NBER WORKING PAPER SERIES

ELICITING THRESHOLDS FOR INTERDEPENDENT BEHAVIOR

Moritz Janas
Nikos Nikiforakis
Simon Siegenthaler

Working Paper 32847
http://www.nber.org/papers/w32847

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2024

We thank seminar participants at the University of Bologna, University of Cologne, Middlebury College, Norwegian Business School, New York University, University of Arizona, University of Texas at Dallas, University of Verona, and audiences at the 2024 Social Norms Workshop Ascona, 2024 Barcelona Summer Forum, the 2024 Dynamics of Social Change workshop at NYU Abu Dhabi, the 2023 World ESA conference in Lyon, and the 2023 ESA Africa conference in Cape Town. The project was pre-registered at AEARCTR-0010895 on March 03, 2023. IRB approval has been obtained by the NYUAD (HRPP-2022-74) and UT Dallas (IRB-22-582) Institutional Review Boards. MJ and NN gratefully acknowledge 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). The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Eliciting Thresholds for Interdependent Behavior
Moritz Janas, Nikos Nikiforakis, and Simon Siegenthaler
NBER Working Paper No. 32847
August 2024
JEL No. C83,C90,D63,D70

ABSTRACT

Threshold models have been widely used to analyze interdependent behavior, yet empirical research identifying people's thresholds is nonexistent. We introduce an incentivized method for eliciting thresholds and use it to study support for affirmative action in a large, stratified sample of the U.S. population. Most Asian, Black, Hispanic, and White men and women condition their support for affirmative action on the number of others supporting it. In line with preregistered hypotheses, thresholds are influenced by one's perceived benefits and pressure to conform. We demonstrate how our method can offer unique insights for policy design and enhance understanding of social dynamics.

Moritz Janas
New York University Abu Dhabi
Social Science Division
P.O. Box 129 188
Abu Dhabi
United Arab Emirates
moritz.janas@nyu.edu

Nikos Nikiforakis
New York University Abu Dhabi
Social Science Division
P.O. Box 129 188
Abu Dhabi, United Arab Emirates
nikos.nikiforakis@nyu.edu

Simon Siegenthaler
Jindal School of Management
University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080
simon.siegenthaler@utdallas.edu

1 Introduction

Our willingness to take an action often depends on the number of others who take the same action. From adhering to social norms to adopting new technologies, investing in specific stocks, participating in protests, and purchasing consumption goods, we frequently look at the actions of others when making decisions. Economists have long recognized that such interdependence can arise from the presence of network externalities (Katz and Shapiro, 1985, 1986), peer effects (Manski, 1993; Durlauf and Ioannides, 2010; Boucher et al., 2024), reputational concerns (Akerlof, 1980; Bernheim, 1994; Dvorak et al., 2024), and informational asymmetries (Banerjee, 1992; Bikhchandani et al., 1992; Dvorak and Fischbacher, 2024), while sociologists and psychologists have traditionally placed greater emphasis on the role of social norms (Coleman, 1990; Bicchieri, 2006, 2016) and personality traits such as attitudes toward conformity (Asch, 1952; Cialdini and Goldstein, 2004), respectively. Irrespective of its origins, understanding the social dynamics resulting from the interdependence is crucial for comprehending social and economic outcomes, anticipating change, and designing policies that promote welfare (e.g., Glaeser et al., 2003).

The idea that people condition their choices on the number of others taking the same action is central in a class of models with a long history in the social sciences: threshold models. Threshold models first gained prominence through the work of Mark Granovetter (Granovetter, 1978) and Thomas Schelling (Schelling, 1978). Since then, economists, philosophers, political scientists, and sociologists have used them to explore a diverse array of topics, including riots (Granovetter, 1978), racial segregation (Schelling, 1978; Clark, 1991; Zhang, 2011), consumption (Granovetter and Soong, 1986), public good provision (Oliver et al., 1985; Macy, 1991), policy adoption (Roland and Verdier, 1994; Simmons and Elkins, 2004), revolutions (Kuran, 1995), diffusion of innovation (Jackson and Yariv, 2007; Galeotti and Goyal, 2009; Young, 2009; Centola, 2015), networks (Jackson, 2008; Goyal, 2023), norm change (Bicchieri, 2016; Efferson et al., 2015, 2020; Andreoni et al., 2021), political polarization (Sunstein, 2018; Ehret et al., 2022), and climate action (Constantino et al., 2022; Berger et al., 2023).[1]

[1]Threshold models are sometimes referred to as models of social influence due to the central role of others' choices in one's decisions. Two alternative classes of models used for analyzing problems

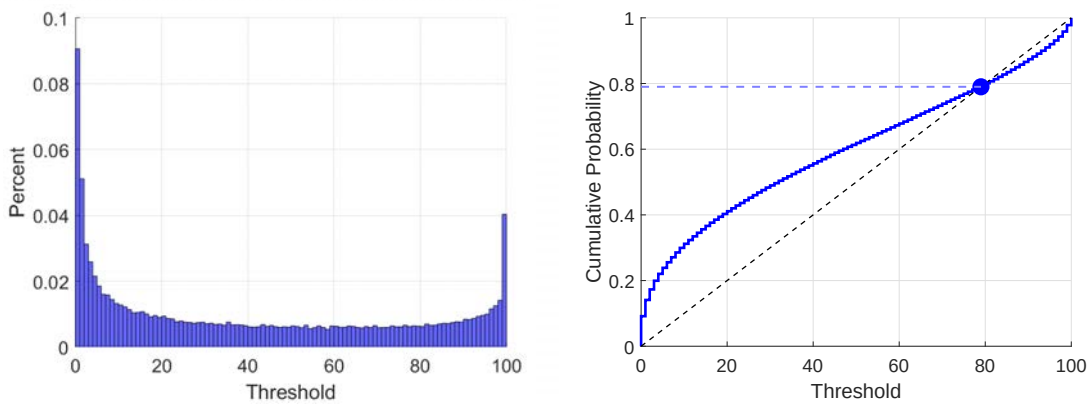
The widespread and enduring use of threshold models can be attributed to the straightforward manner in which they simplify the analysis of complex interdependence problems, without relying on unrealistic behavioral assumptions. In threshold models, each individual is characterized by a threshold, t_i^a, indicating the share of others that must take action a before individual i does the same. Individual thresholds, therefore, can range between 0 (unconditional "supporters" who will choose a even if none else in their group does so) and 100 (unconditional "opponents" who will not choose a even if everyone else in their group does). Between these extremes are individuals whose choice depends on the share of others selecting a. Individuals are assumed to rationally set their thresholds at that point where the perceived benefit to them from choosing a exceeds the perceived cost. Once individual i observes that the share of others choosing a exceeds her threshold, she too chooses a. Starting with a probability distribution of thresholds, threshold models predict the equilibrium number that will select a using best-response dynamics.

An example will help demonstrate how threshold models work. Assume you are invited to a faculty meeting where a policy for adopting affirmative action (AA) in hiring is presented. You are asked to show your support for the policy by a show of hands. You weigh the benefits and costs of the policy as well as the reputational consequences of publicly supporting (or not) the policy to decide whether or not you will condition your support on the support of others in the room. Others in the room of course perform similar calculations leading to the distribution of thresholds seen in Figure 1a. In this example, 9% have a threshold of 0 meaning that they will raise their hand even if zero others do, while 4% have a threshold of 100 meaning that they will not raise their hand even if everyone else does. The remaining 87% exhibits interdependent behavior. Granovetter (1978) shows that the equilibrium is given by the intersection of the cumulative distribution function (CDF) from above with the 45-degree line: so long as the CDF is above the 45-degree line, the best-response dynamic implies that more individuals will raise their hand and, in doing

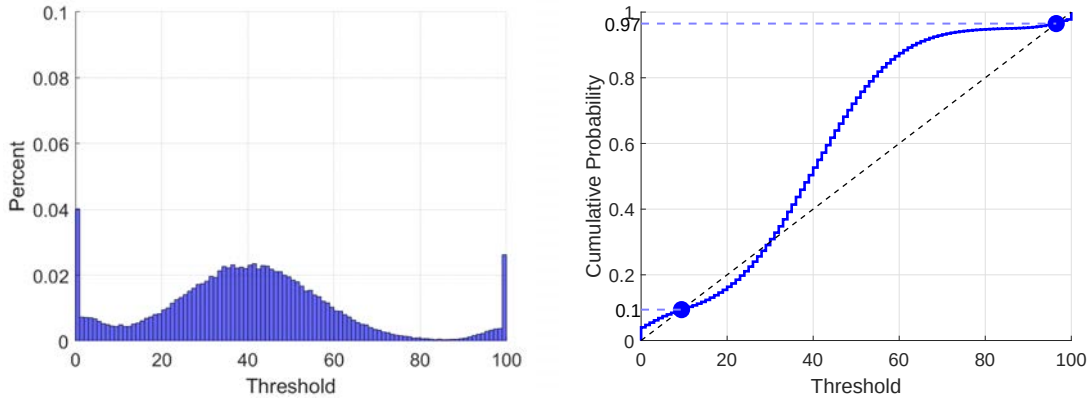
of strategic complementarities under complete information are models of social learning and social contagion. These models have similarities but also differences to models of social influence. For a detailed discussion of the three modeling approaches, see Young (2009). Threshold models also have similarities with global games – games of *incomplete* information with strategic complementarities where players face private signals about the underlying fundamentals (Carlsson and van Damme, 1993). They have been used most frequently to study bank runs, currency attacks, and financial crises.

so, encourage others to do the same. In this example, therefore, 78% of participants are predicted to raise their hand in support of the AA policy.

Figure 1: Examples of Threshold Distributions



(a) Unique Equilibrium



(b) Two Equilibria

Despite the widespread use of threshold models for nearly half a century in theoretical work, there remains a puzzling lack of empirical evidence supporting their application. Specifically, research identifying individual thresholds is nonexistent.[2] Only a handful of empirical studies exist showing that threshold models can accurately predict aggregate-level behavior in laboratory experiments (Centola et al.,

[2]Our paper is the first to elicit thresholds in the context of threshold models. Laboratory studies have shown that a majority of participants use threshold strategies in experimental tests of global games (Heinemann et al., 2004, 2009; Szkup and Trevino, 2020). Unlike in threshold models where individuals condition their choices on the share of others taking an action, in global games individuals condition their choices on the signal they receive about the state of the world.

2018; Andreoni et al., 2021; Ehret et al., 2022). However, these studies do not elicit individual thresholds but rather assume a distribution or induce them. The lack of empirical research on threshold distributions is surprising given the importance of distributions in threshold models.[3] Moreover, it implies that some of the most fundamental questions relating to threshold models remain unanswered: Do people have thresholds for interdependent behavior? What are the determinants of individual thresholds? The answers to these questions are essential to anticipate how policy interventions will affect social outcomes.

