization, with stronger identity cues and targeted media environments reinforcing more extreme positions at both ends of the spectrum.

4.3 The role of forward-looking beliefs

The model highlights a fourth determinant of thresholds: forward-looking beliefs, Δr_i. Individuals who expect their support for change to prompt others to follow should set lower thresholds. To measure such beliefs, participants were asked (after choosing their threshold) to guess how many of the 99 other group members chose thresholds in each of four bins (0–20, 21–50, 51–80, and 81–100), with accuracy financially rewarded.

Beliefs about others' behavior are not mechanically linked to one's own threshold choices: depending on the strategic environment and context—reflected by the underlying threshold distribution—expecting many low-threshold peers could either discourage early action (a free-riding logic) or encourage it (a momentum logic). While both channels are theoretically possible, the free-riding case has an internally conflicting implication: if others are not expected to act, one is then supposed to take the lead, a strategy that carries a high risk of ending up in a minority. By contrast, complementary belief-based behavior arises more naturally: if enough others are perceived as willing to act early, one's own early action becomes safer and more likely to matter.

In our data, this complementary pattern dominates. Respondents who believe

Table 5: Leading Change—Structural Estimates of Model Parameters

	(1) Proxy of Δb_i: Ind. Benefits Index	(2) Proxy of Δb_i: REG Avg. of Benefits Index
Social Alignment ($\hat{\beta}$)	0.846***	0.907***
	(0.050)	(0.117)
Social Pressure ($\hat{\gamma}$)	0.189***	0.234**
	(0.050)	(0.073)
Forward-Looking Beliefs ($\hat{\mu}$)	-0.051***	-0.061***
	(0.012)	(0.013)
Error SD ($\hat{\sigma}$)	0.430***	0.444***
	(0.005)	(0.005)
Observations	7,972	7,972
Subjects	3,986	3,986

Notes: Maximum likelihood estimation of model parameters with standard errors clustered by subject, ** $p < 0.05$, *** $p < 0.01$. We use the benefits index to proxy perceived benefits. Model (1) uses each individual's own index value for Δb_i. Model (2) relies on the REG-level average benefits index to avoid endogeneity with individuals' threshold choices.

that more of their peers choose low thresholds (0–20) also select significantly lower thresholds themselves (-0.213 percentage points per one-point increase in the believed share, OLS $p < 0.001$). This is consistent with anticipatory behavior: individuals who expect a strong base of instigators see their own early action as more likely to help push the group toward social change. We also find evidence for pluralistic ignorance: participants underestimate how many others hold low thresholds. On average, participants believe that 17.8% of others choose thresholds between 0 and 20, whereas the true share is nearly twice as high (34.6%, $p < 0.001$). Conversely, participants overestimate the prevalence of intermediate thresholds (21–80): they believe the share is 60.7%, compared to an actual share of 39.6% ($p < 0.001$).

To integrate forward-looking beliefs with the other determinants of thresholds, we estimate the linear utility model

$$U_i(a_i) = b_i(a_i) - \beta_i r(a_i) - \gamma_i r(a_i) \mathbb{1}_{a_i=1}, \tag{1}$$

which also underlies Figure 2. Optimal thresholds are then (see Online Appendix A),

$$t_i^{a,*} = \frac{\beta + \gamma - \Delta b_i}{2\beta + \gamma} + \frac{\beta + \gamma}{2\beta + \gamma}\epsilon_i, \quad \epsilon_i \sim N(\mu, \sigma^2), \tag{2}$$

where we replace the forward-looking term Δr_i with a shock ϵ_i. Treating expectations as a stochastic component captures heterogeneous beliefs under limited information about the threshold distribution and provides the random variation needed for maximum-likelihood estimation without adding an ad hoc error term.

To estimate the parameters, we proxy Δb_i using each individual's elicited benefits index (Table 5, model 1) or, alternatively, their REG group's average benefits index (model 2), which is exogenous to any given individual. We then exploit variation in observed threshold choices to estimate the parameters governing social alignment ($\hat{\beta}$), social pressure ($\hat{\gamma}$), and forward-looking beliefs ($\hat{\mu}$).

The estimated mean of the belief shock, $\hat{\mu}$, is negative: individuals choose thresholds that are 5.1–6.1 percentage points lower than they would if behaving myopically. This suggests that participants anticipate their own support for change will encourage others to follow suit. The estimates also confirm that social alignment ($\hat{\beta}$) and perceived social pressure ($\hat{\gamma}$) are significant behavioral drivers. While we do not wish to overstate the precision of the structural estimates, the model provides a coherent interpretation of the observed threshold choices.

5 Thresholds and institutional change

By revealing *when* individuals are willing to act, threshold data complement choice data—which capture only *whether* individuals act—and survey-based measures of preferences and beliefs. Threshold data are particularly valuable when there is a tension between the societal status quo and underlying preferences (Andreoni et al., 2021). A classic case is misalignment between formal and informal institutions, where latent support for an action may persist even as legal frameworks shift (North, 1990).

Our main experiment was conducted at a time when federal support for affirmative action was at a high point. Specifically, data collection concluded shortly before the U.S. Supreme Court's June 29, 2023 decision in *Students for Fair Admissions v. Harvard*, which overturned longstanding precedent permitting race-conscious admissions. The subsequent reelection of President Donald J. Trump in late 2024 and his January 2025 executive order terminating affirmative action in federal contracting—along with intensified federal scrutiny of DEI practices in the private sector—marked a pronounced reversal in the formal institutional stance toward affirmative action (Associated Press, 2025).

How might such a change in the formal institutional environment affect individuals' thresholds for action in light of our model? A classic argument in the literature holds that when institutional avenues for influence narrow, individuals increasingly rely on voice—informal or collective expressions of support or dissent—to compensate (Hirschman, 1972). In line with this view, our framework suggests that changes in formal institutional support may alter individuals' behavioral thresholds even in the absence of changes in underlying preferences or beliefs. In particular, while we have no strong reason to expect perceived benefits of affirmative action to change discretely following the institutional shift, what plausibly changes is the level of formal support itself, and with it the mapping from perceived benefits to thresholds.

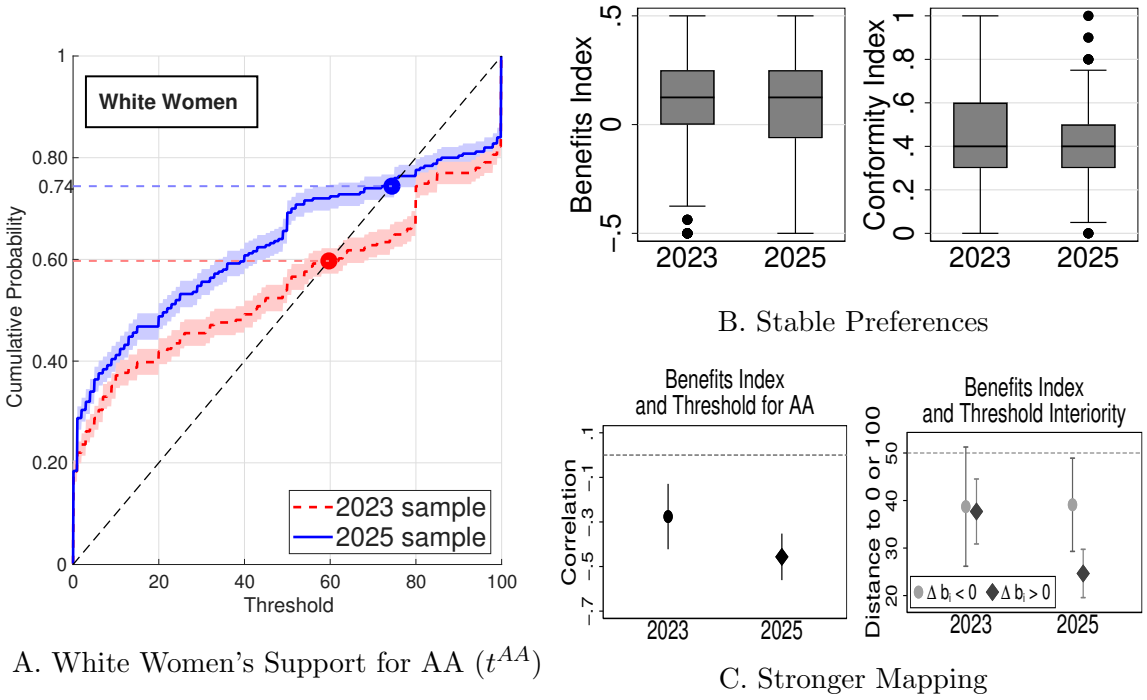
To formalize this intuition, recall that in our baseline model thresholds depend negatively on the perceived marginal benefit of action: as Δb_i increases, t_i decreases (Figure 2). We extend the model by allowing perceived benefits to depend on the prevailing institutional environment. Let S denote the prevailing level of formal institutional support for AA. Individuals can take an advocacy action that affects affirmative action support in addition to S. Let $B_i(\cdot)$ denote individual i's perceived benefit from the overall level of support. The sign of $B_i'(S)$ may be positive or negative, reflecting support or opposition at the prevailing institutional baseline S. We assume $B_i(\cdot)$ is concave, so perceived benefits flatten as support approaches an individual's ideal point (diminishing marginal effects). In our study, choosing pro-AA advocacy corresponds to adding a small increment $\kappa > 0$ to the effective level of support, so the decision-relevant benefit is

$$\Delta b_i(S) \equiv B_i(S + \kappa) - B_i(S) \approx \kappa B_i'(S).$$

When formal institutional support weakens (a decline in S), concavity implies that $B_i'(S)$ increases, raising $\Delta b_i(S)$ and thereby lowering thresholds for pro-AA advocacy—even if absolute attitudes remain unchanged.

This extension yields a clear hypothesis. A decline in formal institutional support for affirmative action should shift thresholds for pro-AA advocacy downward on average. The reason is an increased marginal return to informal action when formal channels weaken. The effect should be heterogeneous. Individuals who favor higher levels of affirmative action should respond more strongly to a decline in formal support. As S moves further from their preferred level, the marginal benefit of advocacy

Figure 5: Changing Thresholds Despite Stable Preferences



A. White Women's Support for AA (t^{AA})

B. Stable Preferences

C. Stronger Mapping

Notes: **A.** Thresholds shift downward (in favor of affirmative action) from 2023 (red, dashed) to 2025 (blue, solid); $N \approx 250$ in both samples. **B.** Distributions of perceived benefits and conformity remain stable over time. **C.** The correlation between perceived benefits and thresholds increases in magnitude from -0.28 to -0.46 ($p = 0.03$). Thresholds of individuals in favor of AA ($\Delta b_i > 0$) move closer to 0 ($p = 0.002$), whereas the thresholds of individuals opposed to AA ($\Delta b_i < 0$) remain at the same distance from 100 ($p = 0.96$).

increases. This strengthens the mapping between perceived benefits and thresholds. By contrast, individuals opposed to affirmative action prefer lower levels of S. For them, a decline in institutional support moves the baseline closer to their preferred level. In this region, concavity implies flatter marginal valuations and more muted threshold responses.

To test these predictions, in June 2025 we recruited a new nationally representative sample of White women—a group with historically mixed views on affirmative action. This timing is particularly informative because informal institutions, consisting of norms, conventions, and shared expectations (North, 1990), tend to adapt slowly (Andreoni et al., 2021; Kamm et al., 2021). The abrupt shift in the formal institutional environment therefore creates a natural setting to examine how thresholds adjust when formal support changes more quickly than underlying norms.

Figure 5 summarizes the results. As shown in Panel A, there is a substantial shift in White women's thresholds between 2023 and 2025. The average threshold for supporting movement toward affirmative action decreased from 44.92 in 2023 to 36.86 in 2025, accompanied by a leftward shift of the entire distribution (K–S, $p = 0.03$), while its overall shape remained similar. The implied equilibrium support for AA increased from 60% to 74%. Hence, reduced formal support for AA is associated with an *increased* willingness of White women to support AA.

Consistent with the model, most of the change reflects increased support among Democrats. Specifically, Democrats account for 72% of the total shift (a 10.03-point reduction in their average thresholds), Independents for 27% (a 5.94-point decline), while Republicans contribute 8% (a 1.49-point decline). The residual difference reflects a sample composition effect: changes in partisan makeup—specifically, a higher share of Republicans and a lower share of Independents—account for 7% less support for AA.[10]

Further evidence supports the mechanism implied by the model. Both the benefits index and the conformity index are nearly identical across the 2023 and 2025 samples (Figure 5B), indicating stable underlying preferences. What changed was the strength of the mapping between perceived benefits and thresholds. In 2025, perceived benefits were more strongly associated with individuals' thresholds (Figure 5C, left panel; $p = 0.03$). The shift is concentrated among AA supporters ($\Delta b_i > 0$): their thresholds moved closer to zero in the direction implied by their perceived benefits (Figure 5C, right panel; $p = 0.002$). In contrast, individuals opposed to affirmative action ($\Delta b_i < 0$) exhibit no comparable change; their distance from 100 remains flat across years.

Data on beliefs reinforce this interpretation. White women's incentivized guesses about how many of the other 99 would ultimately support the pro–affirmative action organization are very similar in 2023 and 2025 (42.46 vs. 40.45, $p = 0.429$).[11] Beliefs about social sanctions also remained stable: the perceived share of others who would

[10] We decompose the change in average thresholds into within-group adjustments and composition effects. For each partisan group, we multiply its 2023 population share by its change in average thresholds between 2023 and 2025. The sum of these within-group contributions is compared to the overall change; the residual is the composition effect reflecting changes in the relative shares of Democrats, Independents, and Republicans between waves.

[11] We elicited two types of beliefs: (i) how many others in one's group of 100 would ultimately support AA (reported above), and (ii) the shares of respondents participants believed would choose thresholds in each bin (0–20, 21–50, 51–80, with 81–100 as the residual). Beliefs remained unchanged across all bins ($p > 0.180$), indicating that the full belief distribution was stable over time.

confront individuals advocating for AA was 39.04% in 2023 and 37.28% in 2025 ($p = 0.450$). With preferences and beliefs unchanged, the observed decline in thresholds isolates a shift in behavioral readiness rather than a reaction to changing expectations.

Taken together, the findings in this section underscore the distinction between preferences, beliefs, and thresholds. Thresholds capture individuals' readiness to act, which is not contained in standard preference measures, and they do so without relying on changes in beliefs. As a result, threshold elicitation provides a useful tool for studying behavioral interdependence. The next section shows how information about the distribution of thresholds can offer novel insights into the aggregate effects of interventions and incentive changes.

6 Using threshold data to anticipate changes in aggregate outcomes

In this section, we show how threshold distributions can be used to reason about aggregate outcomes under interdependence. As noted in the introduction, aggregate outcomes need not respond smoothly to changes in incentives or interventions. When individuals condition their behavior on others' actions, social dynamics can amplify or dampen the effects of a given change, making aggregate responses difficult to infer solely from individual-level preferences. The threshold framework does not imply that aggregate outcomes are fragile; rather, it clarifies when aggregate behavior is stable and when small individual-level changes can have large effects.

Targeting near the equilibrium

Threshold models predict that the aggregate impact of a given incentive change depends critically on how close the thresholds of affected individuals lie to the prevailing equilibrium. As shown above, thresholds differ substantially across REG and political groups. As a result, identical interventions can generate very different aggregate outcomes depending on which segments of the population they target and how those segments are positioned relative to the equilibrium.

Figure 6A provides an example by showing the distribution of t^{AA} after a targeted intervention that reduces thresholds among respondents identified as Republican or Strong Republican or among an equally-sized group of Democrats or Strong Democrats. In both cases, thresholds in the targeted group are lowered by 50 points,

a magnitude that, under the structural estimates, corresponds to an increase in the benefits index to its maximum level.

In the U.S. population, equilibrium support for AA rises from 50% in the absence of an intervention to 68% when thresholds are lowered among Democrats. Targeting Independents instead yields two equilibria at 74–80%, while targeting Republicans produces two equilibria at 77–81%. These differences arise because Republicans are more concentrated near the equilibrium than Democrats. Importantly, the mechanism is not ideological extremity per se: Independents generate shifts of similar magnitude because, despite being less ideologically extreme, their thresholds are likewise clustered near the equilibrium.

Multiple equilibria and tipping points

Threshold models emphasize the role of equilibrium structure in shaping aggregate responses. Identical shifts in individual thresholds can generate smooth adjustments in populations with a unique equilibrium, but trigger transitions between equilibria, or tipping points, when multiple equilibria are present. As a result, aggregate effects depend not only on who is affected by an intervention, but on whether the underlying threshold distribution admits multiple equilibria.