This paper takes a first step toward addressing this gap in our knowledge by presenting an incentivized method for eliciting individual thresholds for interdependent behavior. We use this method to elicit individual thresholds for supporting affirmative action in a large sample of the US population, stratified over race/ethnicity (Asian, Black, Hispanic, and White), and gender, and explore their determinants. To guide our empirical analysis, we construct a simple model assuming that individuals weigh the benefits of supporting affirmative action against the costs and use it to derive (and preregister) behavioral hypotheses. We elicit thresholds in different conditions which vary (*i*) one's reference group (the U.S. population vs. people from the same racial/ethnic/gender group), (*ii*) whether the individual's behavior is observed by others or not, and (*iii*) whether an individual is called to support or oppose AA. Additionally, we collect information on individuals' attitudes toward AA, conformity, and perceived norm strength, which we use in our analysis.

We find that a large majority of individuals – from 63.33% to 87.92% – across racial, ethnic, and gender (REG) groups and treatments, condition their support for affirmative action on the number of others who also support it. This finding is of significance because threshold models could be said to be falsified if all or most thresholds are at 0% or 100% such that people's behavior would not be contingent on that of others (Granovetter, 1978). As predicted by our model, thresholds are influenced by whether individuals stand to benefit from affirmative action (as proxied by their REG group), their general attitudes towards the fairness of AA, and the pres-

[3]As Granovetter (1978) points out in the abstract of his seminal paper: "Stress is placed on the importance of exact distributions for outcomes. Groups with similar average preferences may generate very different results" (p. 1420). To illustrate this point, Figure 1b presents a distribution with the same mean as the distribution in Figure 1a. Nevertheless, the model now admits two equilibria due to the mass of thresholds around 40: one equilibrium has 10% of individuals supporting AA while the other has 97%.

sure – internal and external – to conform. Specifically, in line with our hypotheses, we find that thresholds for supporting affirmative action decrease as an individual's perceived benefits from AA increase. We also find that thresholds are more likely to be interior (and are closer to 50) the narrower the individual's reference group is and the greater their sensitivity to conformity. Using a structural model, we find that, consistent with threshold models, individuals expect their support for AA to encourage others to follow suit.

The next section presents the threshold elicitation method, the experimental design, and a simple model from which we derive testable hypotheses. Section 3 presents the main empirical findings from our study. In section 4, we show how the elicited threshold distributions can be used to generate predictions in real-world settings and inform the design of policies. Finally, we conclude in section 5 with a discussion of future research.

2 The experiment

2.1 An incentivized method for eliciting thresholds

Each individual in threshold models is characterized by a threshold, t_i^a, indicating the share of others that must take action a before i does the same, with $t_i^a \in [0, 100]$, and $a_i \in \{0, 1\}$ where 1 indicates support for a. We start by introducing an incentivized method for eliciting t_i^a.

Let g_τ^a denote the share of supporters of a at time τ. Individual i chooses $a_i = 1$ if and only if $g_\tau^a \geq t_i^a$. That is, i will choose a at time $\tau+1$ if and only if i observes that the share of others selecting a at time τ meets i's threshold for support. Thresholds are heterogeneous, where F is the cumulative distribution function of thresholds. The dynamic is governed by $g_{\tau+1}^a = F(g_\tau^a)$ and $g_0^a = 0$. So, in the status quo, no one supports action a, and the share of supporters increases until an equilibrium $g^{a,*}$ is reached such that $g_\tau^a = F(g_\tau)$ (Granovetter, 1978).

The two-step method for eliciting individual thresholds is as follows. First, we assign each participant to a group of n individuals and we promise to donate $\$x$ for each group member to a charity *opposing* a. In other words, the default is that we will donate $\$n * x$ to a status quo organization ($a_i = 0$ for all i). Second, each group member i is asked to choose whether they would like to change their donation to

an organization *supporting* a by denoting the share of others who must support a before i does too, $t_i^a \in [0, 100]$.[4] To incentivize threshold choices, group member i changes their donation to $a_i = 1$ if and only if her threshold is equal to or smaller than $g^{a,*}$.

The way we use to incentivize choices is a defining feature of our method that distinguishes it from other "strategy methods" often found in the experimental economics literature (e.g., Mitzkewitz and Nagel, 1993; Brandts and Charness, 2011; Fischbacher and Gächter, 2010). Strategy methods ask participants to make conditional decisions for each possible information set, but their choices cannot affect the choices of other individuals. By contrast, our threshold elicitation method allows individuals to influence others' choices. Hence, participants can form forward-looking beliefs about how their choices will impact the likelihood that others support a. Our method is thus suitable for situations where individuals observe the fraction of others supporting change and switch their support when their threshold is reached, despite being uncertain about the eventual number of supporters.

2.2 Thresholds for supporting affirmative action

The method described in the previous section can be used to elicit thresholds in a wide range of situations including technology adoption, participating in collective action, adhering to social norms, and more. The key requirement is that the action space is binary. Here, we use it to study threshold distributions for supporting affirmative action policies in the United States.

There are three reasons for studying thresholds for supporting affirmative action (AA). First, AA has significant socioeconomic implications (Holzer and Neumark, 2000), which remain a topic of continuous discussion (Bleemer, 2022).[5] Second, individuals' perceived benefits from AA are likely to be influenced by a person's racial/ethnic/gender (REG) group. This provides natural exogenous variation, enabling us to test the predictions of our model. Third, the strong correlation between

[4] An alternative would be to give each individual an endowment which they could either keep or donate to the organization. Such design, however, would introduce income effects. Specifically, in a highly diverse sample such as ours, REG membership would be expected to be correlated both with income as well as attitudes toward AA.

[5] Data collection for this project was completed before the landmark ruling by the U.S. Supreme Court on June 29, 2023, which determined that race-based affirmative action programs in college admissions violate the Equal Protection Clause of the Fourteenth Amendment.

perceived benefits and one's REG group makes for a tough test of threshold models. The reason is that, as we discuss in the model section, thresholds are less likely to be interior when perceived benefits either for or against an action are substantial.

We recruit a large number of participants (see next section) and place them in groups of 100 individuals. For each individual in the group, we promise to donate \$1 to one of two organizations. The organizations are the American Association for Access, Equity, and Diversity (pro-affirmative action) and the American Civil Rights Institute (anti-affirmative action). For each group, one of the organizations is randomly selected as the status quo ($a_i = 0$) such that taking action ($a_i = 1$) represents changing one's allocation of money to the other organization. All participants receive information about the organizations' missions and can visit the official websites.

Each group member is asked to choose whether they would like to change their donation to the other organization by denoting the share of others who must support a before group member i does too, $t_i^a \in [0, 100]$. Choosing 100 guarantees supporting the status quo organization because there can be at most 99 others with $a_i = 1$. Choosing 0 guarantees supporting the alternative organization because an individual chooses $a_i = 1$ even if no one else does. An interior number allows participants to condition their donation on others' behavior. If $t_i^a \leq g^{a,*}$, i's donation is switched to organization a.[6]

2.3 The sample

The sample consists of 4,086 individuals. A roughly equal number of Asian, Black, Hispanic, and White men and women participated in the study. The reason for using a sample stratified over race, ethnicity and gender is that individuals' perceived benefits from affirmative action (AA) are expected to correlate with their racial, ethnic, and gender (REG) group membership: individuals from underrepresented groups are likely to perceive greater benefits from AA policies. Therefore, stratification over REG groups offers natural exogenous variation for testing the predictions

[6]Participants received clear instructions explaining the incentivization. We separately explained the consequences of choosing an interior number, a threshold of 0, and a threshold of 100. The median participants spent 117 seconds (25^{th} percentile: 64 sec; 75^{th} percentile: 200 sec) considering their response to the first threshold question. Attention checks were used to screen out inattentive participants.

Table 1: Sample

		Education		Age Group				Region			
REG Group	Total	No College	College	21-24	25-34	35-44	45-65	Mid-west	North-east	South	West
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	278	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
Total	4086	2036	2050	286	1014	1014	1772	664	721	1605	1095

Notes: REG refers to racial/ethnic/gender group. Education, region, and age quotas are derived from the 2021 American Community Survey (ACS). Participants who declined to state their race/ethnicity or identified as non-binary are listed as unassigned. No data was collected for American Indians, Alaska Natives, Native Hawaiians, and other Pacific Islanders (1.3% of the US population) as we would not have been able to obtain a sufficient number of observations to generate threshold distributions.

of our model effectively.[7] Sampling weights were used such that the sample for each REG group is representative of the population from which it was drawn on education, age, and geographical region. Quotas are derived from the American Community Survey. Table 1 provides an overview of the sample.

Data were collected by the survey company Ipsos, using an interface that was programmed in SophieLabs (Hendriks, 2012). The median completion time for the study was 14.27 minutes. Monetary incentives were designed to ensure participants were engaged and consisted of three components: the standard participation reward paid to Ipsos panel members (iSay points redeemable for various vouchers and gift cards including Amazon, Starbucks, and Walmart), the donations we made on the participants' behalf to the two non-governmental organizations to incentivize threshold choices ($4,086), and 98 Amazon vouchers ($4,900) to incentivize belief elicitation tasks.

[7]If we had opted for a random sample of the U.S. population instead, we would have limited power to identify differences in underrepresented REG groups.

2.4 Theoretical framework

We present a simple framework to derive behavioral hypotheses and guide the empirical analysis. We assume individual i's utility is given by

$$U_i(a_i) = v_i(a_i) - \beta_i C(a_i) - \gamma_i C(a_i) \mathbb{1}_{a_i=1} \tag{1}$$

where $a_i \in \{0, 1\}$ indicates the organization i supports, with $a_i = 0$ ($a_i = 1$) indicating that i supports the default (alternative) organization. The variable $v_i(a_i) \in \mathcal{R}$ captures the perceived benefits from each action, both personal benefits and social benefits. Social benefits depend on the congruence of an individual's values with the supported organization. The second and third terms in (1) represent non-conformity costs. Specifically, variable $C(a_i) = \frac{1}{n-1} \sum_{j \neq i} \mathbb{1}_{a_j \neq a_i}$ represents the fraction of others who choose an organization different from the one selected by i. Variable β_i captures i's personal preference for conforming to others. If $\beta_i > 0$, i suffers a disutility which increases with the number of those who support it. Variable γ_i captures the external pressure i feels to select the default organization. This pressure also increases with the number of others selecting the default organization. Such pressure can arise from fear of social sanctions or image concerns.[8]

Individual i's optimal threshold corresponds to the fraction of others supporting a such that $U_i(0) = U_i(1)$. The optimal threshold ensures that individual i chooses $a_i = 1$ as soon as enough others have abandoned the status quo given i's perceived benefits and conformity preferences. The optimal threshold is:

$$t_i^* = \frac{\beta_i + \gamma_i - \Delta v_i}{2\beta_i + \gamma_i} - \frac{\beta_i + \gamma_i}{2\beta_i + \gamma_i} m_i(t_i^*) \tag{2}$$

where $\Delta v_i \equiv v_i(1) - v_i(0) \in \mathcal{R}$ is the net perceived benefit from choosing $a_i = 1$, and $m_i(t_i^*) \geq 0$ captures i's expected marginal impact on the actions chosen by the others. To be precise, denote by $g^{a_i,*}(a_i)$ the fraction of others who choose $a_i = 1$ in equilibrium when individual i chooses $a_i = 0$ or $a_i = 1$. Individual i's marginal

[8] As can be seen in equation (1), we assume that only individuals who deviate from the default organization face external pressure. There are different reasons for this modeling choice. From a conceptual point of view, this asymmetry captures the observation that, in many instances, breaking with the status quo attracts social disapproval while continuing to follow the status quo behavior, at least for a while, is socially accepted. From a practical perspective, as we will see, this assumption fits better our experimental design.

impact on others is $m_i(t_i^*) \equiv g_i^{a,*}(1) - g_i^{a,*}(0) \geq 0$.