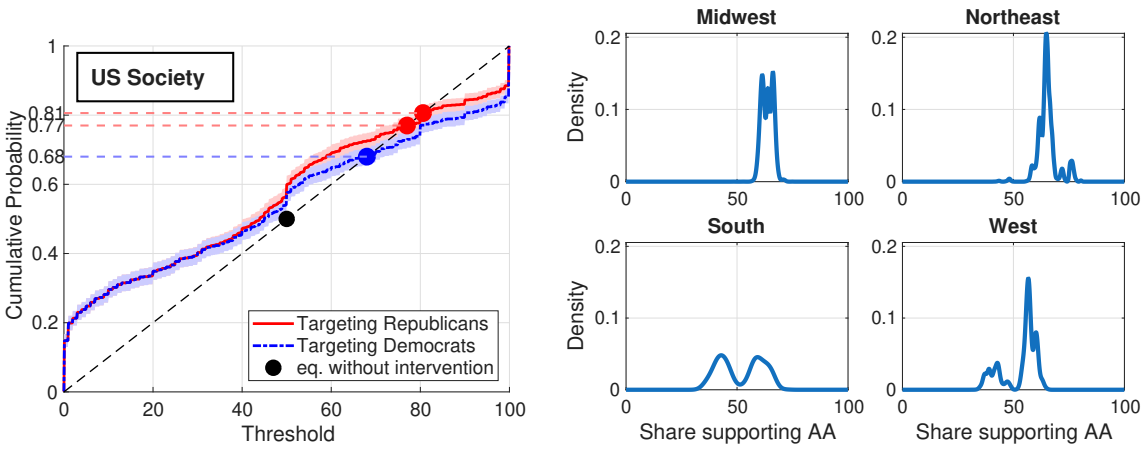
Figure 6b visualizes this logic using different U.S. regions. For each region, we simulate 10,000 groups, record all equilibrium outcomes, and plot the resulting density of equilibrium shares supporting affirmative action. The Northeast, for example, exhibits a single, tightly concentrated mass point at 64%, indicating a unique and stable equilibrium: aggregate outcomes vary little across realizations, and interventions are therefore unlikely to generate discontinuous changes. By contrast, the South displays two distinct mass points, with modes of 42% and 59%, corresponding to multiple equilibria. As in the canonical multiple-equilibrium case illustrated in Figure 1b, this bimodality implies that modest shifts in thresholds can induce transitions between equilibria.[12]

The regional differences in equilibrium structure reflect differences in underlying threshold distributions across subpopulations.[13] We focus on regions because they are

[12]When the equilibrium is unique and tightly concentrated, aggregate outcomes are robust: adding noise or allowing for measurement error in thresholds has essentially no effect on predicted behavior. In settings with multiple equilibria, precise point predictions are structurally difficult; what the framework allows one to anticipate instead is the potential for rapid behavioral shifts in response to modest perturbations.

[13]The South and West have larger Hispanic shares and smaller White shares than the Midwest and

Figure 6: Designing Effective Interventions



A. Targeted Intervention

B. Empirical Equilibrium Structure

Notes: **A.** The figure shows the distribution of t^{AA} after a targeted intervention that lowers thresholds among Republicans or, alternatively, among an equally sized group of Democrats. **B.** The figure displays the distribution of equilibrium AA support across 10,000 simulated societies for each U.S. region. Thresholds are drawn from region-specific samples and aggregated using threshold dynamics. Equilibrium multiplicity arises from the underlying threshold distribution, while equilibrium selection depends on whether early low-threshold supporters trigger additional support among higher-threshold individuals or fail to do so. Appendix B details the simulation procedure.

a natural and policy-relevant unit for assessing the scalability of interventions. Social and political interventions are often deployed uniformly across geographic areas, yet their aggregate effects may differ sharply across regions. The threshold framework clarifies how identical individual-level shifts can produce smooth aggregate changes in some regions but trigger social tipping when there are multiple equilibria. Threshold distributions are therefore useful for assessing the external validity of treatment effects across populations.

Social networks

Network structure shapes whose behavior individuals observe and how quickly actions diffuse. An individual's position within the network determines whether and when thresholds are triggered, as it governs the flow of information and the visibility of others' actions. Our main analysis already captures broad patterns of social exposure by eliciting thresholds under representative and within-REG reference groups.

Northeast (Hispanic: 19–31% vs. 9–15%; White: 54–59% vs. 66–78%). This greater compositional heterogeneity is consistent with the higher prevalence of multiple equilibria in those regions.

When more detailed network information is available, the same framework can readily incorporate it, as we illustrate next.

To proxy respondents' social environments, we elicited the gender, racial or ethnic, and political-affiliation composition of the ten people with whom they most recently exchanged opinions.[14] These elicited networks allow us to incorporate observed segregation into the computation of social equilibria. In a fully segregated society, predicted support for affirmative action reaches 86 percent among Black women but only 31 percent among White men. In a representative society—where exposure mirrors population shares—support converges to 71 and 50 percent, respectively. Using respondents' actual networks yields 80 percent predicted support for Black women (close to the segregated benchmark) and around 50 percent for White men (close to the representative benchmark), with the other REG groups falling in between.

These patterns illustrate how combining threshold data with network information yields insights into the role of social networks across settings. In our sample, further increasing the diversity of social networks would have only minimal aggregate effects because most elicited networks are already demographically mixed enough to generate close-to-representative outcomes. By contrast, rising segregation—especially among White men, who currently operate close to the representative benchmark—would reduce overall support. Eliciting both thresholds and networks thus provides a realistic basis for identifying where social equilibria are robust and where they are vulnerable to social fragmentation. More generally, we found that small perturbations to observed network composition have limited effects on predicted aggregate outcomes, whereas large aggregate shifts can arise when network structure itself changes *systematically*, such as sustained increases in segregation.

7 Conclusion

In many economically and socially relevant settings, individuals condition their actions on the behavior of others, giving rise to dynamics that can amplify or attenuate the effects of incentives, policies, and institutions. In such environments, observed choices alone provide a limited view of underlying support, as they reflect equilibrium

[14] The questions were: *"Among the ten people you most recently met—outside your family—with whom you exchanged opinions, how many do you think identify as (Male / Female / Other)? (Republican / Democrat)? (your racial or ethnic group / not your racial or ethnic group)?"* The third question directly referenced participants' racial or ethnic identity.

behavior rather than individuals' readiness to act under alternative social conditions. Eliciting attitudes is likewise insufficient, as standard opinion measures abstract from how willingness to act depends on others' behavior. This paper contributes a method to study behavioral interdependence directly by eliciting individuals' thresholds for action. While threshold models have long been used to capture interdependence in theory, threshold heterogeneity has not been directly measured in empirical work, and the determinants of thresholds remain largely unexplored.

Applying our method to study support for affirmative action in the United States, we document substantial and systematic heterogeneity in thresholds across racial, ethnic, gender, and political groups, in line with preregistered hypotheses. Thresholds are shaped by perceived benefits and expectations about others' behavior, and they respond sharply to changes in the institutional environment. In particular, the weakening of federal support for affirmative action appears to have altered individuals' readiness to act even as underlying preferences and beliefs remained largely stable. These findings highlight the importance of distinguishing between preferences, beliefs, and thresholds when studying socially interdependent behavior.

Thresholds provide insights that are difficult to obtain from choice or survey data alone. As aggregate outcomes depend on the distribution of thresholds, identical changes can generate very different collective responses depending on which individuals are affected and how they are positioned relative to the prevailing equilibrium. By making threshold distributions observable, our method helps anticipate when interdependence will amplify or dampen the effects of policy interventions, assess the external validity of interventions across populations, and examine how changes in group composition or social exposure shape collective outcomes.

References

Akerlof, G. and R. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics 94*(4), 749–775.

Alesina, A., M. Ferroni, and S. Stantcheva (2021). Perceptions of racial gaps, their causes, and ways to reduce them. *National Bureau of Economic Research #29245*.

Andreoni, J., N. Nikiforakis, and S. Siegenthaler (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences 118*(16), e2014893118.

Associated Press (2025, January 21). Trump administration directs all federal diversity, equity and inclusion staff be put on leave. Accessed: 2025-11-24.

Banerjee, A., E. Breza, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2024). Can a trusted messenger change behavior when information is plentiful? evidence from the first months of the covid-19 pandemic in west bengal. *Review of Economics and Statistics 106*(5), 945–960.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics 107*(3), 797–817.

Baumann, L. and W. Olszewski (2021). Demand cycles and heterogeneous conformity preferences. *Journal of Economic Theory 194*, 105252.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Berger, J., C. Efferson, and S. Vogt (2023). Tipping pro-environmental norm diffusion at scale: opportunities and limitations. *Behavioural Public Policy 7*(3), 581–606.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy 102*(5), 841–877.

Bicchieri, C. (2006). *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C., E. Dimant, S. Gächter, and D. Nosenzo (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior 132*, 59–72.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy 100*(5), 992–1026.

Bleemer, Z. (2022). Affirmative action, mismatch, and economic mobility after California's proposition 209. *Quarterly Journal of Economics 137*(1), 115–160.

Boucher, V., M. Rendall, P. Ushchev, and Y. Zenou (2024). Toward a general theory of peer effects. *Econometrica 92*(2), 543–565.

Brandts, J. and G. Charness (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics 14*, 375–398.

Bursztyn, L., G. Egorov, and S. Fiorin (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review 110*(11), 3522–3548.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American Economic Review 110*(10), 2997–3029.

Card, D., A. Mas, and J. Rothstein (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics 123*(1), 177–218.

Carlsson, H. and E. van Damme (1993). Global games and equilibrium selection. *Econometrica 61*(5), 989–1018.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica 81*(3), 855–882.

Centola, D. (2015). The social origins of networks and diffusion. *American Journal of Sociology 120*(5), 1295–1338.

Centola, D., J. Becker, D. Brackbill, and A. Baronchelli (2018). Experimental evidence for tipping points in social convention. *Science 360*(6393), 1116–1119.

Chinoy, S., N. Nunn, S. Sequeira, and S. Stantcheva (2026). Zero-sum thinking and the roots of U.S. political differences. *American Economic Review*. Forthcoming.

Constantino, S. M., G. Sparkman, G. T. Kraft-Todd, C. Bicchieri, D. Centola, B. Shell-Duncan, S. Vogt, and E. U. Weber (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest 23*(2), 50–97.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550.

Duffy, J., J. Ochs, and L. Vesterlund (2007). Giving little by little: Dynamic voluntary contribution games. *Journal of Public Economics 91*(9), 1708–1730.

Durlauf, S. N. and Y. M. Ioannides (2010). Social interactions. *Annual Review of Economics 2*(1), 451–478.

Efferson, C., S. Vogt, A. Elhadi, H. E. F. Ahmed, and E. Fehr (2015). Female genital cutting is not a social coordination norm. *Science 349*(6255), 1446–1447.

Efferson, C., S. Vogt, and E. Fehr (2020). The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour 4*(1), 55–68.

Ehret, S., S. M. Constantino, E. U. Weber, C. Efferson, and S. Vogt (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour 6*(12), 1669–1679.

Fatas, E., S. P. H. Heap, and D. R. Arjona (2018). Preference conformism: An experiment. *European Economic Review 105*, 71–82.

Fehérová, M., S. Heger, J. Péliová, M. Servátka, and R. Slonim (2022). Increasing autonomy in charitable giving: The effect of choosing the number of recipients on donations. *Economics Letters 217*, 110701.

Fehr, E. and U. Fischbacher (2004). Third party sanctions and social norms. *Evolution and Human Behavior 25*(2004), 63–87.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Fischbacher, U. and S. Gächter (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review 100*(1), 541–556.

Fischbacher, U., S. Gächter, and E. Fehr (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters 71*(3), 397–404.

Galeotti, A. and S. Goyal (2009). Influencing the influencers: a theory of strategic diffusion. *The RAND Journal of Economics 40*(3), 509–532.

Galeotti, A., S. Goyal, M. O. Jackson, F. Vega-Redondo, and L. Yariv (2010). Network games. *The Review of Economic Studies 77*(1), 218–244.

Goeree, J. K. and L. Yariv (2015). Conformity in the lab. *Journal of the Economic Science Association 1*(1), 15–28.

Goldsmith, R. E., R. A. Clark, and B. A. Lafferty (2005). Tendency to conform: A new measure and its relationship to psychological reactance. *Psychological Reports 96*(3), 591–594.

Goyal, S. (2023). *Networks: An economics approach*. MIT Press.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology 83*(6), 1420–1443.

Granovetter, M. and R. Soong (1986). Threshold models of interpersonal effects in consumer demand. *Journal of Economic Behavior & Organization 7*(1), 83–99.

Heinemann, F., R. Nagel, and P. Ockenfels (2004). The theory of global games on test: experimental analysis of coordination games with public and private information. *Econometrica 72*(5), 1583–1599.

Heinemann, F., R. Nagel, and P. Ockenfels (2009). Measuring strategic uncertainty in coordination games. *The Review of Economic Studies 76*(1), 181–221.

Hirschman, A. (1972). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.

Holzer, H. J. and D. Neumark (2000). Assessing affirmative action. *Journal of Economic Literature 38*(3), 483–568.

Hong, S.-M. and S. Faedda (1996). Refinement of the hong psychological reactance scale. *Educational and Psychological Measurement 56*(1), 173–182.

Hong, S.-M. and S. Page (1989). A psychological reactance scale: Development, factor structure and reliability. *Psychological Reports 64*(3), 1323–1326.

Jackson, M. O. (2008). *Social and economic networks*, Volume 3. Princeton University Press.

Jackson, M. O. and L. Yariv (2005). Diffusion on social networks. *Économie Publique 16*, 3–16.

Jackson, M. O. and L. Yariv (2007). Diffusion of behavior and equilibrium properties in network games. *American Economic Review 97*(2), 92–98.

Kamm, A., C. Koch, and N. Nikiforakis (2021). The ghost of institutions past: History as an obstacle to fighting tax evasion? *European Economic Review 132*, 103641.

Katz, M. L. and C. Shapiro (1985). Network externalities, competition, and compatibility. *American Economic Review 75*(3), 424–440.

Katz, M. L. and C. Shapiro (1986). Technology adoption in the presence of network externalities. *Journal of Political Economy 94*(4), 822–841.

Kuran, T. (1995). The inevitability of future revolutionary surprises. *American Journal of Sociology 100*(6), 1528–1551.

List, J. A. and D. Lucking-Reiley (2002). The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. *Journal of Political Economy 110*(1), 215–233.

Macy, M. W. (1991). Chains of cooperation: Threshold effects in collective action. *American Sociological Review 56*(6), 730–747.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives 14*(3), 115–136.

My, K. B., M. Brunette, S. Couture, and S. Van Driessche (2024). Are ambiguity preferences aligned with risk preferences? *Journal of Behavioral and Experimental Economics 111*, 102237.

North, D. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.

Oechssler, J., A. Reischmann, and A. Sofianos (2022). The conditional contribution mechanism for repeated public goods–the general case. *Journal of Economic Theory 203*, 105488.

Oliver, P., G. Marwell, and R. Teixeira (1985). A theory of the critical mass. I. interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology 91*(3), 522–556.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review 83*(5), 1281–1302.

Roland, G. and T. Verdier (1994). Privatization in Eastern Europe: Irreversibility and critical mass effects. *Journal of Public Economics 54*(2), 161–183.

Scheffer, M. (2020). *Critical transitions in nature and society*, Volume 16. Princeton University Press.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton & Company.

Schmidt, K. M. and A. Ockenfels (2021). Focusing climate negotiations on a uniform common commitment can promote cooperation. *Proceedings of the National Academy of Sciences 118*(11), e2013070118.

Simmons, B. A. and Z. Elkins (2004). The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review 98*(1), 171–189.

Szkup, M. and I. Trevino (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior 124*, 534–553.

Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences 108*(29), 11825–11829.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review 99*(5), 1899–1924.

Zhang, J. (2011). Tipping and residential segregation: a unified Schelling model. *Journal of Regional Science 51*(1), 167–193.

Online Appendix

A Comparative statics of the utility-based model of threshold determinants

General result

As in the paper, we consider the utility difference between supporting $a = 1$ and not supporting,

$$\Delta U_i\big(r(0), \Delta b_i, \beta, \gamma, \Delta r_i\big) \equiv U_i(1, r(1), \Delta b_i, \beta, \gamma) - U_i(0, r(0), \Delta b_i, \beta, \gamma),$$

where $r(0)$ is the fraction of others supporting the action when i does not support, and

$$\Delta r_i \equiv r(1) - r(0) \geq 0$$

captures i's perceived marginal impact on the adoption rate: the fraction of others who would support $a = 1$ only if i also does so. Individual i chooses a threshold t_i^a and supports $a = 1$ if and only if $r(0) \geq t_i^a$. The optimal threshold $t_i^{a,*}$ is defined by the indifference condition

$$\Delta U_i\big(t_i^{a,*}, \Delta b_i, \beta, \gamma, \Delta r_i(t_i^{a,*})\big) = 0, \tag{A.1}$$

where $\Delta r_i(t)$ denotes the beliefs about marginal impact evaluated at threshold t.

We first characterize how $\Delta r_i(t)$ can change with the threshold. This argument depends only on the threshold logic and not on the specific utility representation.

Lemma A.1 (Belief monotonicity). Let $t_1 > t_0$ be two thresholds and define $\Delta m_i = \Delta r_i(t_1) - \Delta r_i(t_0)$. Then $\Delta m_i \in [t_0 - t_1, 0]$. In particular, if Δr_i is differentiable, this implies $\Delta r_i'(t) \in [-1, 0]$.