The marginal impact $m_i(t_i^*)$ depends on expectations in complex ways. It depends on the details of F and individual heterogeneity in risk preferences, optimism, etc. Furthermore, since we study transitions between equilibria in large groups, individuals face uncertainty about the threshold distribution, F. We thus model $m_i(t_i^*)$ in the form of an idiosyncratic error term:[9]

$$t_i^* = t_i^{**} + \epsilon_i, \quad \epsilon_i \sim N(\mu, \sigma_\epsilon^2) \tag{3}$$

where

$$t_i^{**} = \frac{\beta_i + \gamma_i - \Delta v_i}{2\beta_i + \gamma_i} \tag{4}$$

and $\mu = -\frac{\beta_i+\gamma_i}{2\beta_i+\gamma_i} m_i(t_i^*)$. Our data will allow us to estimate μ and σ_ϵ^2 together with the other parameters.

The comparative statistics are as follows. An individual's threshold decreases as Δv_i increases because of the greater perceived benefits from supporting action a. A greater β_i pushes thresholds inward, closer to one-half, because of the internal desire to conform to others. A greater γ_i increases thresholds because of the external pressure to conform when publicly challenging the status quo — the latter effect is strongest for individuals with low thresholds.[10] Finally, a negative μ indicates forward-looking beliefs rather than myopic behavior, with a greater absolute value of μ causing lower thresholds as individuals expect others to follow them in choosing $a_i = 1$. Note that t_i^* is a latent variable because empirically observed thresholds are censored at 0% and 100%. We will account for this when estimating the model.

[9] Refer to Battaglini and Palfrey (2024) for an equilibrium analysis of a related model. Their findings provide a micro-foundation for threshold models in that, in Perfect Bayesian Equilibria, players adopt threshold strategies based on the number (and timing) of others who have acted. Furthermore, it is shown that precise equilibrium characterization is analytically intractable except if n grows without bound: "beliefs and behavior are history-dependent, and there are multiple equilibria, so even numerically solving for equilibrium value functions is a daunting task" (p.3).

[10] To support the comparative statics, note that $\frac{\partial t^{**}}{\partial \Delta v_i} = -\frac{1}{2\beta_i+\gamma_i} < 0$. Moreover, $\frac{\partial t^{**}}{\partial \beta_i} = \frac{2\Delta v_i - \gamma_i}{(2\beta_i+\gamma_i)^2}$ is positive if $t^{**} < 0.5$ and negative if $t^{**} > 0.5$. Finally, $\frac{\partial t^{**}}{\partial \gamma_i} = \frac{\beta_i + \Delta v_i}{(2\beta_i+\gamma_i)^2} > 0$ if $t^{**} < 1$: the effect of γ_i is strongest for individuals with high Δv_i. The effect of γ_i reverses only at the boundary for individuals who choose a threshold of 100% (in the absence of the error term).

2.5 Experimental treatments

According to our model, three factors affect an individual's threshold (aside from forward-looking beliefs): her perceived benefits, her attitudes toward conformity, and the pressure she feels from others to conform to the status quo. To test the impact of each factor, we elicit thresholds under different conditions.

First, we vary whether the default organization is anti-AA or pro-AA. We randomize the default organization such that in half of the groups we elicit thresholds for *supporting* AA (t_i^{AA}), and in the other half we elicit thresholds for *opposing* AA (t_i^{NoAA}). This manipulation allows us to exogenously vary individuals' perceived benefits of change, Δv_i, and provides us with a clear hypothesis regarding the threshold choices.

Second, we vary individuals' reference groups. Evidence suggests that people are more likely to care about conforming to the actions of others who are similar to them (e.g., Bicchieri, 2006; Fatas et al., 2018; Ehret et al., 2022). Hence, we expect conformity — the β-parameter — to be higher if group members share more attributes. For this, we elicit two thresholds per individual: a *population threshold*, where participants choose thresholds in a group of 100 people representing the general U.S. population, and an *REG threshold*, where participants choose thresholds in a group homogeneous in gender, race/ethnicity, or both. One of the threshold choices (population or REG) is randomly chosen to determine participants' donations.

Finally, to evaluate the impact of the external pressure to conform — the γ-parameter — on thresholds, we manipulate the privacy of the donations. Twenty percent of participants are allocated in a *Private* condition in which donations to organizations are private information. Eighty percent of the participants are assigned into a *Public* condition where the email addresses and donations of those who supported change ($a_i = 1$) are posted on a publicly accessible website.[11] Additionally,

[11]The website is available at: www.HowPeopleThinkAbout.org/AffirmativeAction. The email addresses posted were the ones that the individuals use for participating in the Ipsos panel. We opted to post the donations only of those who abandoned the default organization rather than all donations in order to meet the requirement of the IRB. Specifically, this design gives every participant the option not to have their email address appear online. This can be achieved by keeping their donation to the default organization. As per our IRB protocol, participants' email addresses were removed from the website six months after the data was collected. Redacted screenshots are available in the Online Appendix F.

participants are informed that we may share the study's results and website on social media. While we expect that the external pressure to conform will increase thresholds in the *Public* condition, the lack of in-person contact with other participants implies that we are likely to be underestimating the external pressure to conform in daily life.

2.6 Individual measures of benefits and conformity

In addition to the exogenous variation in benefits and conformity, we also collect individual measures to enrich our empirical analysis.

For each individual, we create a *Benefits Index* to proxy an individuals' perceived benefits from affirmative action (Δv_i). Specifically, we ask each participant about their agreement to the following statements on a five-point Likert scale: (*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. Participants' responses are aggregated and normalized such that they lie between -0.5 and 0.5, where a greater positive number indicates higher perceived benefits from AA, and a negative number indicates greater perceived harm from AA. As these questions do not directly ask about whether the individuals themselves stand to benefit from AA, the *Benefits Index* should primarily capture perceptions about the social benefits of AA, even though these can be self-serving.

To obtain a measure of β_i, i.e., an individual's inclination to conform to others' choices, we employ an instrument used to capture conformity (e.g., Hong and Page, 1989; Hong and Faedda, 1996; Goldsmith et al., 2005). This measure has been shown to predict an individual's willingness to deviate from social norms (Andreoni et al., 2021). Participants are asked to indicate on a five-point scale their agreement to the 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. We aggregate and normalize the responses such that they lie between 0 and 1, where 1 indicates individuals with the greatest tendency to conform.

Finally, to evaluate the impact of the pressure felt by individuals to conform (γ_i), we measure their perceptions about the strength of norms using an incentivized measure. One way for assessing the strength of a norm is the sanctions imposed on deviators (Fehr and Fischbacher, 2004; Bicchieri, 2006). We first ask participants how likely they would be to confront others who speak out in favor or against AA (depending on the default organization). After that, we ask them to guess the number of others in their group who said that they are likely to confront others. Guesses that are within five percentage points of the true answer add a lottery ticket for one of the 98 Amazon vouchers to the participant's account.[12]

Appendix A provides summary statistics of the elicited measures. The exact wording and order of all measures can be found in Appendix E.

2.7 Behavioral hypotheses

Table 2 offers a summary of how the theoretical framework links with our experimental design. The first column identifies the different threshold components. The second column shows the predicted effects of these components on individual thresholds. The third column presents the exogenous variation that will be used to identify the causal impact of the threshold component on individual thresholds, while the fourth column presents the respective individual measures.

We proceed to present three behavioral hypotheses that will guide our empirical analysis. Our hypotheses were preregistered (AEARCTR-0010895).[13] Recall that t^{AA} denotes thresholds for change toward AA, while t^{NoAA} denotes thresholds for change against AA.

Hypothesis 1 (Perceived Benefits)**:** *The greater the perceived benefits from AA are, the lower (higher) the t^{AA} (t^{NoAA}) will be. Specifically: (i) individuals from underrepresented REG groups will have lower t^{AA} and higher t^{NoAA} than White men; (ii) individuals with a higher Benefits Index will have lower t^{AA} and higher t^{NoAA}.*

[12] We also elicit risk preferences via 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?" Subjects answered on an eleven-point scale. Dohmen et al. (2011) provide evidence that unincentivized questions on risk preferences strongly correlate with incentivized measures. We follow their approach to keep the risk elicitation brief.

[13] Details about the preregistration can be found in Appendix B. There we also present a few results that we preregistered but decided not to discuss in the main body of the paper.

Table 2: Theory and Design Overview

Threshold Component	**Predicted Effect**	**Exogenous Variation**	**Individual Measures**
Perceived Benefits (Δv)	$\frac{\partial t^*}{\partial \Delta v} < 0$	REG group / Status quo org.	Benefits Index
Conformity Internal (β)	$\frac{\partial t^*}{\partial \beta} = \begin{cases} > 0 & \text{if } t^* < 0.5 \\ < 0 & \text{if } t^* > 0.5 \end{cases}$	Reference group	Conformity score
Conformity External (γ)	$\frac{\partial t^*}{\partial \gamma} > 0$	Privacy	Beliefs about norm strength

Notes: The table shows the predicted effects of threshold components on threshold choices. Exogenous variation and individual measures summarize how we test the predictions empirically. REG abbreviates racial/ethnic/gender.

Hypothesis 1 follows from the fact that $\frac{\partial t_i^*}{\partial \Delta v_i} < 0$. Additionally, we hypothesize the White men will be the ones with the lowest perceived benefits from AA. With regards to underrepresented groups, an implication of (i) is that, as long as the perceived benefits from AA are positive, they will have $t^{AA} < t^{NoAA}$.

Hypothesis 2 (Conformity internal)**:** *The greater one's tendency to conform, the more likely thresholds will be interior. Specifically, thresholds will be more likely to be interior when (i) the reference group is narrow (own REG group) than broad (US population), (ii) the greater an individual's Conformity score.*

Hypothesis 2 follows because $\frac{\partial t_i^*}{\partial \beta_i}$ implies that an increase in β_i pushes thresholds inward. Additionally, as mentioned previously, we hypothesize that individuals are more willing to conform to the choices of others who are similar to them.

Hypothesis 3 (Conformity external)**:** *Thresholds will be lower if choices are private than if they are public. Thresholds will also be lower for individuals who perceive the punishment threat for deviating from the default organization to be weaker.*

Hypothesis 3 follows because $\frac{\partial t_i^*}{\partial \gamma_i} > 0$. The hypothesis is tested via variation in the privacy of donations and in the perceived norm strength. We also hypothesize that the decrease in thresholds between the public and private conditions will be greatest for risk-averse individuals, assuming risk aversion increases the perceived threat of punishment.

3 Results

Given the richness of our dataset, before presenting results from regression analyses to formally test Hypotheses 1, 2 and 3, we offer an overview of the data. The left-hand side of Table 3 presents average thresholds across racial, ethnic and gender (REG) groups, and across conditions. The upper half of the table presents average thresholds for the different REG groups, while the lower half presents averages after pooling observations across different REG groups.[14]

The data in the upper half of Table 3 provide some first evidence in support of our hypotheses. In line with Hypothesis 3, we see that thresholds are higher in the Public than in the Private treatment (in 12 out of 16 within-treatment REG comparisons). Also, in line with Hypothesis 1, we see that White men have higher thresholds for supporting AA (t^{AA}) than those in underrepresented groups (in 11 out of 14 within-treatment comparisons) and lower thresholds for opposing AA (t^{NoAA}) (in 13 out of 14 within-treatment comparisons). We also see that for all underrepresented groups (i.e., all REG groups other than White men), the threshold for supporting AA is lower than the threshold for opposing it (in 13 out of 14 within-treatment comparisons). The data in the lower half of Table 3 indicates that the variation in average thresholds is intuitive. As one would expect, women have lower thresholds for supporting AA than men, as do Asian, Black, and Hispanic Americans compared to White Americans. Notably, the largest difference is between Democrats and Republicans, with the latter having thresholds which are far less conducive to supporting AA.

To formally test Hypothesis 1 and Hypothesis 3, Table 4 presents estimates from OLS regressions where the dependent variable is an individual's threshold for supporting AA (t^{AA}).[15] Both hypotheses are supported by the data. As can be seen in columns (1), (5), (6) and (8), in line with Hypothesis 1, thresholds for AA

[14] According to the latest U.S. census, Asian, Black, Hispanic and White constitute 98.7% of the U.S. population. While our sample does not include Native or Alaska American, or Pacific Islander people, for brevity, we refer to the weighted average across REG groups as average of the "U.S. population". This is presented in the last row of Table 3. To calculate it, we used the following weights which are taken from the 2021 wave of the American Community Survey: Asian Female 0.035, Asian Male 0.031, Black Female 0.069, Black Male 0.064, Hispanic Female 0.094, Hispanic Male 0.098, White Female 0.304, White Male 0.305.

[15] Table C.1 in Appendix C presents the same regression models for the thresholds against AA (t^{NoAA}), providing independent confirmatory evidence for Hypothesis 1 and Hypothesis 3.

Table 3: Summary Statistics of Threshold Choices

	Average threshold				Share interior thresholds			
	t^{AA}		t^{NoAA}		t^{AA}		t^{NoAA}	
Reference Group	U.S.	U.S.	U.S.	U.S.	U.S.	REG	U.S.	REG
Condition	Public	Private	Public	Private	Public	Public	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 and share of interior thresholds across groups and conditions. Average thresholds are those for the U.S. population reference group. When necessary average thresholds are calculated using U.S. population weights. For brevity, the share of interior thresholds refers to the the Public condition.

Table 4: Perceived Benefits, Norm Strength, and Thresholds $(\Delta v, \gamma)$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Benefits Index	-38.618***				-36.303***	-35.495***		-35.589***
	(3.480)				(3.534)	(3.631)		(3.666)
Public Donation		5.234***					5.180***	5.981***
		(1.698)					(1.692)	(1.657)
Norm strength		20.510***					20.575***	16.389***
		(3.051)					(3.056)	(3.151)
Asian/Female			-5.787*		-3.475		-6.368**	-5.460*
			(3.088)		(2.946)		(3.072)	(3.011)
Asian/Male			-2.390		-1.397		-3.286	-2.008
			(2.987)		(2.815)		(2.984)	(2.803)
Black/Female			-14.109***		-9.153***		-14.209***	-6.788**
			(2.958)		(2.899)		(2.948)	(2.951)
Black/Male			-13.953***		-9.772***		-14.636***	-7.698***
			(2.917)		(2.817)		(2.910)	(2.850)
Hispanic/Female			-5.566*		-2.193		-7.489**	-6.930**
			(3.061)		(2.910)		(2.977)	(2.930)
Hispanic/Male			-5.932**		-2.553		-7.530***	-1.814
			(2.870)		(2.722)		(2.855)	(2.698)
White/Female			-8.155**		-6.145**		-8.023**	-7.942***
			(3.180)		(3.016)		(3.173)	(3.050)
Democrat				-9.038***		-5.200***		-5.695***
				(1.744)		(1.710)		(1.703)
Republican				5.446**		3.178		2.168
				(2.207)		(2.130)		(2.165)
College								2.917**
								(1.481)
Age								0.066
								(0.060)
Constant	48.618***	32.056***	51.627***	46.857***	52.687***	48.910***	39.721***	38.551***
	(0.812)	(1.921)	(2.234)	(1.441)	(2.054)	(1.418)	(2.890)	(4.160)
Observations	4070	4070	4070	3836	4070	3836	4070	3836
Subjects	2,035	2,035	2,035	1,918	2,035	1,918	2,035	1,918

Notes: OLS regressions on thresholds for AA ($t^{AA} \in [0, 100]$) with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data includes two thresholds per individual (population and REG threshold). Benefits Index (normalized to -0.5 and 0.5) reflects an individual's perceived social benefits of AA policies. Norm strength is measured via participants' expected social sanctions when speaking in favor of affirmative action (normalized between 0 and 1). White males are the omitted category in columns (3), (5). Independents and individuals who do not have a college degree are the omitted groups in columns (4) and (8).

decrease as an individual's perceived benefits from AA increase. Specifically, the estimates indicate that participants with the highest Benefits Index have t^{AA} which are between 35.5 and 38.6 points lower than those of participants with the lowest Benefits Index.[16] In line with Hypothesis 3, the model in column (2) reveals that thresholds are 5.2 points higher in the Public condition, a 16.3% increase relative to the Private condition as indicated by the constant. The effect is robust to controlling for other variables in columns (7) and (8). Additional analysis presented in Appendix B indicates that this effect is strongest for risk-averse subjects. Also in line with Hypothesis 3, t^{AA} is found to increase significantly in the perceived strength of the norm, which we measure by a person's belief about how many others would confront people who speak out in favor of affirmative action (see Section 2.6). Our estimates show that this effect is separate from that of privacy and persists when controlling for REG group membership.

The coefficients in column (3) indicate that, in line with Hypothesis 1, members of underrepresented REG groups all have significantly lower t^{AA} than White men, with the exception of Asian men. Regression (4) shows that Democrats have significantly lower thresholds for supporting AA than Independents, whereas Republicans have significantly higher thresholds. Regression model (5) shows that the coefficient for each REG group decreases when controlling for the Benefits Index, but remains significant for White females and Black participants. It appears that belonging to an underrepresented group affects behavior beyond the perceived benefits of AA. One reason could be that the Benefits Index primarily captures the social benefits of affirmative action, while REG group membership serves as a proxy for an individual's private benefits from AA. Another explanation might be that REG group membership activates identity concerns that influence threshold choices beyond perceived benefits from AA. In regression model (6), we observe a similar effect for the coefficients of elicited political preferences, which can also be a strong indicator of identity (Iyengar et al., 2012; Ehret et al., 2022). When controlling for the Benefits Index, the coefficients of the political preferences decrease, but Democrats still have significantly lower thresholds than Independents and Republicans. The estimates in column (8) demonstrate that these findings remain when controlling for individuals' age and college education.

[16] Appendix C explores the robustness of this result to the way we construct the Benefits Index, showing that the result holds separately for each of the four statements comprising the index.

Table 5: Threshold interiority, 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	(5) $t_i \notin \{0,100\}$	(6) dist. to 0 or 100	(7) $t_i \notin \{0,100\}$	(8) dist. to 0 or 100
Conformity score	0.136***	3.531***					0.117***	2.838**
	(0.029)	(1.191)					(0.029)	(1.198)
REG ref/ce group			0.045***	1.745***			0.045***	1.745***
			(0.005)	(0.232)			(0.005)	(0.232)
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	0.705***	17.541***	0.741***	18.203***	0.680***	17.303***	0.610***	15.293***
	(0.014)	(0.584)	(0.007)	(0.282)	(0.020)	(0.752)	(0.023)	(0.907)
Observations	7972	7972	7972	7972	7972	7972	7972	7972
Subjects	3986	3986	3986	3986	3986	3986	3986	3986

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, $\min(t_i, 100 - t_i)$. Ref. group (REG) is a dummy for whether group members share gender, or race/ethnicity, or both. White males are the omitted category in columns (5) to (8).

The following finding summarizes the results of the tests of Hypothesis 1 and Hypothesis 3.

Finding 1 (**Threshold averages**): *In line with Hypothesis 1 and Hypothesis 3, individuals from underrepresented groups have lower t^{AA} and higher t^{NoAA} than White men. Individuals who believe affirmative action has greater benefits have lower t^{AA}. Finally, thresholds are higher when donations are public.*

Next, we turn our attention to the extent of interdependent behavior in our

sample and Hypothesis 2. Like before, we first examine the variation across REG groups. The right-hand side of Table 3 presents the share of individuals with interior thresholds, i.e., $1 \leq t_i \leq 99$. This measure is important because threshold models would fail to provide meaningful insights when most group members have thresholds of either 0 or 100.

We find that a large majority of individuals — depending on the condition between 69.24% and 74.67% of the U.S. population — have interior thresholds. The remaining 25.33% to 30.76% of individuals have thresholds that are either 0 or 100. We also find that, in line with Hypothesis 2, the share of interior thresholds is larger when the reference group is narrower, i.e., when individuals condition their thresholds on members of the same REG group. This result holds in 16 out of 16 within-group comparisons. Finally, we note that White men and women have a lower share of interior thresholds than others in our sample. One conjecture is that this result reflects underlying differences in attitudes toward conformity. To explore the reasons behind this finding and to provide a formal test of Hypothesis 2, we turn to a regression analysis.

We consider two dependent variables in the regression analysis: (*i*) a dummy variable indicating whether a threshold is interior, i.e., between 1 and 99, (*ii*) a continuous variable measuring a threshold's distance from 0 or 100, whichever distance is smaller. Table 5 shows that in line with Hypothesis 2, the greater an individual's conformity score, the more likely an individual's threshold is interior (Column 1), and the greater the distance from 0 or 100 (Column 2). Also in line with Hypothesis 2, the estimates in columns (3) and (4) indicate that REG thresholds are more likely to be interior and more distanced from 0 or 100 than U.S. population thresholds. In other words, interdependence is significantly larger in narrower reference groups.