Proof. The upper bound $\Delta m_i \leq 0$ follows because, by increasing their threshold from t_0 to t_1, individual i cannot become pivotal for any additional person: some people for whom i was previously pivotal may no longer be, but the reverse cannot occur. For the lower bound, consider two cases. If $t_0 + \Delta r_i(t_0) \leq t_1$, then $\Delta r_i(t_1) = 0$ and $\Delta m_i = -\Delta r_i(t_0) \geq t_0 - t_1$ by construction. If instead $t_0 + \Delta r_i(t_0) > t_1$, then for any individual j with threshold $t_j > t_1$, i is pivotal for j at t_0 if and only if i is pivotal for

j at t_1 (conditional on $a_i(t_0) = a_i(t_1) = 1$). Hence, only individuals with thresholds in $[t_0, t_1]$ can stop being influenced when i raises their threshold, implying $\Delta m_i = t_0 - t_1$ in this worst-case configuration. Overall, $t_0 - t_1 \leq \Delta m_i \leq 0$. If Δr_i is differentiable at t, divide the inequality $t_0 - t_1 \;\leq\; \Delta r_i(t_1) - \Delta r_i(t_0) \;\leq\; 0$ by $t_1 - t_0 > 0$ and let $t_1 \to t_0 = t$. The resulting limit is the derivative of Δr_i, implying $\Delta r_i'(t) \in [-1, 0]$. □

We now derive the comparative statics of $t_i^{a,*}$ with respect to the parameters $x \in \{\Delta b_i, \beta, \gamma\}$. Differentiating (A.1) with respect to parameter x gives

$$\frac{\partial \Delta U_i}{\partial r(0)} \frac{dt_i^{a,*}}{dx} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \Delta r_i'(t_i^{a,*}) \frac{dt_i^{a,*}}{dx} + \frac{\partial \Delta U_i}{\partial x} = 0,$$

so

$$\frac{dt_i^{a,*}}{dx} = -\frac{\frac{\partial \Delta U_i}{\partial x}}{\frac{\partial \Delta U_i}{\partial r(0)} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \Delta r_i'(t_i^{a,*})}, \qquad x \in \{\Delta b_i, \beta, \gamma\}. \tag{A.2}$$

To interpret the denominator, it is helpful to note how the model is constructed. First, an increase in the current adoption rate $r(0)$ raises the relative attractiveness of supporting: if more people are already on the side of change, joining them reduces the penalty from being in a minority. This implies

$$\frac{\partial \Delta U_i}{\partial r(0)} > 0.$$

Second, an increase in the marginal impact Δr_i makes supporting at least weakly more attractive: if acting brings along more followers, the social alignment and social pressure benefits of acting are larger, while the payoff from not acting is unchanged. Thus

$$\frac{\partial \Delta U_i}{\partial \Delta r_i} \geq 0.$$

Moreover, the effect of $r(0)$ on ΔU_i is at least as large as the effect of Δr_i. A higher $r(0)$ affects the utilities of acting and not acting in opposite directions (it makes joining more attractive and remaining passive less attractive), whereas Δr_i affects only the payoff from acting. Formally, note that

$$\frac{\partial \Delta U_i}{\partial \Delta r_i} = \frac{\partial U_i(1, r(1), \cdot)}{\partial r(1)},$$

because Δr_i enters ΔU_i only through $r(1) = r(0) + \Delta r_i$. In contrast,

$$\frac{\partial \Delta U_i}{\partial r(0)} = \frac{\partial U_i(1, r(1), \cdot)}{\partial r(1)} \; - \; \frac{\partial U_i(0, r(0), \cdot)}{\partial r(0)},$$

since $r(0)$ affects both the acting and non-acting payoffs. Since the model implies

$$\frac{\partial U_i(1, r)}{\partial r} \geq 0 \quad \text{and} \quad \frac{\partial U_i(0, r)}{\partial r} \leq 0,$$

we obtain

$$0 \leq \frac{\partial \Delta U_i}{\partial \Delta r_i} \Big/ \frac{\partial \Delta U_i}{\partial r(0)} \leq 1.$$

Define

$$\rho_i(t) \equiv \frac{\partial \Delta U_i / \partial \Delta r_i}{\partial \Delta U_i / \partial r(0)}.$$

By the previous argument, $\rho_i(t) \in [0, 1]$. We can now rewrite the denominator in (A.2) as

$$\frac{\partial \Delta U_i}{\partial r(0)} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \, \Delta r_i'(t_i^{a,*}) = \frac{\partial \Delta U_i}{\partial r(0)} \left[\, 1 + \rho_i(t_i^{a,*}) \, \Delta r_i'(t_i^{a,*}) \, \right].$$

We rule out the degenerate case in which a single individual both fully internalizes and fully determines the marginal adoption rate, i.e., $\rho_i(t_i^{a,*}) = 1$ and $\Delta r_i'(t_i^{a,*}) = -1$. Since $\rho_i(t_i^{a,*}) \in [0, 1]$ and, by Lemma A.1, $\Delta r_i'(t_i^{a,*}) \in [-1, 0]$, the bracket term lies in $[1 - \rho_i(t_i^{a,*}), 1] \subset (0, 1]$. Hence

$$0 < \frac{\partial \Delta U_i}{\partial r(0)} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \, \Delta r_i'(t_i^{a,*}) \leq \frac{\partial \Delta U_i}{\partial r(0)}.$$

Two implications follow. First, the denominator in (A.2) is strictly positive, so the sign of $dt_i^{a,*}/dx$ is the opposite of the sign of $\partial \Delta U_i / \partial x$. Second, endogenous beliefs lower the denominator, implying that the magnitude of $dt_i^{a,*}/dx$ is weakly larger. Forward-looking beliefs about marginal impact can therefore amplify how strongly the threshold responds to changes in Δb_i, β, or γ, but they can never reverse the direction of these responses.

Linear utility specification

We now illustrate the general result in the linear specification used in Figure 2. Let individual i's utility be given by

$$U_i(a_i) = b_i(a_i) - \beta\, r(a_i) - \gamma\, r(a_i)\, \mathbb{1}_{a_i=1}, \tag{A.3}$$

which is a specific functional form of the general utility function $U_i(a_i, r(a_i), \Delta b_i, \beta, \gamma)$. As before, $a_i \in \{0, 1\}$ indicates whether i supports the status quo ($a_i = 0$) or the alternative ($a_i = 1$) organization. The variable $b_i(a_i) \in \mathbb{R}$ represents the perceived benefits of each action. The second and third terms in (A.3) represent misalignment costs. Specifically, the variable

$$r(a_i) = \frac{1}{n-1} \sum_{j \neq i} \mathbb{1}_{a_j \neq a_i}$$

indicates the fraction of others who choose an organization different from the one selected by i. The parameter $\beta > 0$ captures i's concern for social alignment: i suffers a disutility which increases with the number of others who choose a different action. The parameter $\gamma > 0$ captures social pressure in favor of the status quo organization: when i supports the alternative ($a_i = 1$), the cost of being in a minority rises with the number of others selecting the status quo.

Individual i must select a threshold t_i^a for supporting a. The optimal threshold represents the minimum proportion of others supporting a beyond which individual i prefers to choose $a = 1$ as well, given their perceived benefits, social alignment concerns, and social pressure. The optimal threshold, $t_i^{a,*}$, is determined as the value of $r(0)$ at which $U_i(0) = U_i(1)$, resulting in:

$$t_i^{a,*} = \frac{\beta + \gamma - \Delta b_i}{2\beta + \gamma} - \frac{\beta + \gamma}{2\beta + \gamma} \Delta r_i(t_i^{a,*}), \tag{A.4}$$

where $\Delta b_i \equiv b_i(1) - b_i(0) \in \mathbb{R}$ represents the net perceived benefit of choosing $a_i = 1$, and $\Delta r_i(t_i^{a,*}) \geq 0$ reflects i's expectation of the (marginal) increase in the fraction of others supporting a when $a_i = 1$ rather than $a_i = 0$. The argument in Lemma A.1 applies here as well and implies $\Delta r_i'(t_i^{a,*}) \in [-1, 0]$ when Δr_i is differentiable.

We are now ready to derive the comparative statics in the linear model. Rear-

ranging (A.4), we can write

$$t_i^{a,*} = \frac{\beta + \gamma - \Delta b_i}{2\beta + \gamma} - \rho_i^{\text{lin}} \, \Delta r_i(t_i^{a,*}), \qquad \text{where} \quad \rho_i^{\text{lin}} \equiv \frac{\beta + \gamma}{2\beta + \gamma}.$$

In this linear specification, the ratio introduced in the general case simplifies to a constant,

$$\rho_i(t_i^{a,*}) = \rho_i^{\text{lin}} = \frac{\beta + \gamma}{2\beta + \gamma} \in (0, 1),$$

so the denominator term from (A.2) becomes

$$\frac{\partial \Delta U_i}{\partial r(0)} + \frac{\partial \Delta U_i}{\partial \Delta r_i} \, \Delta r_i'(t_i^{a,*}) = \frac{\partial \Delta U_i}{\partial r(0)} \Big[1 + \rho_i^{\text{lin}} \, \Delta r_i'(t_i^{a,*}) \Big].$$

Define

$$z \equiv 1 + \frac{\beta + \gamma}{2\beta + \gamma} \, \Delta r_i'(t_i^{a,*}) = 1 + \rho_i^{\text{lin}} \, \Delta r_i'(t_i^{a,*}).$$

Since $\rho_i^{\text{lin}} \in (0, 1)$ and Lemma A.1 implies $\Delta r_i'(t_i^{a,*}) \in [-1, 0]$, we have

$$z \in \Big[1 - \rho_i^{\text{lin}}, \, 1 \Big] = \Big[\tfrac{\beta}{2\beta+\gamma}, \, 1 \Big] \subset (0, 1],$$

so the linear model inherits the same amplification logic as the general case: forward-looking beliefs about marginal impact reduce the denominator via z, but never change its sign.

Implicit differentiation of (A.4) then yields

$$\frac{\partial t_i^{a,*}}{\partial \Delta b_i} = -\frac{1}{(2\beta + \gamma) \cdot z} < 0, \tag{A.5}$$

which shows that thresholds decrease with Δb_i. Next, we have

$$\frac{\partial t_i^{a,*}}{\partial \gamma} = \frac{\beta + \Delta b_i - \beta \Delta r_i(t_i^{a,*})}{(2\beta + \gamma)^2 \cdot z} > 0 \tag{A.6}$$

as long as $t_i^{a,*} < 1$. The derivative becomes negative for $t_i^{a,*} > 1$, but in these cases, the change in γ has no actual impact on i's threshold choice because the threshold

choice is $t_i = 1$ in any case. For social alignment, we obtain

$$\frac{\partial t_i^{a,*}}{\partial \beta} = \frac{2\Delta b_i - \gamma + \gamma \Delta r_i(t_i^{a,*})}{(2\beta + \gamma)^2 \cdot z} \text{ which is } \begin{cases} > 0 & \text{if } t_i^{a,*} < (1 - \Delta r_i(t_i^{a,*}))/2, \\ < 0 & \text{if } t_i^{a,*} > (1 - \Delta r_i(t_i^{a,*}))/2. \end{cases} \tag{A.7}$$

Therefore, an increase in social alignment shifts thresholds toward $(1 - \Delta r_i(t_i^{a,*}))/2$, which is generically close to one half. Even in cases where it is not, a greater β pushes thresholds to some intermediate point, away from the extremes of 0% and 100%. The exact value of this intermediate point can be below one half if individuals expect that by choosing a slightly lower threshold they can induce enough others to support change, thus aligning their choice with more than 50% of others. For example, an individual with a large β may choose a threshold of 47% if she expects this deviation from 50% will result in 52% of others supporting change (thus aligning her choice with more than half of others). In the absence of such forward-looking effects, social alignment pushes thresholds toward exactly 50%. In our data, $\Delta r_i(1/2)$ is less than one percent, and for generic threshold distributions F, it will be small in large groups.

B Additional Results, Analyses, and Details

B.1 Descriptive overview of Threshold averages

To summarize these patterns systematically, Table B.1 presents average thresholds across groups and experimental conditions. Consistent with Hypothesis 3, average thresholds are higher in the Public than in the Private condition. This pattern appears in the U.S. population overall and in 12 of the 16 REG group–based comparisons of t^{AA} and t^{NoAA}. Consistent with Hypothesis 1, White men exhibit higher thresholds for supporting AA (t^{AA}) than underrepresented groups in 11 of 14 comparisons (seven REG groups, each examined under both visibility conditions). Underrepresented groups show lower t^{AA} than t^{NoAA} in 13 of 14 comparisons, whereas the reverse holds for White men. We also observe systematic demographic differences: women have lower t^{AA} than men, and Asian, Black, and Hispanic Americans have lower thresholds than White Americans. Finally, Republicans' average t^{AA} (t^{NoAA}) are far above (below) those of Democrats.

Table B.1 also shows that the share of individuals with interior thresholds ($1 \leq$

Table B.1: Threshold Choices Across Groups and Conditions

	Average Thresholds				Percentage of Interior Thresholds			
Advocacy	Pro-AA (t^{AA})		Anti-AA (t^{NoAA})		Pro-AA (t^{AA})		Anti-AA (t^{NoAA})	
Reference Group	U.S.		U.S.		U.S.	REG	U.S.	REG
Visibility	Public	Private	Public	Private	Public		Public	
Asian/Female	46.22	42.98	54.97	52.82	72.36	75.37	68.59	72.25
Asian/Male	49.32	52.02	57.11	60.19	76.81	81.64	79.10	83.58
Black/Female	35.97	34.22	48.23	53.68	76.96	83.25	69.31	75.74
Black/Male	38.11	29.00	46.15	46.82	75.47	80.66	71.81	80.32
Hispanic/Female	47.78	36.58	53.67	45.98	74.53	79.25	74.43	82.39
Hispanic/Male	45.32	42.85	48.16	42.73	84.06	87.92	82.67	85.64
White/Female	44.93	37.69	53.40	38.93	64.21	70.00	65.96	72.87
White/Male	53.97	39.78	43.58	39.94	63.33	67.62	65.61	68.25
Female	44.41	37.46	52.80	43.13	68.66	74.09	68.16	74.96
Male	49.93	39.77	45.74	42.70	69.79	74.15	70.81	74.39
Asian	47.71	46.89	56.01	56.64	74.50	78.39	73.68	77.74
Black	37.05	31.82	47.27	50.44	76.21	81.93	70.47	77.86
Hispanic	46.56	39.67	50.71	44.41	79.27	83.56	78.86	84.14
White	49.68	38.57	48.47	39.42	63.75	68.75	65.78	70.55
Democrat	36.66	30.24	56.05	48.59	72.82	77.50	67.50	72.47
Independent	53.89	47.87	45.28	36.72	67.19	73.97	67.67	72.95
Republican	57.03	40.16	40.36	37.02	65.71	68.56	73.51	78.87
U.S. Population	**47.25**	**38.47**	**49.25**	**42.92**	**69.24**	**74.12**	**69.50**	**74.67**

Notes: Average thresholds are reported for both advocacy conditions (t^{AA}/t^{NoAA}) and both visibility conditions (Public/Private) for the U.S. population reference group. Interior-threshold shares are reported by advocacy conditions and reference group (U.S./REG) for the Public visibility condition. The U.S. population row uses ACS 2021 weights for REG groups (Asian F 0.035, Asian M 0.031, Black F 0.069, Black M 0.064, Hispanic F 0.094, Hispanic M 0.098, White F 0.304, White M 0.305).

$t_i \leq 99$) aligns with the model's predictions. A substantial majority—between 69.24% and 74.67% of the U.S. population, depending on the condition—hold interior thresholds, while the remaining 25.33% to 30.76% have thresholds of either 0 or 100. Consistent with Hypothesis 2, the share of interior thresholds is higher when individuals can condition their actions on members of their own REG group rather than on broader reference groups. This pattern appears in all 16 comparisons (eight REG groups under two advocacy conditions). White men and women exhibit a lower share of interior thresholds than other REG groups.

B.2 Threshold distributions in each REG group

Figure 4 of the paper shows the distribution of REG thresholds for AA of Black and White men and women (t^{AA}). Figure B.1 displays the remaining REG groups. Additionally, Figure B.2 shows the CDFs of all eight REG groups of t^{NoAA}. There exists a large heterogeneity between REG groups in threshold distributions of t^{AA}, with societal equilibria ranging from 39% for AA (Asian men) to 82% and 91% for AA (Black and Hispanic men). The distribution of t^{AA} of Asian men and Hispanic women shows multiple equilibria. Similarly, the distributions of t^{NoAA} are heterogeneous between REG groups, with societal equilibria ranging from 9% and 18% (Hispanic men and women) to 67% and 88% (Black women and Hispanic men). The distribution for Hispanic men shows great tipping potential with multiple equilibria, one at a low level and two at a high level.