The estimates presented in columns (5) and (6) show that non-White Americans have more interior thresholds than White men. Interestingly, the coefficients in columns (7) and (8) show that these differences largely persist when controlling for individual attitude toward conformity.[17] Therefore, explanations for the less interior thresholds of White men likely involve cultural and identity-based factors rather

[17] Appendix C provides further tests on interiority. Controlling for the Benefits Index does not affect the other estimates. Moreover, sharing race/ethnicity and gender does not lead to more interior thresholds than having only one of these characteristics in common with the other group members.

Table 6: Structural Estimates of Model Parameters

	(1) Proxy of Δv_i: Ind. Benefits Index	(2) Proxy of Δv_i: REG Avg. of Benefits Index
Internal Conformity ($\hat{\beta}$)	0.846***	0.907***
	(0.050)	(0.117)
External Conformity ($\hat{\gamma}$)	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
Clusters	3,986	3,986

Notes: Maximum likelihood estimation of model parameters with standard errors clustered by subject, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. We use the Benefits Index to proxy variation in perceived benefits. Model (1) uses the individual Benefits Index values to proxy Δv_i. To avoid endogeneity between individuals' threshold choices and their Benefits Index, model (2) proxies Δv_i by the average Benefits Index of each REG group (i.e., there are eight difference values of Δv_i and individual-level variation in the Benefits Index is not used). See Appendix C for details.

than differences in psychological traits such as conformity preferences. For example, White males may experience greater cultural polarization than other groups, with strong social identities and targeted media consumption driving them toward more extreme positions at both ends of the spectrum.

Finding 2 summarizes our results concerning the support for Hypothesis 2.

Finding 2 (**Interdependent behavior and conformity**)**:** *The majority of individuals exhibit interdependent behavior. Threshold are more likely to be interior in narrower reference groups and the more conformist individuals are.*

Lastly, recall that besides perceived benefits and external and internal conformity, the model suggests that thresholds depend on forward-looking beliefs. Specifically, if individuals anticipate that their support for change will induce others to follow suit, expression (3) implies a negative value of μ. We explore whether individuals have such forward-looking beliefs by means of a structural model. The estimates are based on standard maximum-likelihood routines. We proxy Δv_i by the Benefits Index and use the variation in threshold choices to estimate internal conformity ($\hat{\beta}$), external conformity ($\hat{\gamma}$), and forward-looking beliefs ($\hat{\mu}$); see Appendix C for details.

The results in Table 6 show that the estimated mean of the error term, $\hat{\mu}$, is indeed negative: individuals choose thresholds that are 5.1 to 6.1 percentage points lower than they would if behaving myopically. This indicates that participants anticipate their support for change will induce others to follow suit. The estimates further confirm that internal conformity ($\hat{\beta}$) and external conformity ($\hat{\gamma}$) are significant drivers of behavior. In Appendix C, we show that the estimates imply that the impact of conformity on threshold choices is about as strong as the impact of the perceived benefits.

4 Applications

Information about the distribution of thresholds and their correlates can offer unique insights to researchers and a structured approach for influencing the spread of specific behaviors (or technologies) for policymakers. In this section, we provide some examples by showing how our method can be used to predict societal outcomes, determine instances when interventions can have outsize effects, identify groups for targeting, and recognize contexts where interventions may be ineffective due to existing social dynamics. We also discuss how our method can be used to study network effects. While we use our data on affirmative action (AA) to illustrate these points, the method can be used in any context where behavior could be captured with threshold models, e.g., diffusion of innovation, participation in collective action, and following fads.

4.1 Predicting equilibria and identifying tipping potential

The most obvious application of our method is in predicting societal outcomes. Figure 2 presents threshold distributions for the U.S. population. Figures 2a and 2b display the probability distribution function and the cumulative distribution function (CDF) of thresholds for AA (t^{AA}), respectively.[18] The societal equilibrium denoting the share of individuals supporting AA is identified by the intersection or tangency points of the CDF with the 45-degree line from above (Granovetter, 1978).

The CDF for the U.S. population in Figure 2b reveals the existence of two societal

[18]It is worth noting that thresholds are not normally distributed, as is often assumed in previous research (e.g., Granovetter, 1978; Young, 2009; Bicchieri, 2016; Andreoni et al., 2021).

equilibria: one where 50% of individuals support AA and another where 61% support AA. Threshold models select the lower equilibrium as the relevant one because this is where the process of behavioral change tends to get stuck.[19] However, the existence of a second equilibrium is of special interest for policymakers as it suggests the possibility of social tipping.

To illustrate this point, Figure 2c shows the CDF of t^{AA} after an intervention that reduces thresholds for a randomly selected subgroup of individuals. (We discuss the channels through which thresholds can be altered below.) We randomly selected 10% of society and reduced their thresholds by a normally distributed shock with mean of 50 and a standard deviation of 10. As can be seen, the intervention increases the predicted equilibrium share of change supporters from 50% to 68%, i.e., by more people than were targeted. If we define the *tipping potential* of an intervention in a population as the predicted change in the share of supporters divided by the share of targeted individuals, the tipping potential in Figure 2c is (68%-50%)/10% = 1.8.

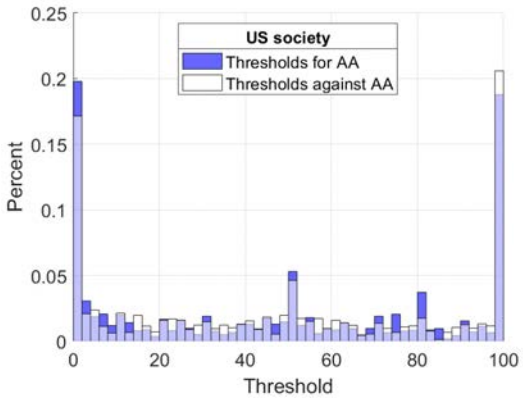
The method introduced in this paper thus allows policymakers to measure the tipping potential of different interventions to help allocate scarce resources and attention. A distribution like example 1 in Figure 1 has limited tipping potential. In contrast, a distribution like example 2 in Figure 1 has enormous tipping potential due to the existence of two equilibria that are far apart, implying drastically different societal outcomes. Because the distribution in example 2 has multiple equilibria, an intervention does not need to increase a specific equilibrium but rather push society past a tipping point to achieve an equilibrium transition.

4.2 Identifying groups of individuals for targeting

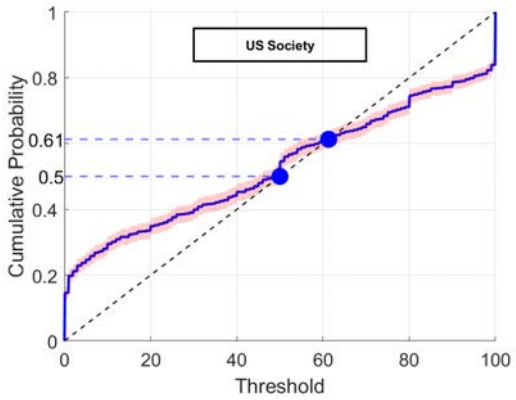
Once a policymaker has determined that he or she wants to promote a certain type of behavior (or the adoption of a given technology), a natural question that arises is who should be targeted. If behavior is interdependent, the social dynamics embedded in threshold models imply that targeting random samples of the population will never be better than targeting specific groups of individuals in terms of impact, all else equal, so long as the "right" group of individuals is identified (Efferson et al., 2020, 2024). Information about the distribution of thresholds in the population of interest allows us to determine the groups of individuals to target.

[19] Appendix D discusses the variance of the simulated societal equilibria.

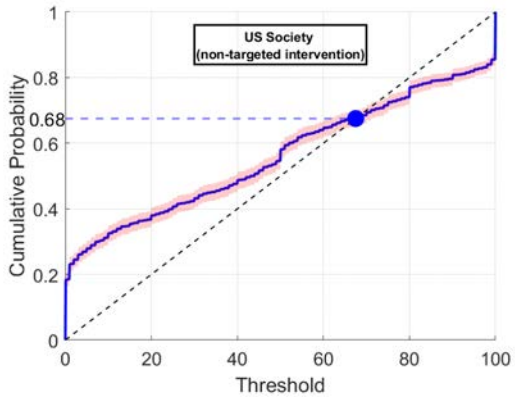
Figure 2: Threshold Distribution in US Society



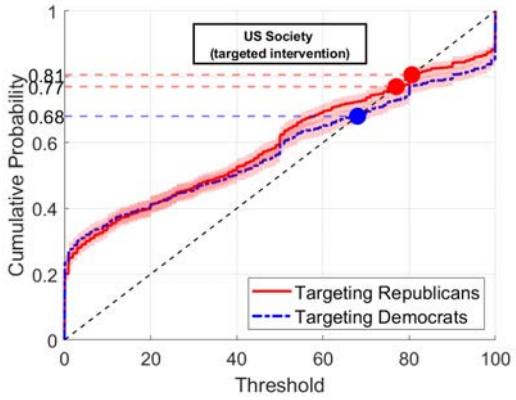
(a) Probability Mass Function



(b) Cumulative Distribution Function



(c) Intervention (Non-Targeted)



(d) Intervention (Targeted)

Notes: Figure (a) shows the distribution of t^{AA} and t^{NoAA} (U.S. population weights, public condition, population reference group). Figure (b) shows the t^{AA} CDF and equilibria; the solid line depicts the median from 10,000 random samples of the elicited thresholds for $n = 1,000$, the shaded region including 90% of simulation outcomes. Figure (c) shows the t^{AA} CDF after a non-targeted intervention that reduces thresholds by a normally distributed shock with a mean of 50 and a standard deviation of 10 for 10% of the individuals (randomly selected). Figure (d) shows the t^{AA} CDF when targeting individuals self-reporting as *Republican/strong Republican* or *Democrat/strong Democrat* (each 18.4 % of the total population) using the same normally distributed shock.

Threshold models suggest that the "right" targets for intervention are those with thresholds *to the right* of the societal equilibrium (pun intended). As a rule of thumb, targeting groups whose members have thresholds *not far above* the equilibrium can be particularly effective, as this approach closes a "gap" in the threshold distribution in the least expensive way. Our method can provide vital information by determining where the societal equilibrium point is and identifying groups with a high fraction of individuals whose thresholds are not far above this equilibrium. For other interventions, however, particularly those expected to have a large impact on individuals' perceived benefits from change, it might be beneficial to target more resistant individuals (e.g., Efferson et al., 2020, 2024). Even in such cases, knowing the predicted societal equilibrium remains crucial because, if the equilibrium is low, targeting those in the intermediate threshold range may have the greatest tipping potential.

To illustrate these points with our data, let us consider the likely impact of a given intervention targeting either Democrats or Republicans. Such targeting is simple in principle given that the two groups tend to rely on different information channels. Importantly, Republicans were found to have a high fraction of individuals with thresholds to the right of the first equilibrium (see Table 3). Figure 2d presents the CDF of t^{AA} after an identical intervention on the two groups. We find that the intervention targeting Republicans increases equilibrium support for AA to 77% while targeting the same share of the population, but focusing on Democrats leads to lower equilibrium support of 68%.[20]

Beyond Democrats and Republicans (who are unlikely to be equally responsive to interventions on AA), our method can be used to determine what other social groups — depending on age, gender, education, urban versus rural, etc. — should be targeted to ensure efficient use of resources.