B.3 Structural estimation

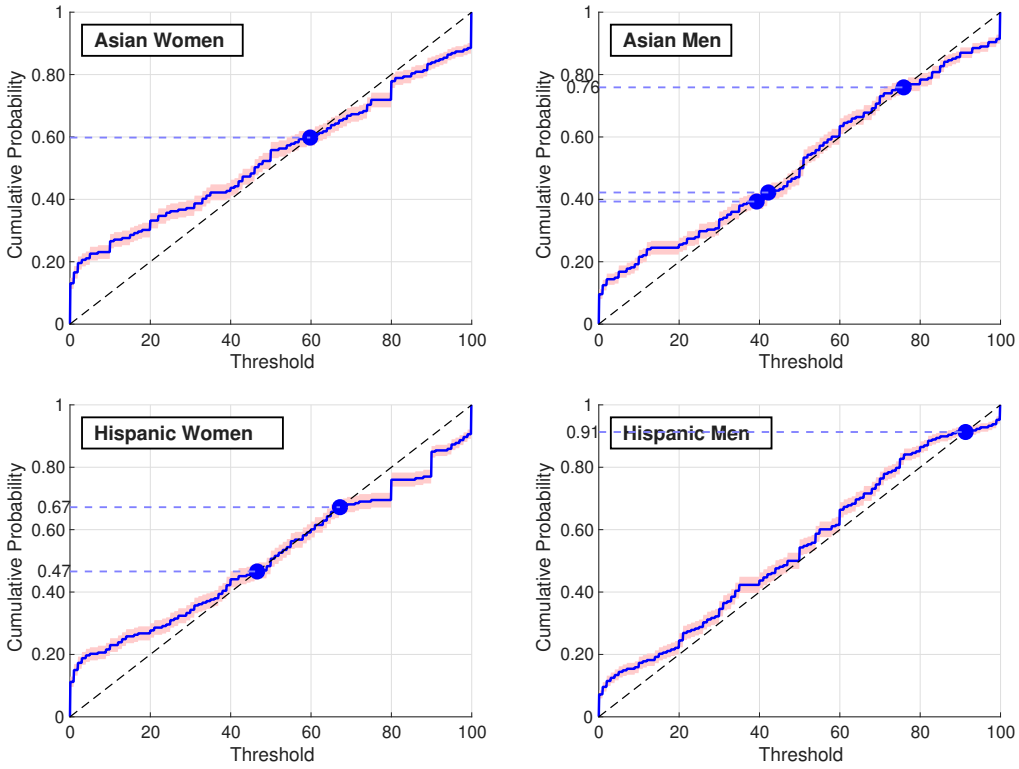
Assuming a linear utility function with homogeneous conformity parameters β and γ across individuals:

$$U_i(a_i) = b_i(a_i) - \beta_i r(a_i) - \gamma_i r(a_i) \mathbb{1}_{a_i=1}, \tag{B.8}$$

the individual's optimal threshold is given by

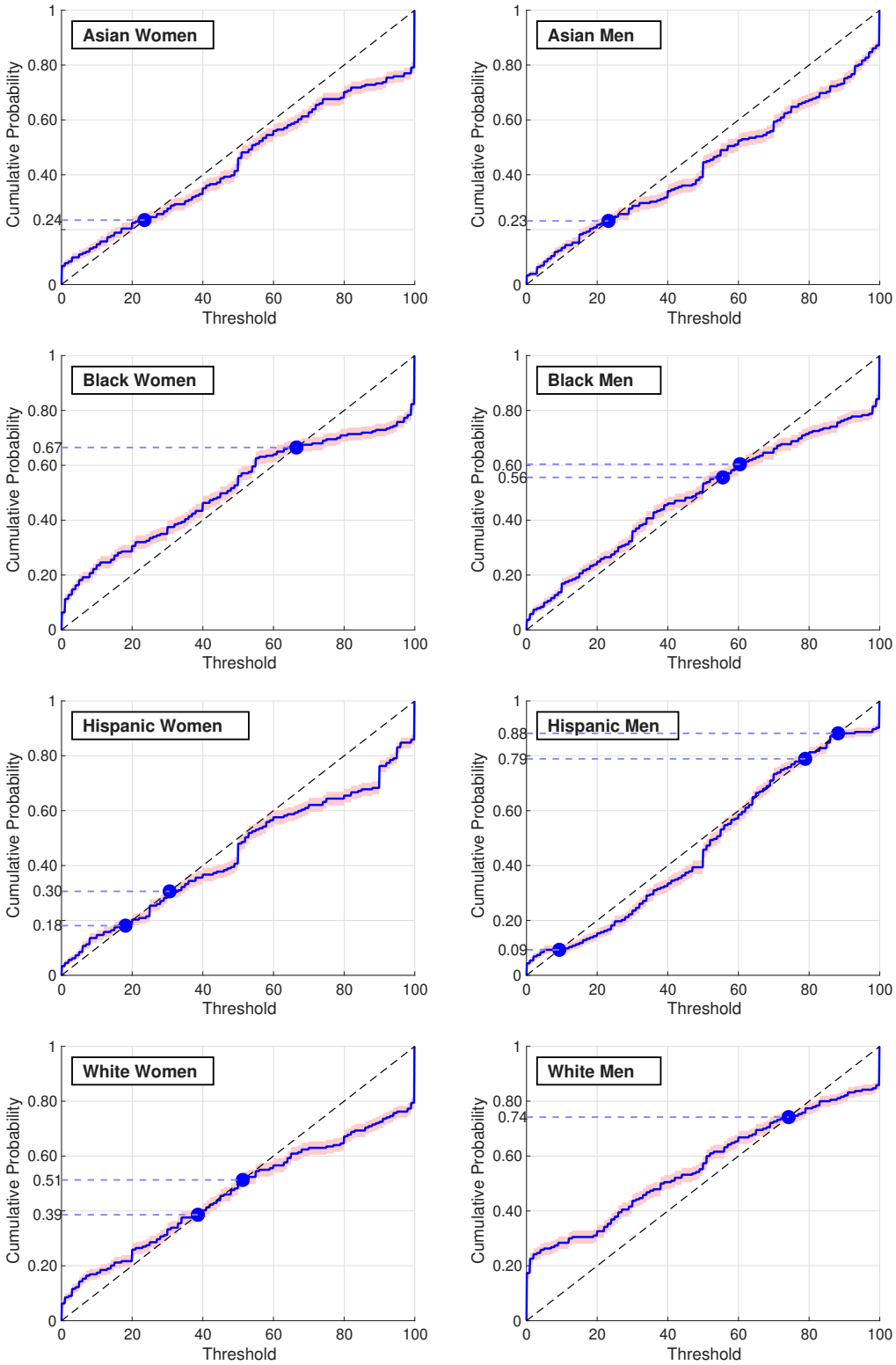
$$t_i^{a,*} = \frac{\beta_i + \gamma_i - \Delta b_i}{2\beta_i + \gamma_i} - \frac{\beta_i + \gamma_i}{2\beta_i + \gamma_i} \Delta r_i, \tag{B.9}$$

Figure B.1: Threshold distributions for AA in REG segregated groups



Notes: Distribution of thresholds for AA (t^{AA}) of different REG segregated groups in the public condition and for narrow (REG) reference groups. Shades depict the 90% confidence intervals of the CDFs when randomly sampling 10,000 times groups of $n = 1,000$. Markers depict societal equilibria.

Figure B.2: Threshold distributions against AA in REG segregated groups



Notes: Distribution of thresholds against AA (t^{NoAA}) of different REG segregated groups in the public condition and for narrow (REG) reference groups. Shades depict the 90% confidence intervals of the CDFs when randomly sampling 10,000 times groups of $n = 1,000$. Markers depict societal equilibria.

where $\Delta r_i \geq 0$ reflects i's expectation of the (marginal) increase in the fraction of others supporting a when $a_i = 1$ rather than $a_i = 0$. To quantify forward-looking behavior and the role of conformity, we estimate the parameters of the threshold expression by assuming that Δr_i is unobserved and replaced by an idiosyncratic error:

$$t_i^{a,*} = \frac{\beta + \gamma - \Delta b_i}{2\beta + \gamma} + \lambda \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(\mu, \sigma_\epsilon^2), \tag{B.10}$$

where $\lambda = \frac{\beta+\gamma}{2\beta+\gamma}$.
We use standard maximum likelihood routines to estimate the model by maximizing the sum of individual log-likelihood contributions:

$$\begin{aligned} L_i = \mathbf{1}_{0<t_i<1} \cdot \frac{1}{\lambda\sigma_\epsilon} \cdot \phi\left(\frac{t_i - t_i^{**} - \lambda\mu}{\lambda\sigma_\epsilon}\right) \\ + \mathbf{1}_{t_i=1} \cdot \Phi\left(\frac{t_i^{**} + \lambda\mu - 1}{\lambda\sigma_\epsilon}\right) \\ + \mathbf{1}_{t_i=0} \cdot \Phi\left(\frac{-t_i^{**} - \lambda\mu}{\lambda\sigma_\epsilon}\right), \end{aligned} \tag{B.11}$$

where

$$t_i^{**} = \frac{\beta + \gamma - \Delta b_i}{2\beta + \gamma},$$

and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density and cumulative distribution functions, respectively. The first term in equation (B.11) corresponds to interior thresholds, while the second and third capture censoring at 1 and 0.

We estimate the model parameters taking Δb_i as given. The benefits index is the proxy for Δb_i. We use two approaches. First, we proxy Δb_i by the individual benefits index values. That is, the individual benefits index values replace Δb_i in (B.11). Second, we proxy Δb_i by the average benefits index value of individual i's REG group (i.e., each individual in a REG group is assigned the same value of Δb_i). This approach allows us to estimate individual parameters based on the variation in the benefits index between the REG groups, which is exogenous to a given individual's threshold choice.

The variation in Δb_i allows us to estimate β (social alignment). The social pressure, γ, is identified through the variation in threshold choices between the private and public conditions. Individual heterogeneity is accommodated through the error term with mean μ and standard deviation σ_ϵ. As discussed in the paper, we interpret

μ as forward-looking beliefs about how many others will follow an individual who takes action.

Table 5 of the paper shows the results of the estimation. In the paper, we discuss forward-looking beliefs. Here, we also consider another question: How substantial is the estimated conformity? Perceived individual and social benefits of AA, captured by variation in Δb_i, push thresholds toward the extremes of 0% and 100%. Social alignment, β_i, pulls thresholds toward the midpoint and dampens the effect of the perceived benefits of AA. Social pressure, γ_i, increases thresholds, particularly for people with low thresholds. The marginal change in thresholds in response to a change in Δb_i is

$$\frac{\partial t_i^*}{\partial \Delta b_i} = \frac{1}{2\beta_i + \gamma_i}. \tag{B.12}$$

Given the empirical estimates in Model (1) of Table 5, $\frac{\partial t_i^*}{\partial \Delta b_i} \approx 0.53$. For model (2), we obtain $\frac{\partial t_i^*}{\partial \Delta b_i} \approx 0.49$. These numbers imply that thresholds change by about half a point per percentage point change in Δb_i. Put differently, on average, the distance between the thresholds of two individuals with diametrically opposed views on affirmative action is 50% (i.e., conformity concerns prevent perceived benefits from shifting thresholds by 100%). In this sense, because conformity cuts in half the potential impact of perceived benefits, conformity and perceived benefits have equal weight in determining individual thresholds. Of course, the individual heterogeneity captured by the standard deviation of the error ($\hat{\sigma}$) allows for more varied thresholds, including at the extremes.

B.4 Details on simulated outcomes in Figure 6B

Simulation inputs and regional composition. Figure 6 in the main text reports the distribution of equilibrium AA support obtained from simulations of a threshold model calibrated separately for four U.S. regions. Thresholds are drawn from region-specific samples that reflect differences in racial and ethnic composition. In particular, the Midwest sample consists of 3.5% Asians, 9.9% Blacks, 8.6% Hispanics, and 78.0% Whites; the Northeast sample of 7.3% Asians, 11.5% Blacks, 15.3% Hispanics, and 65.9% Whites; the South sample of 3.7% Asians, 18.2% Blacks, 19.0% Hispanics, and 59.1% Whites; and the West sample of 10.6% Asians, 4.3% Blacks, 30.8% Hispanics,

and 54.3% Whites. These region-specific threshold samples serve as inputs to the simulations.

Simulation procedure. For each region, we simulate 10,000 societies of size 1,000. In each simulation, individual thresholds are drawn from the corresponding regional sample. Societies form sequentially: individuals are added one at a time, and after each addition all individuals compare the current share of AA supporters to their thresholds and update their choices accordingly. This process continues until no further changes occur, yielding an equilibrium level of AA support. After all 1,000 individuals have been added, we record the resulting equilibrium from each simulation and plot the distribution of equilibrium outcomes by region in Figure 6B.

B.5 Decision frames

The use of donations to elicit thresholds builds on a large literature that employs donation choices to incentive-compatibly measure support for socioeconomic causes (e.g., Bursztyn et al., 2020; Alesina et al., 2021; Fehérová et al., 2022). Existing studies differ in how they frame the underlying trade-off: whether respondents choose between multiple organizations, decide between supporting one organization or doing nothing, or incur a personal cost for supporting an organization. Here, we report results from additional treatments implementing these alternative framings. The overall pattern confirms that the threshold elicitation method captures stable underlying interdependent behavior, independent of the specific choice framing.

Because our threshold elicitation requires a binary action space, we implement three versions of binary donation choices that reflect different real-world decision frames in the 2025 data collection:

- **Competing Causes**: Respondents choose between supporting organization A or organization B. This is the frame used in the 2023 data collection and throughout the main part of our study. It has two key advantages. First, there are no income effects. Second, there is no ambiguity about where the money goes: it either supports cause A or cause B. This setup mirrors many real-world choices, such as voting or selecting a news source, where support for one option implies rejection of the other.

Table B.2: 2025 Data—Summary Statistics of Threshold Choices

	Average threshold		Share interior thresholds	
Reference Group	REG	U.S.	REG	U.S.
Competing Causes	40.30	36.86	66.40	65.60
Support vs. Neutral	43.82		66.67	
Personal Cost	43.93		56.54	

Notes: Threshold data from the 2025 data collection (Public condition). Respondents are White women, sampled to match the U.S. population in terms of age, region, and education. The number of observations is 250 per condition. Competing Causes: donate to cause A or to the opposing cause B. Support vs. Neutral: donate to cause A or take no action. Personal Cost: donate to cause A or keep the money.

- **Support vs. Neutral**: Respondents choose between supporting an organization (at no personal cost) or doing nothing. Like the A vs. B frame, this avoids income effects. However, it introduces uncertainty about what happens if the respondent remains neutral. The money stays with the experimenter and may be used for research, future data collection, or other purposes. This frame reflects real-life situations in which individuals can choose to remain silent rather than explicitly endorse or oppose a cause—for example, staying neutral in a political conversation.

- **Personal Cost**: Respondents choose between keeping a monetary amount or donating it to a cause. This frame introduces income effects. Unlike the other two frames, it blends preference-related motivations with financial incentives. While it complicates interpretation, it closely reflects everyday decisions in which supporting a cause comes with a personal cost.

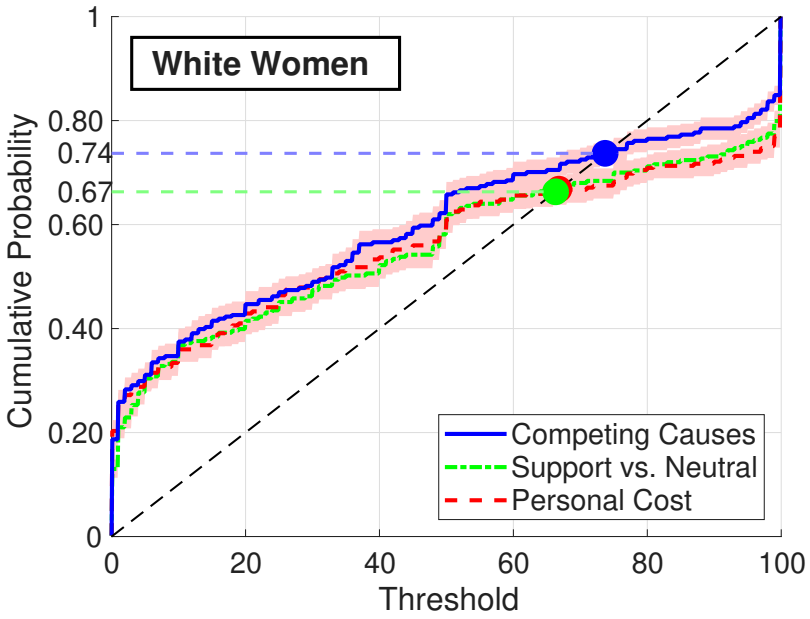
All three decision frames are useful. Our main design uses the Competing Causes (A vs. B) frame to isolate preference-based motivations and maximize experimental control. However, our threshold elicitation method can be applied to all three. To demonstrate this, we included all three decision frames in the 2025 data collection.

Results appear in Table B.2. Figure B.3 shows the corresponding threshold distributions. We find that, in the context of affirmative action, all three frames yield broadly similar distributions. Thresholds appear robust to moderate changes in the decision framing. Still, we observe small but consistent and intuitive shifts worth

noting.

First, average thresholds increase from 40.30 in the Competing Causes (A vs. B) condition to 43.82 in the Support vs. Neutral and to 43.93 in the Personal Cost conditions. While these difference in averages are not statistically significant ($p = 0.315$), Figure B.3 suggests that this shift comes from individuals with high thresholds. Specifically, there are fewer respondents with a threshold of 100 in the Competing Causes frame. When the alternative is to remain neutral or to keep the money, respondents less supportive of AA are more likely to refrain from donating even if most or all others do. Second, Figure B.3 and Table B.2 show a lower share of interior thresholds in the Personal Cost frame. This difference is statistically significant ($p = 0.018$). The presence of income effects may thus reduce conditional choices and weaken interdependence.

Figure B.3: Changing Thresholds—The Impact of Framing



Notes: Distribution of thresholds for AA (t^{AA}) for White women. The figure shows thresholds for different decision frames. Competing Causes: donate to cause A or to the opposing cause B. Support vs. Neutral: donate to cause A or take no action. Personal Cost: donate to cause A or keep the money.

Formally, changes in the decision frame affect the utility comparison between supporting change ($a_i = 1$) and the status quo ($a_i = 0$), primarily by shifting the perceived benefit of action, Δb_i; see Section 2. In the Competing Causes frame,

Δb_i reflects the value of supporting one cause over another, allowing both strong supporters and opponents of a cause to express their preferences symmetrically. In the Support vs. Neutral frame, however, individuals who oppose the cause no longer have an action that affirms their stance. Their only way to avoid supporting the cause is to set a high threshold. A threshold of 100 now means "neutral" whereas before it meant "oppose." This creates upward bunching: individuals who may have expressed moderate opposition through interior thresholds now shift toward the upper end of the distribution, not due to changes in underlying preferences but a less fine-grained way to express opposition. As a result, the threshold distribution becomes more asymmetric and more polarized at the upper end. This pattern is broadly consistent with our data, which show an increase in the share of maximum thresholds under the Neutral frame compared to the A vs. B frame.

The Personal Cost frame introduces a private cost to supporting change. In the model of Section 2, this reduces the net perceived benefit, such that $\Delta b_i^{\text{net}} = \Delta b_i - \text{cost}_i$. All else equal, this lowers the incentive to support the cause and raises optimal thresholds. The data offer some support for this (relative to the Competing Causes frame), though thresholds do not increase as much as the introduction of a personal cost might suggest.

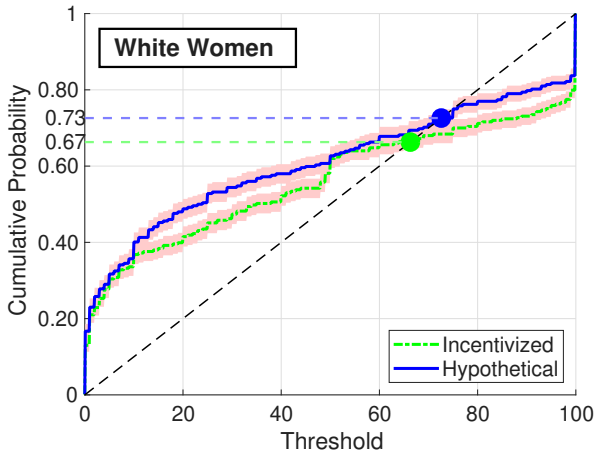
It is important to recall that we are measuring thresholds, not actions. Thresholds reflect interdependent beliefs. If a respondent expects few others to support the cause, then even with a low threshold, their donation will not be triggered, and the personal cost is not realized. In other words, choosing a threshold of 0 carries more weight when the action involves a personal monetary cost. It signals strong commitment and preferences. The stronger signal, in turn, may lead others to follow by choosing lower thresholds themselves, partly offsetting the first-order effect of monetary costs raising thresholds. Such indirect effects make thresholds hard to predict and require empirical measurement.

B.6 Testing for hypothetical bias

Knowing the extent to which elicited thresholds depend on monetary incentives is of importance. If results replicate without incentives, it facilitates the method's application across policy settings. The concern is hypothetical bias, a much-debated tendency of respondents to shift their answers in the absence of real consequences.

To test this, we implemented an additional treatment in which choice-dependent monetary incentives were removed. Participants faced the same threshold choice, still received the standard, flat participation fee, but were explicitly told that all donations linked to thresholds were hypothetical.

Figure B.4: Incentivized vs. Hypothetical Thresholds



Notes: Distribution of thresholds for affirmative action (t^{AA}) among White women. The sample comes from the 2025 data collection wave and includes 250 participants in an incentivized treatment and 250 in a hypothetical treatment, matched on demographic characteristics. In the incentivized condition, participants made real donation decisions to a pro–affirmative action cause (incentivized treatment). In the hypothetical condition, all monetary incentives were removed. Both treatments used the public framing, but in the hypothetical condition, participants were not exposed to the website where donations would otherwise be posted.

Figure B.4 compares the threshold distributions across the incentivized and hypothetical treatments. The two are statistically indistinguishable (Kolmogorov–Smirnov test, $p = 0.373$; mean difference, $p = 0.174$). The main determinants of thresholds—perceived benefits and conformity—remain strong and in the same direction. The overall pattern suggests that the elicitation method performs robustly even without monetary incentives. Researchers and policymakers may reasonably weigh the benefits of simplicity and scalability against the gains from providing fully incentivized settings.

B.7 Supplementary analysis for Section 4

Test of more nuanced model predictions for norm strength: The model predicts two effects of higher γ_i: (i) it raises thresholds because sanctions are only

incurred by change supporters; (ii) it pushes thresholds closer to 50 because incurred sanctions are proportional to the number of others who choose the other organization. The two effects both predict external pressure to increase thresholds for participants with optimal thresholds $t^* < 0.5$. The effects are countervailing for participants with $t^* > 0.5$. Effect (i) dominates (ii) as long as $t^* \leq 1$. Thus, in the paper, we test the primary effect (i). Here, we test effect (ii). It predicts that the effect of the norm strength variable on thresholds should be stronger for individuals who favor change than those who disfavor change according to their benefits index. Table B.3 confirms this prediction, as the coefficient is approximately double in size for individuals who favor change.

Table B.3: Heterogeneity of Norm Strength Effect

	$\Delta b_i > 0$	$\Delta b_i > 0$	$\Delta b_i > 0$	$\Delta b_i < 0$	$\Delta b_i < 0$	$\Delta b_i < 0$
Norm strength	0.213***	0.205***	0.204***	0.104***	0.111***	0.110***
	(0.033)	(0.033)	(0.033)	(0.038)	(0.038)	(0.038)
Constant	29.788***	32.147***	29.976***	52.180***	55.493***	51.160***
	(1.576)	(2.840)	(3.220)	(1.744)	(3.394)	(3.797)
Control Public			✓			✓
Controls REG & status quo		✓	✓		✓	✓
Observations	1,714	1,714	1,714	1,573	1,573	1,573
R^2	0.024	0.034	0.035	0.005	0.020	0.024

Notes: OLS regressions for individuals' thresholds $t^{AA}, t^{NoAA} \in (0, 100)$. Norm strength reflects participants' expectations about whether or not others are likely to confront someone speaking out in favor/against affirmative action policies $\in (0, 100)$. Columns (1) to (3) contain data of participants with a benefits index in favor of change, and columns (4) to (6) contain data of participants with a benefits index in favor of the status quo. Individuals with benefits index of 0 are omitted. Standard errors are depicted in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

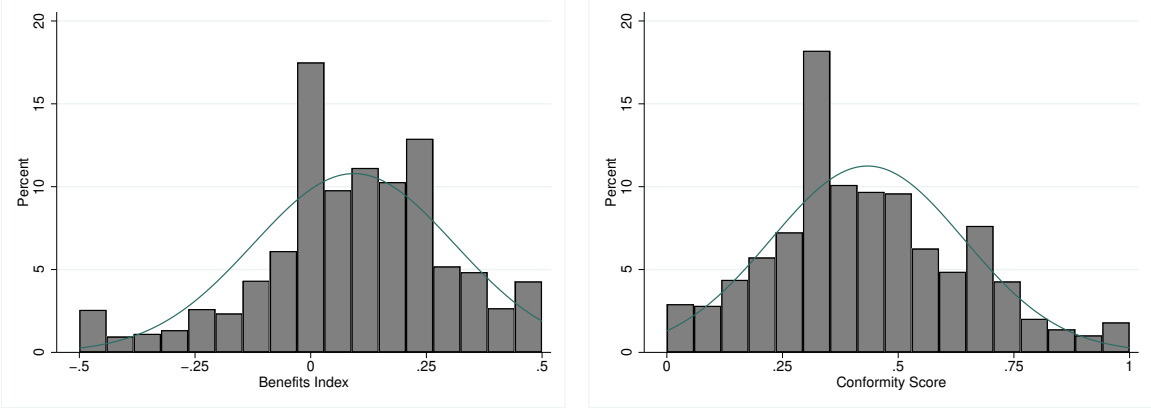
Expanding the analysis from Table 3: Table 3 presents the regressions supporting Hypotheses 1 and 3. Here, we investigate whether the results are driven by participants with interior thresholds ($t \in [1, 99]$), non-interior thresholds ($t \in \{0, 100\}$), or both. Tables B.4 and B.5 demonstrate that we obtain similar results for both groups of participants. For participants with interior thresholds, higher scores on the benefits index are associated with reduced thresholds for AA. Additionally, the public condition and perceived norm strength are found to increase thresholds, and Democrats exhibit lower thresholds for AA compared to Independents and Republicans. For those with non-interior thresholds, the same patterns emerge, although the

coefficient for norm strength is insignificant in some models. This finding is intuitive, as individuals with thresholds of 0 or 100 are less likely to be influenced by beliefs regarding norm strength.

Interiority of thresholds across REG groups controlling for the benefits index: The regression analyses in Table 4 columns (7) and (8) of the paper revealed differences in the interiority of thresholds between White men and most other REG groups. To see whether the benefits index Δb_i can explain this effect (i.e., White men having more extreme views), we report the regressions in Table B.6. The regressions show that the REG group differences prevail even when controlling for perceived benefits.

B.8 Benefits index, conformity index, Risk aversion and Norm strength

Figure B.5: Distribution of Benefits and Conformity Indices



Left: Benefits index. High value indicates a favorable view of affirmative action policies. Right: conformity index. High value indicates a preference for aligning one's behavior with the majority of others.

Individual items of the benefits index: The benefits index aggregates four items, each a five-point scale for agreement on the following statements: (i) affirmative action programs help decrease institutional injustice; (ii) affirmative action does more harm than good to minority groups; (iii) affirmative action is itself a form of discrimination; (iv) affirmative action enhances organizational performance in the long run. Table B.7 demonstrates that each question separately robustly replicates the effect of

Table B.4: Interior thresholds: Perceived Benefits and Social Pressure Shift Threshold Levels

	Both Defaults (1)	Both Defaults (2)	Both Defaults (3)	status quo Anti-AA (4)	status quo Pro-AA (5)	status quo Anti-AA (6)	status quo Pro-AA (7)
Perceived Benefits (Δb_i)							
Benefits index	-11.751***					-12.024***	12.632***
	(3.547)					(3.633)	(3.498)
status quo Pro-AA	-0.904						
	(1.078)						
Benefits index	25.789***						
× status quo Pro-AA	(4.911)						
Social Pressure (γ_i)							
Public		3.601***	4.283**			3.214**	2.993*
		(1.142)	(2.155)			(1.501)	(1.592)
Social Sanctions			29.008***			19.053***	17.250***
			(4.287)			(2.935)	(3.102)
Public			-2.772				
× Social Sanctions			(4.869)				
REG groups							
Asian/Female				-1.164	4.501	-2.625	-1.614
				(2.812)	(2.760)	(2.871)	(2.850)
Asian/Male				2.211	11.176***	1.217	8.181***
				(2.729)	(2.836)	(2.690)	(2.850)
Black/Female				-10.470***	-3.034	-6.912**	-4.354
				(2.669)	(2.719)	(2.767)	(2.828)
Black/Male				-9.788***	-0.958	-7.324***	-1.062
				(2.643)	(2.730)	(2.699)	(2.773)
Hispanic/Female				-0.452	6.877**	-6.038**	-2.095
				(2.837)	(2.920)	(2.786)	(2.835)
Hispanic/Male				-0.256	4.939*	0.671	3.826
				(2.611)	(2.576)	(2.590)	(2.626)
White/Female				-6.465**	-0.583	-8.044***	-1.667
				(2.924)	(2.886)	(2.930)	(2.883)
Democrat						-3.440**	0.401
						(1.527)	(1.757)
Republican						2.101	-0.265
						(1.980)	(1.985)
College						2.854**	6.316***
						(1.355)	(1.419)
Age						-0.069	-0.081
						(0.055)	(0.056)
Constant	45.376***	41.977***	30.418***	47.324***	42.623***	40.051***	34.185***
	(0.788)	(1.004)	(1.872)	(2.044)	(1.996)	(3.885)	(3.662)
Observations	6,088	6,088	6,088	3,131	2,957	2,935	2,750
Subjects	3,225	3,225	3,225	1,654	1,571	1,554	1,462

Notes: OLS regressions on interior thresholds ($t \in [1, 99]$) with s.e. clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data include up to two thresholds per individual (U.S. population and REG reference groups). The benefits index (normalized to -0.5 to 0.5) reflects perceived social benefits of AA policies. Social Sanctions are measured using participants' incentive-compatible expectations about whether others would confront them for speaking in favor of affirmative action (normalized between 0 and 1). Columns 4 and 6 report thresholds for supporting AA; columns 5 and 7 report thresholds for opposing AA. White men are the omitted REG group in columns 4–7. Independents and individuals without a college degree are the omitted categories in columns 6 and 7.

Table B.5: Extreme thresholds: Perceived Benefits and Social Pressure Shift Threshold Levels

	Both Defaults (1)	Both Defaults (2)	Both Defaults (3)	status quo Anti-AA (4)	status quo Pro-AA (5)	status quo Anti-AA (6)	status quo Pro-AA (7)
Perceived Benefits (Δb_i)							
Benefits index	-80.747***					-69.395***	83.491***
	(5.986)					(7.651)	(6.839)
status quo Pro-AA	5.415**						
	(2.690)						
Benefits index	169.987***						
× status quo Pro-AA	(8.249)						
Social Pressure (γ_i)							
Public		5.729	4.385			11.410**	3.093
		(3.949)	(7.102)			(5.073)	(4.759)
Social Sanctions			-8.497			5.233	-18.878**
			(15.012)			(8.184)	(7.625)
Public			4.254				
× Social Sanctions			(16.354)				
REG groups							
Asian/Female				-14.895*	31.970***	-7.958	12.402*
				(7.741)	(7.335)	(7.154)	(6.632)
Asian/Male				-11.663	25.445***	-4.287	13.864**
				(8.171)	(8.055)	(7.310)	(6.798)
Black/Female				-19.327**	27.344***	0.886	9.571
				(8.483)	(7.182)	(8.890)	(6.490)
Black/Male				-21.595***	26.805***	-3.839	5.554
				(7.684)	(7.501)	(6.894)	(6.935)
Hispanic/Female				-16.919**	25.634***	-4.246	8.467
				(8.162)	(8.152)	(7.083)	(8.394)
Hispanic/Male				-22.634**	14.075	-6.881	6.520
				(8.981)	(9.232)	(7.366)	(7.702)
White/Female				-10.103	27.693***	-3.808	13.134**
				(7.415)	(7.517)	(6.444)	(6.218)
Democrat						-9.916**	3.539
						(4.999)	(4.352)
Republican						1.850	-4.140
						(5.708)	(5.635)
College						2.411	5.194
						(4.097)	(3.764)
Age						0.320**	-0.351**
						(0.155)	(0.141)
Constant	51.814***	52.121***	55.093***	59.777***	44.366***	31.246***	64.581***
	(1.900)	(3.589)	(6.452)	(4.969)	(5.522)	(10.709)	(9.741)
Observations	1,884	1,884	1,884	939	945	901	874
Subjects	1,123	1,123	1,123	558	565	537	524

Notes: OLS regressions on extreme thresholds ($t \in \{0, 100\}$) with s.e. clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The data include up to two thresholds per individual (U.S. population and REG reference groups). The benefits index (normalized to -0.5 to 0.5) reflects perceived social benefits of AA policies. Social Sanctions are measured using participants' incentive-compatible expectations about whether others would confront them for speaking in favor of affirmative action (normalized between 0 and 1). Columns 4 and 6 report thresholds for supporting AA; columns 5 and 7 report thresholds for opposing AA. White men are the omitted REG group in columns 4–7. Independents and individuals without a college degree are the omitted categories in columns 6 and 7.

Table B.6: Threshold interiority, conformity and reference groups - Details

	(1) $t_i \notin \{0, 100\}$	(2) dist. to 0 or 100
Benefits index	0.019	1.819
	(0.034)	(1.200)
REG ref/ce group	0.045***	1.745***
	(0.005)	(0.232)
Conformity index	0.115***	2.648**
	(0.029)	(1.205)
Asian/Female	0.050*	1.461
	(0.027)	(1.062)
Asian/Male	0.111***	2.632**
	(0.026)	(1.033)
Black/Female	0.074***	0.676
	(0.027)	(1.051)
Black/Male	0.095***	1.588
	(0.026)	(1.029)
Hispanic/Female	0.109***	1.015
	(0.026)	(1.041)
Hispanic/Male	0.162***	6.282***
	(0.025)	(1.021)
White/Female	0.029	-1.294
	(0.027)	(1.018)
Constant	0.611***	15.334***
	(0.023)	(0.904)
Observations	7,972	7,972
Subjects	3,986	3,986

Notes: OLS regressions with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in (1) is whether or not a threshold is interior, $0 < t_i < 100$. The dependent variable in (2) is the distance from the extreme points, $\min(t_i, 100 - t_i)$. benefits index (normalized to -0.5 and 0.5) reflects an individual's perceived social benefits of AA policies. Ref. group (REG) is a dummy for whether group members share gender, or race/ethnicity, or both. White men are the omitted category.

perceived benefits on threshold choices, including when having all four statements in one regression model.