4.3 Determining the content of interventions

After determining what issue to intervene on and having identified the groups of individuals to target, the next step for a policymaker is to determine the content of

[20]Note that each group comprises 18.4% of the sample. Interestingly, targeting the 18.4% of Democrats in Figure 2d has the same impact on the societal equilibrium as the non-targeted intervention on 10% of society in Figure 2c. Randomly targeting 18.4% of the U.S. population would lead to an equilibrium with 72% of supporters.

the intervention. For example, should they use pecuniary incentives or should they simply "nudge" individuals? If they decide to nudge, should they provide information about the benefits from change to the individual or should they share success stories from other groups/societies to boost expectations for change? Information about the determinants of thresholds can help answer such questions.

Within the framework of our model interventions can modify perceived benefits (Δv), internal conformity (β), external conformity (γ), or beliefs about one's efficacy in driving social change (μ). For example, Field et al. (2021) provide financial incentives to liberalize gender norms in rural India, which influence thresholds by altering perceived benefits. Bursztyn et al. (2020) and Bursztyn et al. (2023) provide information to correct misperceptions about gender norms that hinder progress toward gender equality. Correcting such misperceptions can change thresholds through the external conformity parameter (γ), especially in the case of pluralistic ignorance, when a majority of individuals privately reject a norm but publicly conform to it because they believe others accept it.[21] The Saleema campaign, on the other hand, was a nudge aimed to eradicate female genital cutting in Sudan by manipulating primarily the internal pressure to conform (β) (Bicchieri, 2016). Lastly, the case of Hollywood-producer Harvey Weinstein illustrates the importance of beliefs about whether changing one's behavior will cause others to follow suit (μ).[22]

When altering the perceived benefits of change either by providing information (e.g., about the consequences of smoking) or offering financial incentives, the model's qualitative predictions are straightforward: the thresholds of targeted individuals decrease, leading to an increase in the number of change supporters in equilibrium. By contrast, the effects of altering conformity parameters through

[21] In our context, beliefs about norm strength are relatively accurate. Specifically, 35.12% [32.85%] of individuals responded that they would confront someone who speaks in favor of AA [against AA], while the average belief about the share of others who would confront is 40.88% [39.43%]. This indicates some degree of pluralistic ignorance, but it is limited. In line with this finding, individuals report beliefs that moderately underestimate the number of change instigators (individuals with $t_i^a < 20$) in their groups.

[22] For nearly thirty years, a "code of silence" allowed Weinstein to sexually harass and assault women in the movie industry. His actions were perpetuated by the victims' fear of retaliation. Everyone remained silent, believing that speaking out would not make a difference. This changed in October 2017, when The New York Times and The New Yorker published articles documenting multiple allegations. These publications not only exposed Weinstein but also coordinated allegations, fostering the belief that if one woman spoke out, others would follow. This shift in belief empowered over 80 women to come forward with their stories.

interventions (e.g., promoting in-group identity) are ex-ante unclear. Conformity shifts thresholds toward the midpoint of 50%, which can either promote or hinder change. To illustrate, using our data and model estimates (see Table 6), we find that a counterfactual increase in the internal pressure to conform (β) leads to a decrease in support for AA among Black men but increased support among White men. Conversely, decreasing conformity leads to lower support for affirmative action among Asian men while increasing support among Black women. Precise information about threshold distributions allows one to assess the influence of conformity in a specific context.

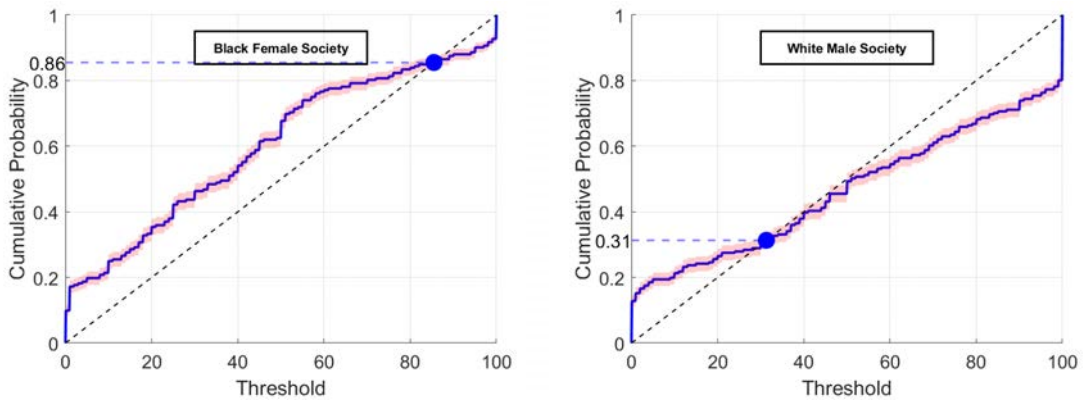
What about manipulating beliefs about the prospects of change? To illustrate using our data, consider a policymaker who considers promoting AA by making individuals more optimistic about their ability to promote change in their organization (μ), e.g., by designing a workshop with success stories of AA in peer organizations. Utilizing the structural estimates in Table 6, we can evaluate the impact of varying μ. If we were to double the degree of forward-looking behavior by increasing μ from -5.1 to -10.2, the average predicted support for AA rises from approximately 50% to 74% — a substantial increase. This counterfactual analysis underscores the significant role of forward-looking behavior in driving norm change. Interventions that increase μ could involve the creation of platforms for individuals to voice their ideas and lead initiatives.

The examples explored in this section illustrate how different interventions target various threshold determinants. It is important to note that the discussed interventions can often impact multiple threshold components, and their effects are not limited to a single interpretation. We do not wish to claim that our model captures all possible channels through which interventions may operate. Nonetheless, this discussion demonstrates that the threshold elicitation method allows for a detailed and theoretically rigorous assessment of the impact of different types of interventions.

4.4 The role of social networks

When using threshold models to make forecasts about social outcomes, one needs to consider the role of social networks. To demonstrate the influence of networks, we turn to our data. Recall that we elicited two thresholds for each individual in our

Figure 3: Threshold Distributions in REG segregated groups



Notes: Distribution of thresholds for AA (t^{AA}) of Black females (left) and White males (right) 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.

sample that differed in the individual's reference group: a U.S. population threshold and an REG threshold where individuals were assigned to groups with people sharing the same race/ethnicity/gender. We saw in Table 3 that REG thresholds were more likely to be interior. What about the predicted support within REG groups?

Figure 3 plots the CDFs of the REG thresholds for Black females and White males. A few things stand out. First, for both groups, the model predicts a unique equilibrium, unlike in Figure 2b. Second, the predictions are markedly different. Reflecting differences in perceived benefits, the model predicts a support of AA of 86% among Black females versus 31% among White males. Importantly, these numbers differ significantly from the predicted behavior of each REG group in the representative network (i.e., the U.S. population thresholds and reference group), where the predicted share of Black females supporting AA is 71.24%, and the share of White males supporting AA is 50.16%. These results suggest that an increase in segregation entails greater polarization of publicly expressed views. Appendix D contains the thresholds for the other REG groups in segregated networks, including one for Hispanic males (t^{NoAA}) with enormous tipping potential resembling the distribution of example 2 in Figure 1.

The analysis above highlights the importance of identifying the appropriate reference group for individuals. To gather information about the individuals' actual

networks and evaluate their impact, we asked participants: *"Among the ten people you most recently met — outside your family — with whom you exchanged opinions, how many do you think identify as [participant's racial/ethnic group]; [participant's gender]?"*. Incorporating these network data in the simulations has a relatively small impact on the predicted aggregate support for AA: the average predicted share of individuals supporting AA is 54.82% in the U.S. representative network (an average that lies between the two equilibria identified in Figure 2b), and 52.83% when accounting for the elicited network structure of individuals. Details for the network-weighted simulations can be found in Appendix D.

5 Conclusion

The research in this paper was spurred by the observation that despite the widespread use of threshold models to analyze a wide range of issues — from adhering to social norms, to adopting new technologies, participating in protests, and determining which consumption goods to purchase — there exists a puzzling lack of empirical evidence supporting their application. To start closing the gap between theoretical and empirical research, we have introduced an incentivized method for eliciting individual thresholds for interdependent behavior. We used this method to study support for affirmative action in a large sample of the US population, stratified over race, ethnicity, and gender, and explore their determinants. Our analysis revealed that a substantial majority of individuals across demographic groups condition their support for affirmative action on the number of others supporting it. Furthermore, in line with our preregistered hypotheses, the thresholds elicited are influenced by individuals' perceived benefits and pressure to conform.

The paper has both theoretical and practical implications. We demonstrated how our method can be leveraged for policy design. Specifically, using the data collected, we showed how information about the distribution of thresholds and their determinants can be used to forecast the impact of interventions, predicting social tipping points, determine who to target to increase the efficacy of an intervention, and design the content of interventions. From a theoretical perspective, our findings contribute to a deeper understanding of threshold models by providing empirical validation of key assumptions. We have shown that, even in an issue polarizing the

U.S. population, individuals do have thresholds that influence their behavior, which can be quantified and analyzed. By demonstrating the role of perceived benefits and conformity pressures, our analysis bridges the gap between theoretical constructs and real-world behavior, enriching the theoretical framework with practical insights. Furthermore, our method allows for the exploration of the heterogeneity in thresholds across different demographic groups, offering a nuanced view of how social influence operates in diverse populations.

Having taken just a first step in bridging the gap between empirical and theoretical research on threshold models, we believe there exist several promising avenues for future research. First, it is important to explore the application of our method across various populations and contexts, such as political participation, technology adoption, and environmental protection, to assess the generalizability and robustness of our findings.[23] Second, future studies should test the predictions obtained from theoretical models using information about the exact distribution of thresholds in a population. Third, exploring the conditions under which the threshold question used in our study can be deployed without incentives would be valuable. This would facilitate empirical research on thresholds by enabling researchers to include these questions in surveys.

In our increasingly interconnected world, the study of interdependent behavior appears to be crucial for understanding how social dynamics shape individual actions and anticipating sudden shifts in collective outcomes. By demonstrating how researchers can elicit individual thresholds and use the information to predict behavior, we hope to have provided scholars and policymakers with useful insights they can leverage to design effective interventions for promoting positive societal change.

References

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.

Andreoni, J., N. Nikiforakis, and S. Siegenthaler (2021). Predicting social tipping

[23] In certain contexts, eliciting thresholds from the target population may be challenging. For instance, ethical considerations would prevent researchers from eliciting individuals' thresholds for participating in protests against an authoritarian regime due to potential risks to the participants.

and norm change in controlled experiments. *Proceedings of the National Academy of Sciences 118*(16), e2014893118.

Asch, S. E. (1952). *Social psychology*. Englewood Cliffs, NJ: Prentice Hall.

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

Battaglini, M. and T. R. Palfrey (2024). Dynamic collective action and the power of large numbers. *NBER Working Paper*.

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.

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., A. W. Cappelen, B. Tungodden, A. Voena, and D. Yanagizawa-Drott (2023). How are gender norms perceived? *NBER Working Paper* (w31049).

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.

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

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

Centola, D. (2018). *How behavior spreads: The science of complex contagions*, Volume 3. Princeton University Press.

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

Cialdini, R. B. and N. J. Goldstein (2004). Social influence: compliance and conformity. *Annual Review of Psychology 55*, 591–621.

Clark, W. A. (1991). Residential preferences and neighborhood racial segregation: A test of the schelling segregation model. *Demography 28*, 1–19.

Coleman, J. S. (1990). *Foundations of social theory*. Harvard University Press.

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.

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

Dvorak, F. and U. Fischbacher (2024). Social learning with intrinsic preferences. *arXiv preprint arXiv:2402.18452*.

Dvorak, F., U. Fischbacher, and K. Schmelz (2024). Incentives for conformity and anticonformity. *Available at SSRN 3754595*.

Efferson, C., S. Ehret, L. von Flue, and S. Vogt (2024). When norm change hurts. *Philosophical Transactions of the Royal Society B 379*(1893), 20220268.

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.

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

Field, E., R. Pande, N. Rigol, S. Schaner, and C. Troyer Moore (2021). On her own account: How strengthening women's financial control impacts labor supply and gender norms. *American Economic Review 111*(7), 2342–2375.

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.

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

Glaeser, E. L., B. I. Sacerdote, and J. A. Scheinkman (2003). The social multiplier. *Journal of the European Economic Association 1*(2-3), 345–353.

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.

Hendriks, A. (2012). Sophie-software platform for human interaction experiments. *University of Osnabrück, Osnabrück*.

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.

Iyengar, S., G. Sood, and Y. Lelkes (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly 76*(3), 405–431.

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

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

Katz, M. L. and C. Shapiro (1985). Network externalities, competition, and compatibility. *The 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.

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.

Mitzkewitz, M. and R. Nagel (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory 22*, 171–198.

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.

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.

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

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

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.

Sunstein, C. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

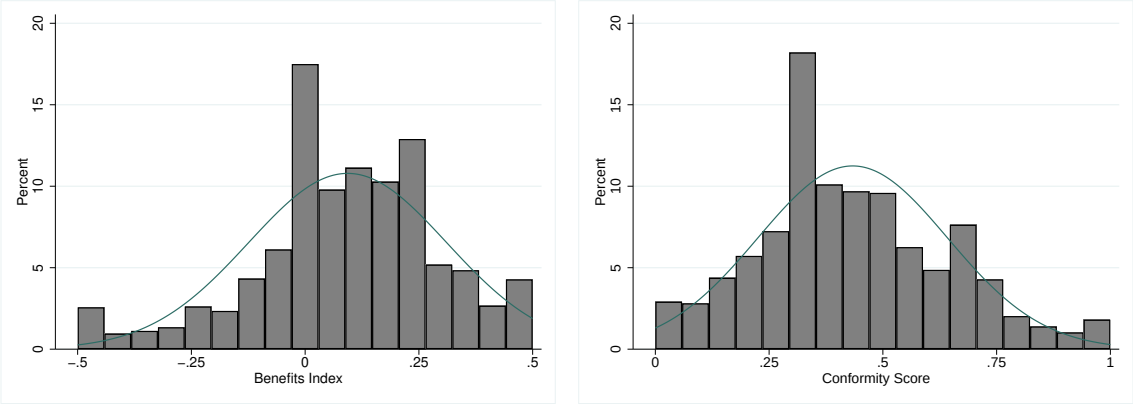
Szkup, M. and I. Trevino (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior 124*, 534–553.

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.

A Overview of Benefits and Conformity Indices

Figure A.1: Distribution of Benefits and Conformity Indices



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

Figure A.1 shows histograms of the Benefits Index and the Conformity score. A higher value on the Benefits Index indicates a more favorable view of affirmative action policies. A higher value of the Conformity score indicates a preference for aligning one's behavior with the majority of others. Table A.1 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 males) to 0.15 (Black females). These numbers indicate an average attitude in favor of AA. Further analysis shows that Asian and White males 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 females and males 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 score range from 0.40 (White males) to 0.48 (Hispanic males). These numbers indicate that most people are moderate conformists, but some value independence.

Table A.1: Averages of elicited measures

	Asian F	Asian M	Black F	Black M	Hisp. F	Hisp. M	White F	White M
Benefits Index	.09	.06	.15	.14	.12	.11	.07	.02
	(.18)	(.20)	(.21)	(.20)	(.19)	(.19)	(.23)	(.28)
Conformity score	.44	.39	.43	.43	.47	.48	.43	.40
	(.20)	(.19)	(.21)	(.21)	(.217)	(.23)	(.20)	(.20)
Risk aversion	.41	.33	.42	.32	.32	.29	.47	.36
	(.26)	(.25)	(.27)	(.25)	(.26)	(.22)	(.26)	(.25)
Norm strength	40.89	41.62	38.40	40.50	46.68	45.04	38.69	37.45
(default anti-AA)	(24.29)	(23.80)	(24.05)	(23.71)	(28.26)	(23.94)	(24.24)	(25.14)
Norm strength	40.57	40.75	38.36	36.33	49.67	39.15	35.33	36.18
(default pro-AA)	(26.57)	(23.47)	(24.21)	(25.58)	(29.30)	(23.98)	(22.38)	(23.99)

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 Score $\in [0, 1]$ 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?"*. Standard deviation in parentheses.

B 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 score (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 default organization is anti-AA and higher thresholds if the default 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 2.7). 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. Finding 1 and Finding 2 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 default organization is anti-AA and higher thresholds if the default 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 default organization is anti-AA and higher thresholds if the default 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 4 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 (4) and (6).

We test Hypothesis H2c in Table B.1 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 score (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 B.1: Conformity, city size, and thresholds

	(1) Conformity score	(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	7972	7972	7972
Subjects	3986	3986	3986

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 score. 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 B.2: 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	7972	7972	7972	7972
Subjects	3986	3986	3986	3986

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. Finding 2 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. As discussed in Footnote 17 of the paper, this hypothesis cannot be supported. Table B.2 columns (1) and (2) provide the supporting analysis. 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 B.2 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 4 of the paper tests 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 B.3: 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	7972	7972	7972
Subjects	3986	3986	3986

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 B.3 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 B.3. The hypothesis can be rejected.[24]

[24]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 B.3 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.

C Supplementary analysis for Section 3

C.1 Threshold averages

Replicating Table 4 for t^{NoAA}: Table C.1 shows that the results summarized in Finding 1 of the paper are replicated for thresholds against AA (t^{NoAA}). In line with Hypothesis 1 and Hypothesis 3, individuals from underrepresented groups have higher t^{NoAA} than White men. Individuals who believe affirmative action has greater benefits have higher t^{NoAA}. Finally, t^{NoAA} are higher when donations are public and the against-AA norm is perceived to be stronger.

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 C.2 demonstrates that each question separately robustly replicates the effect of perceived benefits on threshold choices, including when having all four statements in one regression model.

Test of more nuanced model predictions for external pressure: 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 C.3 confirms this prediction, as the coefficient is approximately double in size for individuals who favor change.

Interiority of thresholds across REG groups controlling for the Benefits

Table C.1: Thresholds against AA: Perceived Benefits, Norm Strength, and Thresholds $(\Delta v, \omega)$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Benefits Index	40.504***				41.363***	37.211***		38.577***
	(3.321)				(3.383)	(3.485)		(3.492)
Public Donation		3.470*					3.355*	3.584**
		(1.817)					(1.808)	(1.795)
Norm strength		0.162***					0.151***	0.073**
		(0.031)					(0.031)	(0.032)
Asian/Female			12.132***		8.744***		11.362***	3.648
			(3.054)		(2.931)		(3.081)	(3.097)
Asian/Male			14.005***		12.088***		13.280***	10.283***
			(3.018)		(2.869)		(3.022)	(2.924)
Black/Female			5.808*		0.613		5.502*	0.233
			(3.058)		(2.925)		(3.082)	(3.054)
Black/Male			5.215*		-0.205		5.239*	0.281
			(3.009)		(2.874)		(2.993)	(2.935)
Hispanic/Female			10.484***		6.094**		8.430***	0.157
			(3.137)		(2.992)		(3.113)	(3.124)
Hispanic/Male			6.130**		2.562		5.547*	2.530
			(2.858)		(2.731)		(2.870)	(2.780)
White/Female			7.271**		4.965*		7.436**	4.780
			(3.148)		(2.929)		(3.156)	(2.948)
Democrat				4.444**		1.570		1.696
				(1.927)		(1.878)		(1.872)
Republican				-5.355**		-2.439		-2.284
				(2.183)		(2.111)		(2.128)
College								7.567***
								(1.549)
Age								-0.145**
								(0.062)
Constant	47.195***	41.616***	43.134***	48.687***	42.792***	46.225***	35.044***	40.030***
	(0.765)	(2.014)	(2.190)	(1.591)	(2.002)	(1.540)	(2.793)	(4.014)
Observations	3902	3902	3902	3624	3902	3624	3902	3624
Subjects	1,951	1,951	1,951	1,812	1,951	1,812	1,951	1,812

Notes: OLS regressions on thresholds against AA ($t^{NoAA} \in [0, 100]$) with standard errors clustered by subject in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data includes two thresholds per individual (population and REG threshold). Benefits Index (normalized to -0.5 and 0.5) reflects an individual's perceived social benefits of AA policies. Norm strength is measured via participants' expected social sanctions when speaking against affirmative action (normalized between 0 and 1). White males are the omitted category in columns (3), (5). Independents and individuals who do not have a college degree are the omitted groups in columns (4) and (8).

Table C.2: 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 default	✓	✓	✓	✓	✓
Observations	8172	8172	8172	8172	8172
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 default, 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.3: Heterogeneity of Norm Strength Effect

	$\Delta v_i > 0$	$\Delta v_i > 0$	$\Delta v_i > 0$	$\Delta v_i < 0$	$\Delta v_i < 0$	$\Delta v_i < 0$
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 & Default		✓	✓		✓	✓
Observations	1714	1714	1714	1573	1573	1573
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.4: 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)
Conformity score	0.115***	2.648**
	(0.029)	(1.205)
REG ref/ce group	0.045***	1.745***
	(0.005)	(0.232)
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	7972	7972
Subjects	3986	3986

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 males are the omitted category.

Index: The regression analyses in Table 5 of the paper revealed differences in the interiority of thresholds between White males and most other REG groups. To see whether the Benefits Index Δv_i can explain this effect (i.e., White males having more extreme views), we report the regressions in Table C.4. The regressions show that the REG group differences prevail even when controlling for perceived benefits.