Creation of conformity index: Hong and Page (1989) created a 14-item *psychological reactance scale*. This scale is designed to measure the dimensions *(i)* Freedom of Choice, with the statements *I become angry when my freedom of choice is restricted; I become frustrated when I am unable to make free and independent decisions; I am contented only when I am acting of my own free will; The thought of being dependent on others aggravates me*, *(ii)* Conformity Reactance, with the statements

Table B.7: Agreement to statements in AA-questionnaire and threshold choice

	(1)	(2)	(3)	(4)	(5)
Q1: *AA programs help to decrease institutional injustice*	-26.947***				-18.997***
	(1.856)				(1.681)
Q: *AA does more harm than good to minority groups*		21.758***			11.564***
		(1.715)			(1.680)
Q: *AA is itself a form of discrimination*			19.692***		6.248***
			(1.667)		(1.648)
Q: *AA enhances organizational performance in the long run*				-23.519***	-5.210***
				(1.923)	(1.811)
Constant	61.314***	35.775***	36.325***	59.134***	52.212***
	(1.392)	(0.973)	(0.970)	(1.432)	(1.484)
Control for status quo	✓	✓	✓	✓	✓
Observations	8,172	8,172	8,172	8,172	8,172
Subjects	4,086	4,086	4,086	4,086	4,086
R^2	0.049	0.038	0.034	0.036	0.070

Notes: OLS regressions for individuals' thresholds $\in [0, 100]$. Thresholds normalized by status quo, such that a lower threshold makes a change towards the pro-AA organization more likely. The level of agreement to each statement is coded as $\in \{0, 0.25, 0.5, 0.75, 1\}$, with 0 as *completely disagree*, and 1 *completely agree*. Standard errors are clustered at the subject level and depicted in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

When something is prohibited, I usually think that's exactly what I am going to do; Regulations trigger a sense of resistance in me; I find contradicting others stimulating, *(iii)* Behavioral Freedom, with the statements *It disappoints me to see others submitting to society's standards and rules; When someone forces me to do something, I feel like doing the opposite; I resist the attempts of others to influence me; It makes me angry when another person is held up as a role model for me to follow*, and *(iv)* Reactance to Advice and Recommendations, with the statements *I consider advice from others to be an intrusion; Advice and recommendations usually induce me to do just the opposite; It irritates me when someone points out things which are obvious to me*. Hong and Faedda (1996) refine the scale to an 11-item scale by omitting the three statements *The thought of being dependent on others aggravates me; It disappoints me to see others submitting to society's standards and rules*, and *I am contented only when I am acting of my own free will*.

We elicit agreement with the 11-item scale suggested by Hong and Faedda (1996), using Hong and Page (1989)'s classification of dimensions. Table B.8 relates these dimensions to threshold choices by replicating the first two columns of Table 4 for each dimension. As we are interested in the drivers of interior thresholds, we aggregate the *dis*agreement levels for the dimensions *Freedom of Choice* and *Behavioral Freedom*. Thus, higher scores in these dimensions capture individuals' willingness to be influenced by others. The regressions in Table B.8 reveal that individuals with a lower concern for *Freedom of Choice* are more likely to have interior thresholds, while the *Behavioral Freedom* dimensions does not significantly affect whether thresholds are interior. The *conformity index* used in the article is created by aggregating the disagreement levels in the *Freedom of Choice* and *Behavioral Freedom*, as these dimensions capture interdependent behavior. For the dimensions *Conformity Reactance* and *Reactance to Advice and Recommendations*, higher agreement scores are interpreted as a tendency to choose actions contradicting societal norms (i.e., pro- or anti-AA in our context). In line with this interpretation, Table B.8 shows that, on average, participants who agree with the items under these dimensions have more interior thresholds; see Goldsmith et al. (2005) for a discussion of the relationship between psychological reactance and conformity.

Overview of indexes: Figure B.5 shows histograms of the benefits index and the conformity index. A higher value on the benefits index indicates a more favorable

Table B.8: Threshold interiority and (dis)agreement to dimensions of Psychological Reactance

	(1) $t_i \notin \{0, 100\}$	(2) dist. to 0 or 100	(3) $t_i \notin \{0, 100\}$	(4) dist. to 0 or 100	(5) $t_i \notin \{0, 100\}$	(6) dist. to 0 or 100	(7) $t_i \notin \{0, 100\}$	(8) dist. to 0 or 100
Freedom of Choice (disagreement)	0.175***	5.091***						
	(0.023)	(0.966)						
Conformity Reactance			0.166***	7.260***				
			(0.026)	(1.093)				
Behavioural Freedom (disagreement)					0.039	0.568		
					(0.026)	(1.079)		
Reactance to Advice and Recommendations							0.112***	5.117***
							(0.027)	(1.164)
Constant	0.703***	17.302***	0.654***	14.295***	0.745***	18.796***	0.689***	15.671***
	(0.011)	(0.428)	(0.019)	(0.764)	(0.014)	(0.593)	(0.020)	(0.818)
Observations	7,972	7,972	7,972	7,972	7,972	7,972	7,972	7,972
Subjects	3,986	3,986	3,986	3,986	3,986	3,986	3,986	3,986

Notes: OLS regressions with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in (1), (3), (5) and (7) is whether or not a threshold is interior, $0 < t_i < 100$. The dependent variable in (2), (4), (6) and (8) is the distance from the extreme points, $\min(t_i, 100 - t_i)$. The different dimensions are created using the extent of (dis)agreement to the statements in Hong and Page (1989) as described in the text, and normalized to lie between zero and one.

view of affirmative action policies. A higher value of the conformity index indicates a preference for aligning one's behavior with the majority of others. Table B.9 shows the averages of the two indices across REG groups. It also shows the average risk attitudes and perceived norm strength.

The averages of the benefits index range from 0.02 (White men) to 0.15 (Black women). These numbers indicate an average attitude in favor of AA. Further analysis shows that Asian and White men express a greater agreement than other groups with the statements that affirmative action policies may harm rather than help minority groups and that affirmative action policies represent a different form of discrimination. White women and men express less agreement than other groups with the statements that affirmative action decreases institutional injustice and enhances organizational performance in the long run. The averages of the conformity index range from 0.40 (White men) to 0.48 (Hispanic men). These numbers indicate that most people are moderate conformists, but some value independence.

Table B.9: Averages of elicited measures

	Asian F	Asian M	Black F	Black M	Hisp. F	Hisp. M	White F	White M	Dem.	Ind.	Rep.
Benefits index	.09	.06	.15	.14	.12	.11	.07	.02	.16	.06	-.03
	(.18)	(.20)	(.21)	(.20)	(.19)	(.19)	(.23)	(.28)	(.20)	(.22)	(.26)
Conformity index	.44	.39	.43	.43	.47	.48	.43	.40	.44	.43	.40
	(.20)	(.19)	(.21)	(.21)	(.217)	(.23)	(.20)	(.20)	(.21)	(.20)	(.21)
Risk aversion	.41	.33	.42	.32	.32	.29	.47	.36	.40	.35	.39
	(.26)	(.25)	(.27)	(.25)	(.26)	(.22)	(.26)	(.25)	(.26)	(.26)	(.26)
Norm strength	40.89	41.62	38.40	40.50	46.68	45.04	38.69	37.45	40.09	41.92	37.47
(status quo anti-AA)	(24.29)	(23.80)	(24.05)	(23.71)	(28.26)	(23.94)	(24.24)	(25.14)	(24.99)	(25.56)	(24.10)
Norm strength	40.57	40.75	38.36	36.33	49.67	39.15	35.33	36.18	35.91	41.61	37.25
(status quo pro-AA)	(26.57)	(23.47)	(24.21)	(25.58)	(29.30)	(23.98)	(22.38)	(23.99)	(22.83)	(27.00)	(24.04)

Notes: Benefits index $\in [-.5, .5]$ is constructed by aggregating and normalizing the agreement levels to the four questions (i) affirmative action programs help decrease institutional injustice; (ii) affirmative action does more harm than good to minority groups; (iii) affirmative action is itself a form of discrimination; (iv) affirmative action enhances organizational performance in the long run. Conformity index $\in [0, 1]$ is constructed by aggregating the answers to questions related to conformist behavior (Hong and Page, 1989; Goldsmith et al., 2005), with higher scores depicting more conformist behavior. Risk aversion $\in [0, 1]$ is elicited via the question of Dohmen et al. (2011). Norm strength $\in [0, 100]$ are the answers to the question *"How many in the group of 100 Americans do you think said that they would somewhat likely or very likely confront a person who publicly speaks in favor of affirmative action policies [against affirmative action policies] on the previous page?"*. The last three columns report the averages for a representative, weighted sample disaggregated by political affiliation (Democrats, Independents, and Republicans). Standard deviation in parentheses.

C Pre-registration

Our study was preregistered at the AEA Social Science Registry (AEARCTR-0010895) before any data was collected.

Sample: We planned to recruit 4,000 participants, 500 per race/ethnicity and gender (REG) group. The final sample consists of 4,086 participants, with 3,986 participants willing to share their race/ethnicity and gender. For each REG group, the sample includes between 484 and 507 participants. These numbers closely align with the pre-registration.

Hypotheses: Below we list the preregistered hypotheses and discuss where we address them in the paper.

Hypothesis 1—Correlation of thresholds with elicited preferences and beliefs

H1a Participants with a higher conformity index (measured via the conformity questionnaire) have a higher probability of interior thresholds and select thresholds closer to 50.

H1b Participants with a higher elicited intrinsic preference for affirmative action have lower thresholds if the status quo organization is anti-AA and higher thresholds if the status quo is pro-AA.

H1c Participants who expect high sanctions and are in the *Public* treatment (measured via the question *How many in the group of 100 do you think said that they would somewhat likely or very likely confront a person who publicly speaks in favor of [against] affirmative action on the previous page?*) have higher thresholds than participants who expect high sanctions and are in the *Private* treatment. Within the *Public* treatment, participants with low expected sanctions have higher thresholds than participants with high expected sanctions.

Hypothesis H1a is included in Hypothesis 2 of the paper (section 3.2). Hypothesis H1b refers to the benefits index and is included in Hypothesis 1 of the paper. Hypothesis H1c refers to norm strength and is included in Hypothesis 3 of the paper. Result 1, Result 2 and Result 3 provide evidence in favor of these hypotheses.

Hypothesis 2—Correlation of thresholds with observable characteristics

H2a Individuals belonging to groups that are more likely to benefit from affirmative action, on average and c.p., have a lower threshold if the status quo organization is anti-AA and higher thresholds if the status quo is pro-AA. For example, being female, Black, or Hispanic is expected to push thresholds toward more affirmative action through the intrinsic preference parameter.

H2b Supporters of the Democratic party have stronger intrinsic preference for affirmative action and therefore (c.p.) lower thresholds if the status quo organization is anti-AA and higher thresholds if the status quo is pro-AA compared to supporters of the Republican party.

H2c Participants who grew up in smaller towns have a higher conformity level and therefore (c.p.) have a higher probability of interior thresholds and select thresholds closer to 50.

Table 3 of the paper provides a test of Hypothesis H2a by demonstrating that individuals who belong to an underrepresented group have lower thresholds for AA (t^{AA}) and that the benefits index mediates the effect. Hypothesis H2b is tested in the same table, columns (6) and (7).

We test Hypothesis H2c in Table C.10 of this appendix. We find that the premise of the hypothesis is incorrect: in our data, growing up in smaller cities leads to a *lower* conformity index (on average). As a consequence, we find that growing up in smaller cities leads to fewer interior thresholds, rejecting the hypothesis. Put differently, Hypothesis H2c is a joint hypothesis of (i) the effect of smaller city size on conformity (the premise), and (ii) the effect of conformity on the interiority of thresholds (the test of the model). The data supports the model prediction (conformity effect on thresholds) but rejects the premise.

Hypothesis 3—Reference groups

H3a There are more interior thresholds for the narrower reference group (the second threshold question) than for the broader reference group (the first threshold question). The distance to threshold 50 is smaller for the narrower reference group (the second threshold question) than for the broader reference group (the first threshold question).

Table C.10: Conformity, city size, and thresholds

	(1) Conformity index	(2) $t_i \notin \{0, 100\}$	(3) dist. to 0 or 100
City size < 25k	-0.018**	-0.055***	-2.485***
	(0.008)	(0.017)	(0.650)
Constant	0.438***	0.774***	19.534***
	(0.004)	(0.007)	(0.281)
Observations	7,972	7,972	7,972
Subjects	3,986	3,986	3,986

Notes: OLS regressions with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in (1) is the conformity index. In (2) whether or not a threshold is interior, $t_i \notin \{0, 100\}$. The dependent variable in (3) is the distance to the extreme values, $min(t_i, 100 - t_i)$. City size < 25k is a dummy variable for whether the individual spent most of their childhood in a small city/town (population less than 25,000).

Table C.11: Conformity and reference groups

	(1) $t_i \notin \{0, 100\}$	(2) dist. to 0 or 100	(3) $t_i \notin \{0, 100\}$	(4) dist. to 0 or 100
Reference group	0.044***	1.694***	0.053***	1.957***
(gender or race)	(0.007)	(0.309)	(0.006)	(0.296)
Gender × Race	0.003	0.153		
	(0.014)	(0.588)		
Agree to statement			0.138***	4.648***
			(0.013)	(0.572)
Agree to statement			-0.021**	-0.563
× Pop. Ref. Group			(0.009)	(0.475)
Constant	0.741***	18.203***	0.690***	16.455***
	(0.007)	(0.282)	(0.009)	(0.358)
Observations	7,972	7,972	7,972	7,972
Subjects	3,986	3,986	3,986	3,986

Notes: OLS regressions with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in (1) and (3) is whether or not a threshold is interior, $t_i \notin \{0, 100\}$. The dependent variable in (2) and (4) is the distance to the extreme values, $min(t_i, 100 - t_i)$. Reference group (gender or race) is a dummy for whether group members share gender or race/ethnicity. Gender × Race captures the interaction effect, i.e., when groups share gender, race/ethnicity, and are similar in other dimensions. Agree to statement is a dummy variable that equals one if a participant chooses agree or strongly agree to *I am more likely to conform to the opinion of others who are [male/female]/[Asian/Black/Hispanic/White] than to the general US population.*.

H3b The effect of **H3a** is larger for the *Similar* treatment than the *Gender* or *Race* treatments.

H3c The effect of **H3a** is larger for the participants who are more in agreement with the statement that they are more conformist toward the given reference group than towards US society in general.

Hypothesis H3a is included in Hypothesis 2 of the paper. Result 3 supports the hypothesis. H3b states that sharing both race/ethnicity and gender with the reference group would result in even more interior thresholds than having a reference group that shares either race/ethnicity or gender alone. Table C.11 columns (1) and (2) provide the supporting analysis and show that this hypothesis cannot be supported. Interestingly, the results show that making a choice in the most narrow reference group (when group members are of the same race/ethnicity and gender) does not increase the interiority of thresholds compared to when they have in common only one of these characteristics; see the insignificant interaction term. Table C.11 columns (3) and (4) test H3c. The results show that stronger agreement with the statement *"I am more likely to conform to the opinion of others who are [male/female]/[Asian/Black/Hispanic/White] than to the general US population"* indeed increases the interiority of REG thresholds (thresholds chosen in the narrow reference group). However, the effect persists for the population threshold. The statement seems to measure conformity in general rather than conformity towards one's in-group specifically.

Hypothesis 4—Public versus private donations

H4 Thresholds in the *Public* treatment are higher than thresholds in the *Private* treatment.

Hypothesis H4 is included in Hypothesis 3 of the paper. Table 3 of the paper tests the hypothesis. Result 2 summarizes the evidence in favor of the hypothesis.

Hypothesis 5—Risk aversion

H5a Risk aversion increases threshold choices in the *Public* treatment compared to the *Private* treatment.

H5b Higher expected sanctions (measured via the question *How many in the group of 100 do you think said that they would somewhat likely or very likely confront a person who publicly speaks in favor of [against] affirmative action on the previous page?*) lead to a stronger increase in thresholds for more risk averse participants.

Table C.12: Risk aversion and threshold choice

	(1)	(2)	(3)
Public	4.575***	1.682	
	(1.266)	(1.752)	
Risk averse		-9.732***	0.634
		(2.207)	(2.087)
Public		4.987**	
× Risk averse		(2.516)	
Norm strength			0.222***
			(0.028)
Norm strength			-0.134***
× Risk averse			(0.045)
Constant	44.016***	49.006***	40.615***
	(1.116)	(1.572)	(1.433)
Observations	7,972	7,972	7,972
Subjects	3,986	3,986	3,986

Notes: OLS regressions on thresholds ($t^{AA}, t^{NoAA} \in \{0, 1, ...100\}$). Data includes two thresholds per individual (population and REG threshold). Public represents the dummy variable for being in the treatment where the individual decision will be posted on our website. Risk averse is a dummy variable with a median split among the above- and below- median answers to the question *"How do you see yourself: Are you a person who is generally willing to take risks, or do you try to avoid taking risks?"*. Standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

H5a predicts that the effect of the Public condition (see H4) is driven by risk-averse subjects. Table C.12 Column (2) shows that the interaction of Public with Risk averse is indeed large and significant. Hypothesis H5b is tested in Column (3) of Table C.12. The hypothesis can be rejected.[15]

[15]Notice that we had no hypothesis about the direct effect of risk attitudes on threshold choices (the hypotheses concern the *interaction* with the Public condition). Table C.12 shows that risk aversion decreases thresholds in the private condition. An ex-post rationale for this effect is that risk attitudes correlate with ambiguity attitudes (e.g., My et al., 2024), and ambiguity-averse participants can reduce uncertainty about their donation outcomes when choosing lower thresholds.

D Experimental Materials

D.1 Instructions

The exact wording of the study reads as follows.

Consent for Participation in a Research Study

Welcome!

We are researchers from New York University and The University of Texas at Dallas. This study is about social attitudes in the United States. Participation takes about **15 minutes**.

Your time and effort are greatly appreciated. In addition to the participation fee, you may receive **bonus earnings** in the form of **Amazon vouchers**. There are 14 questions where you have to enter a guess. For each question you guess accurately, you will enter a draw for one of **98 Amazon vouchers worth $50 each**. All information provided is 100% accurate, and your bonus earnings will be determined precisely as described.

There are no risks to participation beyond those of everyday life. Your participation is voluntary. You may stop participating at any time. However, if you choose to do so, you will not be able to restart the study and will not receive any compensation.

For questions about your rights, you may contact the Institutional Review Boards of New York University at IRBnyuad@nyu.edu or The University of Texas at Dallas at (972) 883-4575. For questions about this research, you may contact Dr. Moritz Janas at mmj9701@nyu.edu.

In this survey, some tasks and questions will be about affirmative action, and some choices include a donation decision. In addition, some questions will be about ethnicity, religion and political opinions, and a "Prefer not to answer" option will be available. To receive the bonus earnings, you will also be asked to verify your email address. The latter may be temporarily posted on a public website in a way that

humans can read it, but automated computer programs cannot. Whether or not this happens will be entirely under your control. This information will be discarded after 6 months. Your survey answers will be combined with the answers from other participants in academic publications and presentations such that your anonymity is preserved.

Are you 18 years or older, understand the statements above, and accept to participate in the research survey?
[*Yes, proceed to study*; *No, I want to leave the study*]

page break

Please answer the questions below.

Which U.S. state do you currently live in?
[*drop-down menu with all US states and oversea territories*]

How would you describe the place where you **currently** live?
[*city with more than 500'000 inhabitants*; *city with 100'000 to 500'000 inhabitants*; *city with 50'000 to 100'000 inhabitants*; *city/town with 25'000 to 50'000 inhabitants*; *city/town/village with 10'000 to 25'000 inhabitants*; *city/town/village with less than 10'000 inhabitants*]

How would you describe the place where you spent **most of your childhood** (age 0 to 18)?
[*city with more than 500'000 inhabitants*; *city with 100'000 to 500'000 inhabitants*; *city with 50'000 to 100'000 inhabitants*; *city/town with 25'000 to 50'000 inhabitants*; *city/town/village with 10'000 to 25'000 inhabitants*; *city/town/village with less than 10'000 inhabitants*]

page break

Please answer the questions below.

In what year were you born?
[*Open cell, 1900-2010*]

Are you ...
[*Male*; *Female*; *other*]

page break

Please answer the question below.

What racial or ethnic group best describes you?
[*Asian or Asian American*; *Black or African American*; *Hispanic or Latino*; *Native American or Alaskan Native*; *White or Caucasian*; *Middle Eastern*; *Other*; *prefer not to answer*;]

page break

Please answer the questions below.

What is the highest level of education you have completed?
[*No formal schooling*; *Primary school*; *Secondary school (High school)*; *Technical/vocational training*; *University degree (Bachelor)*; *Postgraduate (Masters, Ph.D.)*]

What is the highest level of education completed by **your father**? If you are not sure, please provide your best guess.
[*No formal schooling*; *Primary school*; *Secondary school (High school)*; *Technical/vocational training*; *University degree (Bachelor)*; *Postgraduate (Masters, Ph.D.)*; *not applicable, e.g. I did not know my father*]

What is the highest level of education completed by **your mother**? If you are not sure, please provide your best guess.
[*No formal schooling*; *Primary school*; *Secondary school (High school)*; *Technical/vocational training*; *University degree (Bachelor)*; *Postgraduate (Masters, Ph.D.)*; *not applicable, e.g. I did not know my mother*]

page break

Please answer the question below.

As of today, do you consider yourself to be a Democrat, a Republican, or an Independent?
[*Strongly Democrat*; *Democrat*; *Leaning Democrat*; *Independent*; *Leaning Republican*; *Republican*; *Strongly Republican*; *prefer not to answer*]

page break

Among the **ten people** you met most recently, **that are outside your family and with whom you exchanged opinions**, how many were male and female?

Male [*Open cell, 0-10*]
Female [*Open cell, 0-10*]
Other [*Open cell, 0-10*]

page break

Among the **ten people** you met most recently, **that are outside your family and with whom you exchanged opinions**, how many do you think self-identify as **Republican** or **Democrat**? (Please provide your best guess.)

Republican [*Open cell, 0-10*]
Democrat [*Open cell, 0-10*]
Other [*Open cell, 0-10*]

page break

Please provide your best guess: among the **ten people** you met most recently, **that are outside your family and with whom you exchanged opinions**, how many do you think ...
Identify as [*same racial/ethnic group*][16]
[*Open cell, 0-10*]

Do not identify as [*same racial/ethnic group*]
[*Open cell, 0-10*]

[16]Depending on the selection of the participant in the earlier question, this question contains one of the following: [*Asian or Asian American*, *Black or African American*, *Hispanic or Latino*, *Native American or Alaskan Native*, *White or Caucasian*, *Middle Eastern*]. This question did not appear to Participants who chose *other* or *prefer not to answer* in the racial/ethnic group question.

page break

Group attitudes in the U.S. population

You are one of 4,000 participants in this study, aged 21-65.

For this part of the study, we will assign you to **a group of 100 people living in the U.S.** (you and 99 others). Group members **represent the population** of the U.S. That is, different genders, races, and age groups are selected proportionally to their share in the U.S. population.

Note that this is the first of two parts in this study. In both parts, you will make a decision that affects donations to organizations. At the end of the study, one of the two parts will be randomly selected to determine your donation.

page break

Organizations and donations

This study examines attitudes toward affirmative action policies in the U.S.

To begin, we will randomly select one of the following two organizations:

- The American Association for Access, Equity and Diversity (AAAED) is a **PRO-affirmative action organization**. It fights for workplaces with equal representation of groups that were discriminated against or overlooked in the past (http://www.aaaed.org/).
- The American Civil Rights Institute (ACRI) is an **ANTI-affirmative action organization**. It fights against hiring procedures that allow for the preferred treatment of different groups based on gender, race, etc. (http://www.acri.org/).

By status quo, **we will donate $1 per person** in your group to the randomly selected organization. You will have the opportunity to change your $1 donation from the status quo organization to the other organization if you so desire.

page break

[This screen is only shown to participants in treatment Public.]

Changing your donation

Donations will be posted on **a public website** (https://www.howpeoplethinkabout.org/AffirmativeAction). The URL will be shared with all study participants and on social media.

Specifically, on the website, we will **post the email addresses** of all participants that changed their donation away from the status quo organization to the other. We will upload the addresses as pictures such that they can be read by humans, but cannot be copied by computer algorithms. Further, we will delete the addresses from the website within 6 months.

Your email address will be **publicly posted only if you change your donation** away from the status quo organization to the other. (Note: we will post the email address to which we sent you the invitation to this survey.) Your email address **will not be posted on the website if you don't change your donation**. Whether or not your email address will be posted on the website is, therefore, entirely under your control.

page break

[This screen is only shown to participants with the Anti-AA organization as status quo.

ANTI Affirmative action

The computer program randomly selected the **ANTI-affirmative action organization** American Civil Rights Institute (ACRI). Thus, by status quo, we will donate $1 per person in your group to this organization.

On the next page, **you will have the opportunity to change your donation** to the PRO-affirmative action organization.

page break

[This screen is only shown to participants with the Pro-AA organization as status quo.]

PRO Affirmative action

The computer program randomly selected the **PRO-affirmative action organization** American Association for Access, Equity and Diversity (AAAED). Thus, by status quo, we will donate $1 per person in your group to this organization.

On the next page, **you will have the opportunity to change your donation** to the ANTI-affirmative action organization.

page break[17]

Determining your donation

Please read carefully

To determine whether you change your $1 donation (from the anti-affirmative action to the pro-affirmative action organization), you must **choose a number between 0 and 100**. The number indicates **how many others in your group have to change their donation to the pro-affirmative action organization** such that you do so too.

Your donation can depend on the choices of others. Specifically, **if you choose a number between 1 and 99**, your donation will depend on the numbers chosen by the other people. For example:

- **If you choose 1**: your donation will change if 1 or more of the other people donate to the pro-affirmative action organization.
- **If you choose 79**[18]: your donation will change if 79 or more of the other people donate to the pro-affirmative action organization.
- ...and so on...

[17] the instructions on this and the subsequent pages are shown for the participants with the Anti-AA organization as status quo. For the instructions to participants with the Pro-AA organization as status quo every *pro-AA* is replaced with *anti-AA* and vice versa.

[18] This number is a randomly generated number between 2 and 99.

If you choose either 0 or 100, your donation will not depend on others' choices. Specifically:

- **If you choose 0**: you donate to the pro-affirmative action organization, even if no one else does.
- **If you choose 100**: you donate to the anti-affirmative action organization, even if no one else does.

Note that by choosing a lower number, you are more likely to change your donation. By extension, you are increasing the likelihood that others change their donation to the pro-affirmative action organization too.

page break

Your response

I will change my donation to the pro-affirmative action organization **if __ or more** of the other 99 Americans in my group do the same.[19]

[After selecting a number on the slider, the following bullet points appear below the slider.]

[When selecting 0] More precisely, when choosing 0:

- *you definitely donate to the pro-affirmative action organization.*
- *you increase everyone else's likelihood to donate the pro-affirmative action organization.*
- *your email address will be posted on the website.*[20]

[19] After selecting a number on the slider, this number appears in this sentence. If one selects 0, this sentence changes to "I will change my donation to the pro-affirmative action organization **even if none of the other** 99 Americans in my group do the same." If one selects 100, this sentence changes to "I will not change my donation to the pro-affirmative action organization **even if all other** 99 Americans in my group change their donation to the pro-affirmative action organization." Figures D.6, D.7, and D.8 provide visual representations of this screen.

[20] Only shown to participants in the Public treatment.

[When selecting a number between 1 and 99] More precisely, when choosing [number]:

- *you donate to the pro-affirmative action organization only if at least one other person chooses 0, at least two other people choose 0 or 1, at least three other people choose 0, 1 or 2, and so on up to the requirement that at least [number] other people choose a number below [number].*
- *you increase the likelihood to donate to the pro-affirmative action organization of others who choose a number above [number].*
- *your email address will be posted on the website if you donate to the pro-affirmative action organization.*[21]

[When selecting 100] More precisely, when choosing 100:

- *you definitely donate to the anti-affirmative action organization.*
- *you do not increase anyone else's likelihood to donate the organization.*
- *your email address will not be posted on the website.*[22]

..... *popup* .....

Are you sure?

You selected that you change your donation to the pro-affirmative action organization if $[X]$ or more of the other 99 Americans in your group do the same.
[*Return to slider*; *Submit*]

page break

Guess well to earn bonus

Out of the other 99 Americans in your group, please guess how many will ultimately change their donation to the pro-affirmative action organization.

[21] Only shown to participants in the Public treatment.
[22] Only shown to participants in the Public treatment.

[*Open cell, 0-99*]

If the actual number of people ultimately donating to the pro-affirmative action organization is between [*X-5*] and [*X+5*], you will enter a draw for one of 98 Amazon vouchers worth $50 each.

page break

Guess well to earn up to four more vouchers

Out of the other 99 Americans in your group, please guess how many...

... chose a number between **"0" and "20"** on the slider.
[*Open cell, 0-100*]

... chose a number between **"21" and "50"** on the slider.
[*Open cell, 0-100*]

... chose a number between **"51" and "80"** on the slider.
[*Open cell, 0-100*]

... chose a number between **"81" and "100"** on the slider.
[*automatically filled out s.t. numbers add to 99*]

For each answer, if the actual number of people is within "5" of your guess, you will enter a draw for one of 98 amazon vouchers worth $50 each.

page break

[Participants in the treatment with reference group race/ethnicity]

Group Change! 100 [Asians or Asian Americans] [Blacks or African Americans] [Hispanics or Latinos] [Native Americans or Alaskan Natives] [Whites or Caucasians] [Middle Easterners]

We will now place you in a group of **100 [Asians or Asian Americans] [Blacks or**

African Americans] [Hispanics or Latinos] [Native Americans or Alaskan Natives] [Whites or Caucasians] [Middle Easterners] living in the U.S. (you and 99 others). Group members represent the typical population of [Asians or Asian Americans] [Blacks or African Americans] [Hispanics or Latinos] [Native Americans or Alaskan Natives] [Whites or Caucasians] [Middle Easterners] living in the U.S.
We will ask you one final time the question we asked you before, but this time, for the group of 100 [Asians or Asian Americans] [Blacks or African Americans] [Hispanics or Latinos] [Native Americans or Alaskan Natives] [Whites or Caucasians] [Middle Easterners].

[Participants in the treatment with reference group gender]
Group Change! 100 [Men] [Women]

We will now place you in a group of **100 [men] [women] living in the U.S.** (you and 99 others). Group members represent the typical population of [men] [women] living in the U.S.
We will ask you one final time the question we asked you before, but this time, for the group of 100 [men] [women].

[Participants in the treatment with reference group similar-to-you]
Group Change! 100 people similar to you

We will now place you in a group of **100 people living in the U.S. that are similar to you** (you and 99 others). This group consists of people of the same gender, similar age group, same ethnical background, living in the same region in the U.S., and having a similar level of education.
We will ask you one final time the question we asked you before, but this time, for the group of 100 people similar to you.

page break[23]

100 [men] [women]

[23] the instructions on this and the subsequent pages are shown for the participants with the gender reference group. For the instructions to participants with the race/ethnicity or similar-to-you reference group, the wording referring to the group changes accordingly.

Recall that the computer program randomly selected the anti-affirmative action organization as the status quo. Because you are now in a new group where everyone is [male] [female], the number you enter to determine whether or not you change your donation may differ from before.

I will change my donation to the pro-affirmative action organization **if** ___ **or more** of the other 99 [men] [women] living in the U.S. in my group do the same.

[After selecting a number on the slider, the following bullet points appear below the slider.]

[When selecting 0] More precisely, when choosing 0:

- *you definitely donate to the pro-affirmative action organization.*
- *you increase everyone else's likelihood to donate the pro-affirmative action organization.*
- *your email address will be posted on the website.*[24]

[When selecting a number between 1 and 99] More precisely, when choosing [number]:

- *you donate to the pro-affirmative action organization only if at least one other person chooses 0, at least two other people choose 0 or 1, at least three other people choose 0, 1 or 2, and so on up to the requirement that at least [number] other people choose a number below [number].*
- *you increase the likelihood to donate to the pro-affirmative action organization of others who choose a number above [number].*
- *your email address will be posted on the website if you donate to the pro-affirmative action organization.*[25]

[24]Only shown to participants in the Public treatment.
[25]Only shown to participants in the Public treatment.

[When selecting 100] More precisely, when choosing 100:

- *you definitely donate to the anti-affirmative action organization.*
- *you do not increase anyone else's likelihood to donate the organization.*
- *your email address will not be posted on the website.*[26]

page break

Guess well to earn bonus

Out of the other 99 [men] [women] in your group, please guess how many will ultimately change their donation to the pro-affirmative action organization.
[*Open cell, 0-99*]

If the actual number of [men] [women] ultimately donating to the pro-affirmative action organization is between [*X-5*] and [*X+5*], you will enter a draw for one of 98 Amazon vouchers worth $50 each.

page break

Guess well to earn up to four more vouchers

Out of the other 99 [men] [women] in your group, please guess how many...

... chose a number **between "0" and "20"** on the previous screen.
[*Open cell, 0-100*]

... chose a number **between "21" and "50"** on the previous screen.
[*Open cell, 0-100*]

... chose a number **between "51" and "80"** on the previous screen.
[*Open cell, 0-100*]

[26] Only shown to participants in the Public treatment.

... chose a number **between "81" and "100"** on the slider.
[*automatically filled out s.t. numbers add to 100*]

For each answer, if the actual number of people is within "5" of your guess, you will enter a draw for one of 98 Amazon vouchers worth $50 each.

page break

Comprehension question: an opportunity for another voucher!

You will enter a draw for one of 98 vouchers worth $50 each **if you answer both questions correctly (you have one attempt only)**.

1. Is the following statement correct?
If someone chooses the number 0, they will change their donation irrespective of the choices of others in the group (that is, even if no one else changes their donation).
[*This statement is correct.*; *This statement is incorrect.*]

2. What happens if someone chooses the number 14?
[*They will change their donation irrespective of the choices of others in the group.*; *Whether they change their donation to the pro-affirmative action organization depends on the choices of others in the group.*; *They will not change their donation to the pro-affirmative action organization irrespective of the choices of others in the group.*]

page break

What do you prefer?
Please select your preferred option in each of the scenarios below. Choose carefully because we will implement your choice for one of the scenarios with positive probability.

Scenario A – please choose between
[*You receive $10 in vouchers, and we donate $10 to the status quo-AA organization.* ; *You receive no vouchers, and we donate $10 to the nondefault-AA organization.*]

Scenario B – please choose between

[*You receive $5 in vouchers, and we donate $10 to the status quo-AA organization.* ; *You receive no vouchers, and we donate $10 to the nondefault-AA organization.*]

Scenario C – please choose between
[*You receive no vouchers, and we donate $10 to the status quo-AA organization.* ; *You receive no vouchers, and we donate $10 to the nondefault-AA organization.*]

Scenario D – please choose between
[*You receive no vouchers, and we donate $10 to the status quo-AA organization.* ; *You receive $5 in vouchers, and we donate $10 to the nondefault-AA organization.*]

Scenario E – please choose between
[*You receive no vouchers, and we donate $10 to the status quo-AA organization.* ; *You receive $10 in vouchers, and we donate $10 to the nondefault-AA organization.*]

page break

People should speak in favor of affirmative action in public forums.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

People should speak against affirmative action in public forums.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

page break

Guess well to earn two more vouchers

How many in your first group, the group of 100 **Americans**, do you think said that they **agreed or strongly agreed** with the statement 'People should speak in favor of affirmative action' on the previous page?[27]
[*Open cell, 0-100*]

If the actual number of Americans is between [*X-5*] and [*X+5*], you will enter a draw

[27]Participants in with the pro-AA organization as status quo are asked about the statement 'People should speak against affirmative action'

for one of 98 Amazon vouchers worth $50 each.

How many in your second group, the group of 100 **[men]** **[women]**, do you think said that they **agreed or strongly agreed** with the statement 'People should speak in favor of affirmative action' on the previous page?[28]
[*Open cell, 0-100*]

If the actual number of [men] [women] is between [*X-5*] and [*X+5*], you will enter a draw for one of 98 Amazon vouchers worth $50 each.

page break

Confronting others

How likely would you be to confront a person who speaks out in favor of affirmative action policies?
[*Very unlikely*; *somewhat unlikely*; *neither likely nor unlikely*; *somewhat likely*; *very likely*]

How likely would you be to confront a person who speaks out against affirmative action policies?
[*Very unlikely*; *somewhat unlikely*; *neither likely nor unlikely*; *somewhat likely*; *very likely*]

page break

Guess well to earn two more vouchers

How many in the group of 100 **Americans** do you think said that they would **somewhat likely** or **very likely confront** a person who publicly speaks in favor of affirmative action on the previous page?[29]
[*Open cell, 0-100*]

[28]Participants in with the pro-AA organization as status quo are asked about the statement 'People should speak against affirmative action'

[29]Participants in with the pro-AA organization as status quo are asked about 'a person who speaks out against affirmative action policies'.

If the actual number of Americans is between [*X-5*] and [*X+5*], you will enter a draw for one of 98 Amazon vouchers worth $50 each.

How many in the group of 100 **[men]** **[women]** do you think said that they would **somewhat likely** or **very likely confront** a person who publicly speaks in favor of affirmative action on the previous page?[30]
[*Open cell, 0-100*]

If the actual number of [men] [women] is between [*X-5*] and [*X+5*], you will enter a draw for one of 98 Amazon vouchers worth $50 each.

page break

Views regarding affirmative action

Please indicate the extent of your agreement to the following statements. Your individual answers will never be shared or shown anywhere.

AA programs help to decrease institutional injustice.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

AA does more harm than good to minority groups
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

AA is itself a form of discrimination
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

AA (attention-check) please select Disagree here[31]
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

AA enhances organizational performance in the long run.

[30]Participants in with the pro-AA organization as status quo are asked about 'a person who speaks out against affirmative action policies'.

[31]participants who do not pass the attention check are screened out from the survey.

[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

page break

Taken all things into consideration, which statement would you say best describes your stance toward affirmative action?
[*I would publicly support it even if almost no one around me does*; *I would publicly support it only if at least some others around me do*; *I would publicly support it only if around half of those around me do*; *I would publicly support it only if most others around me do*; *I would not publicly support it even if almost all others around me do*]

page break

How would you donate $100?

Imagine you have one hundred dollars to donate in private to either the pro-affirmative action organization or the anti-affirmative action organization.

What amount would you donate to the pro-affirmative action organization?
[*Open cell, 0-100*]

Your selection implies that you would donate $ [*100-X*] to the anti-affirmative action organization.

page break

Abortion
In general, how does your willingness to publicly support women's access to abortion depend on the opinions of the people around you?
[*I would publicly support it even if almost no one around me does*; *I would publicly support it only if at least some others around me do*; *I would publicly support it only if around half of those around me do*; *I would publicly support it only if most others around me do*; *I would not publicly support it even if almost all others around me do*]

Migration
In general, how does your willingness to publicly support migration into the United

States depend on the opinions of the people around you?
[*I would publicly support it even if almost no one around me does*; *I would publicly support it only if at least some others around me do*; *I would publicly support it only if around half of those around me do*; *I would publicly support it only if most others around me do*; *I would not publicly support it even if almost all others around me do*]

Gun control
In general, how does your willingness to publicly support the right to bear firearms depend on the opinions of the people around you?
[*I would publicly support it even if almost no one around me does*; *I would publicly support it only if at least some others around me do*; *I would publicly support it only if around half of those around me do*; *I would publicly support it only if most others around me do*; *I would not publicly support it even if almost all others around me do*]

Working mothers
In general, how does your willingness to publicly support mothers of preschool children working full-time outside the home depend on the opinions of the people around you?
[*I would publicly support it even if almost no one around me does*; *I would publicly support it only if at least some others around me do*; *I would publicly support it only if around half of those around me do*; *I would publicly support it only if most others around me do*; *I would not publicly support it even if almost all others around me do*]

page break[32]

Please indicate the extent to which you agree to the following statements.

Regulations trigger a sense of resistance in me.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

I find contradicting others stimulating.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

When something is prohibited, I usually think "that's exactly what I am going to

[32]For 20% of the participants this screen and the second next screen do not appear here, but right before the page "Group attitudes in the U.S. population".

do."
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

I consider advice from others to be an intrusion.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

I become frustrated when I am unable to make free and independent decisions.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

It irritates me when someone points out things which are obvious to me.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

I become angry when my freedom of choice is restricted.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

Attention check - please select Disagree here.[33]
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

Advice and recommendations induce me to do just the opposite.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

I resist the attempts of others to influence me.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

It makes me angry when another person is held up as a model for me to follow.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

When someone forces me to do something, I feel like doing the opposite.
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

page break

To what extent do you agree to the following:
I am more likely to conform to the opinion of others who are [male] [female] than to

[33]participants who do not pass the attention check are screened out from the survey.

the general US population.[34]
[*Strongly Agree*; *Agree*; *Neither Agree nor Disagree*; *Disagree*; *Strongly Disagree*]

page break

Please answer the question below.

How do you see yourself: Are you a person who is generally willing to take risks, or do you try to avoid taking risks?
[*scale from 0 to 10, above 0 it says "not at all willing to take risks" and above 10 it says "very willing to take risks"*]

page break

Final questions ...

Religion
What is your religious affiliation – are you...
[*Protestant*; *Catholic*; *Mormon*; *Jewish*; *Muslim*; *Agnostic*; *Hindu*; *Buddhist*; *Christian Orthodox*; *Atheist*; *Another religion*; *Unaffiliated*; *prefer not to answer*]

page break

Social Media

On which of the social media platforms below are you active? Please select all that apply.
[*Facebook*; *Instagram*; *TruthSocial*; *Twitter*; *TikTok*; *LinkedIn*; *Snapchat*; *Reddit*; *Other*; *None of the above*]

page break

Thank you for participating!

In order to email you any Amazon vouchers that you won, we need to ensure that you are the rightful recipient.

[34]The wording of this questions depends on the reference group treatment participants are facing.

Please enter below **the same email address** to which we sent you the invitation to this survey.
Email:
[*open field*; *I prefer not to answer. I understand that this makes me ineligible for the bonus earnings (Amazon vouchers)*]

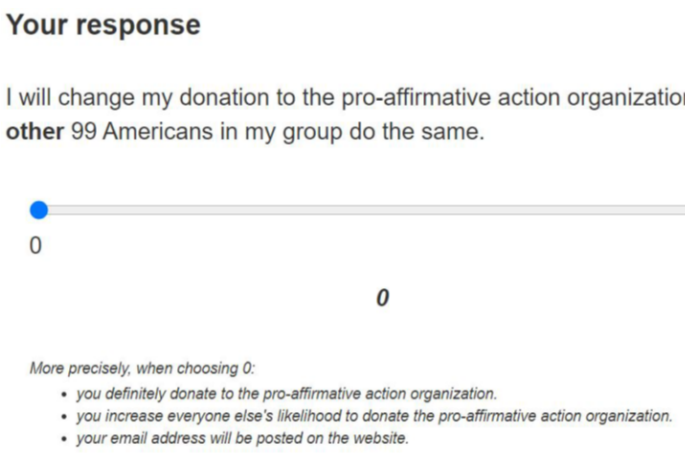
page break

End of the study

In the next days, we will make the donations to the two organizations. If you won any vouchers, we will contact you soon.

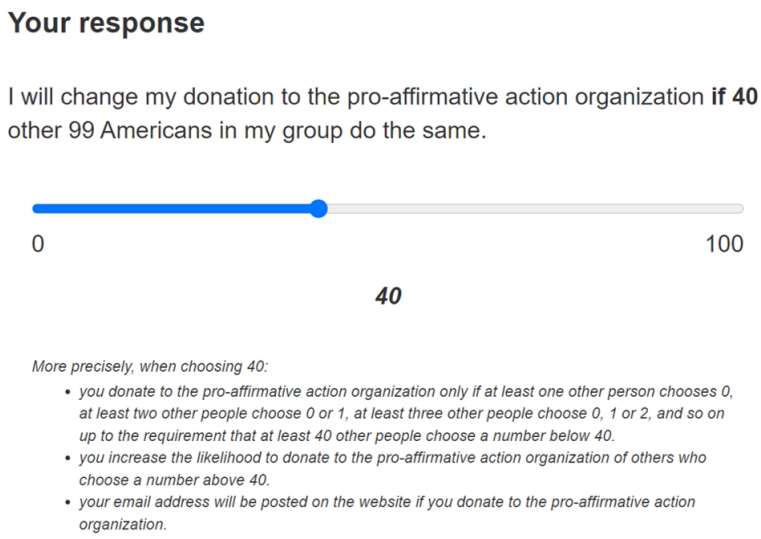
D.2 Screenshots of threshold question

Figure D.6: Screenshot of the threshold decision after participants selected zero.



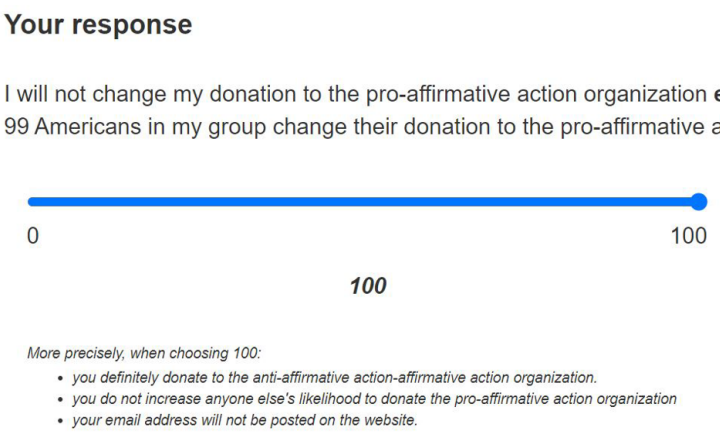
Notes: The sentence and the text below the slider updated in real-time depending on the current decision. The last bullet point was only shown for participants in the Public condition.

Figure D.7: Screenshot of the threshold decision after participants selected forty.



Notes: The sentence and the text below the slider updated in real-time depending on the current decision. The last bullet point was only shown for participants in the Public condition.

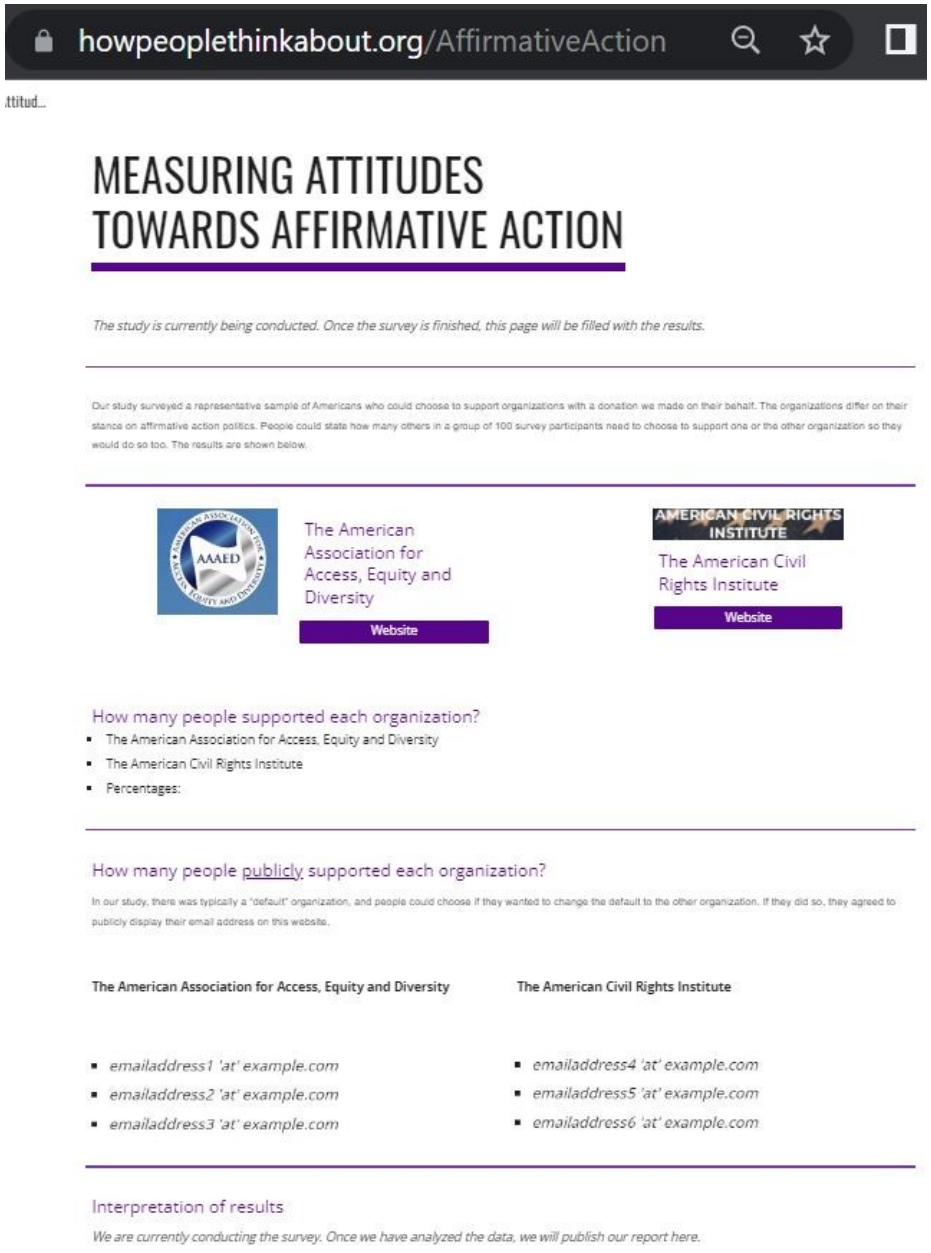
Figure D.8: Screenshot of the threshold decision after participants selected 100.



Notes: The sentence and the text below the slider updated in real-time depending on the current decision. The last bullet point was only shown for participants in the Public condition.

D.3 Website in Public Treatment

Figure D.9: Screenshot of the website during the time the experiment was conducted.



Notes: Participants were provided with the link to the website. After data collection, this website was updated showing the email addresses of the participants who deviated from the status quo in the Public treatment. In line with IRB requirements, the email addresses were removed after six months.