C.2 Structural estimation

To quantify forward-looking behavior and the relative impact of conformity, we estimate the parameters of the optimal threshold expression given in expression (3) of the paper.

We use standard maximum likelihood routines to maximize the sum of

$$L^{\text{tobit}}(t_i) = \mathbf{i}_{0<t_i<1}\phi\left(\frac{t_i - t_i^{**} - \mu}{\sigma_\epsilon}\right) + \mathbf{i}_{t_i=1}\Phi\left(\frac{t_i^{**} + \mu - 1}{\sigma_\epsilon}\right) + \mathbf{i}_{t_i=0}\Phi\left(\frac{-t_i^{**} - \mu}{\sigma_\epsilon}\right) \tag{5}$$

where ϕ and Φ are the standard normal probability and cumulative distribution functions. The first term in (5) is the probability of observing threshold t_i if the optimal threshold t^* is interior and $\epsilon_i \sim N(\mu, \sigma_\epsilon^2)$. The second and third terms are the analogous probabilities for non-interior thresholds, which are censored at 0 and 1.

We estimate the model parameters taking Δv_i as given. The Benefits Index is the proxy for Δv_i. We use two approaches. First, we proxy Δv_i by the individual Benefits Index values. That is, the individual Benefits Index values replace Δv_i in (5). Second, we proxy Δv_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 Δv_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.

Roughly speaking, the variation in Δv_i allows us to estimate β (internal conformity). The external conformity parameter, 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 6 of the paper shows the results of the estimation. In the paper, we discuss forward-looking beliefs. Here, we consider another question: How substantial is the estimated conformity? Perceived benefits of affirmative action, captured by variation in Δv_i, push thresholds toward the extremes of 0% and 100%. Internal

conformity, β_i, pulls thresholds toward the midpoint and dampens the effect of personal views. External conformity, γ_i, increases thresholds, particularly for people with low thresholds. The marginal change in thresholds in response to a change in Δv_i is

$$\frac{\partial t_i^*}{\partial \Delta v_i} = \frac{1}{2\beta_i + \gamma_i}. \tag{6}$$

Given the empirical estimates in Model (1) of Table 6, $\frac{\partial t_i^*}{\partial \Delta v_i} \approx 0.53$. For model (2), we obtain $\frac{\partial t_i^*}{\partial \Delta v_i} \approx 0.49$. These numbers imply that thresholds change by about half a point per percentage point change in Δv_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.

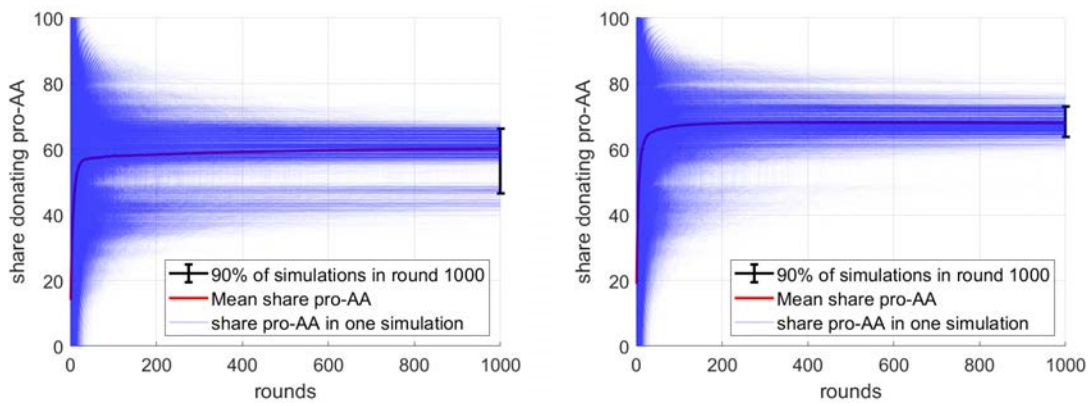
D Supplementary analysis for Section 4

D.1 Variance of equilibria

The cumulative distribution functions (CDF) of thresholds presented in section 4 of the paper allow one to detect all equilibria. However, in any society or group, thresholds will be stochastic and randomly drawn from a population. In general, the variance in predicted outcomes will be higher the closer the threshold CDF is to the 45-degree line. Moreover, variance will be smaller in larger groups, which means the lowest equilibrium, as indicated by the CDF, will emerge most of the time.

Here, we present a simulation approach that allows for path dependence. In the presence of path dependence and multiple equilibria, the higher equilibria can be realized even in large groups because the initial movers determine equilibrium selection. This approach is useful, as it allows us to examine the variance in predicted outcomes even for very large groups. Let $a_\tau \in \{1, 2, ..., n\}$ be the number of active society members in period τ. We begin with only one active individual in the

Figure D.1: Variance of predictions (US Society)



(a) Equilibria of elicited thresholds (b) Equilibria after intervention

Notes: Figures show the equilibrium convergences with path dependence (see description in flow text). Each thin blue line represents one of 10,000 iterations. The thresholds, t^{AA}, are from the Public treatment. Figure (a) uses the unaltered thresholds. Figure (b) uses the threshold after a non-targeted intervention that reduces thresholds by a normally distributed shock with a mean of 50 and a standard deviation of 10 for 10% of the individuals (randomly selected).

first period, $a_1 = 1$. The dynamic rule $g_\tau = F(g_{\tau-1})$, is iterated only among the active individuals. One randomly selected individual is added to the group of active players whenever stability is reached. The process continues until all n individuals are active. The path-dependence model can be interpreted by way of network interactions, where individuals who are active early play a similar role as centrally-positioned players in social networks (e.g., Jackson, 2008; Centola, 2018).

Figure D.1 presents simulations allowing for path dependence. Panel (a) shows that after many iterations, 90% of the simulated outcomes range from a share of 47% to 67% supporting AA. The presence of the two equilibria previously discovered in Figure 2b implies that this variance persists even in large groups. Panel (b) of Figure D.1 shows the same simulations after the policy intervention that lowers the threshold by half for 10% of the individuals. In Figure 2c, we predicted that this would eliminate the lower of the equilibria. The path-dependence approach plotted in Figure D.1(b) indeed shows a much lower variance of societal outcomes, with no second equilibrium, and the interval containing 90% of the simulation ranging from 64% to 73% AA support.

D.2 Additional information on social networks

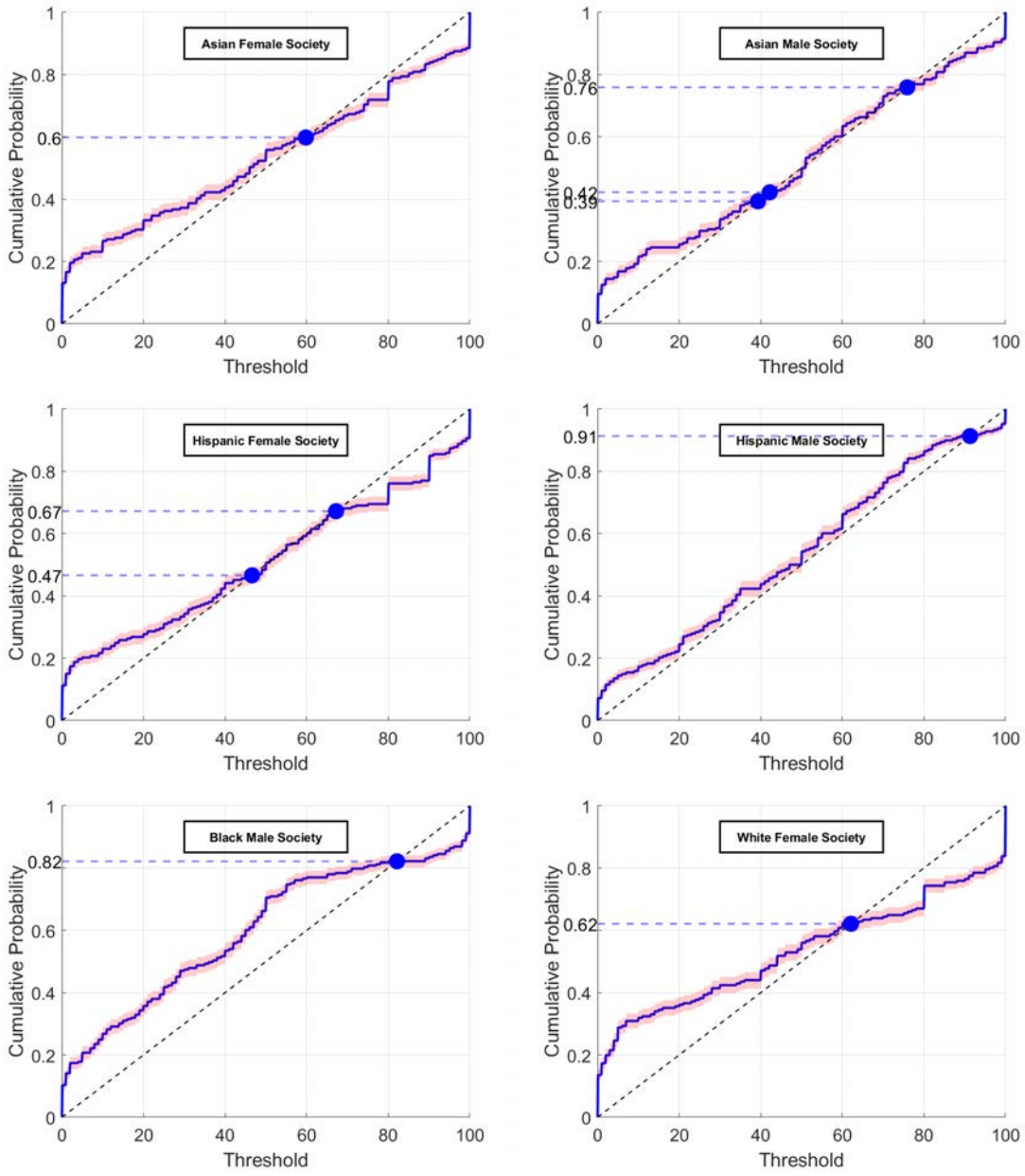
Threshold distributions in each REG group: Figure 3 of the paper shows the distribution of REG thresholds for Black females and White males (t^{AA}). Figure D.2 displays the remaining REG groups. Additionally, Figure D.3 shows the CDFs of all eight REG groups of t^{NoAA}. There exists a large heterogeneity between REG groups in threshold distributions of t^{AA}, with societal equilibria ranging from 39% for AA (Asian males) to 82% and 91% for AA (Black and Hispanic males). The distribution of t^{AA} of Asian males and Hispanic females shows multiple equilibria. Similarly, the distributions of t^{NoAA} are heterogeneous between REG groups, with societal equilibria ranging from 9% and 18% (Hispanic males and females) to 67% and 88% (Black females and Hispanic males). The distribution for Hispanic males shows great tipping potential with multiple equilibria, one at a low level and two at a high level.

Details about elicited social networks: The paper also discusses simulations based on the actual extent of segregation in the U.S., rather than considering segregated REG groups. To be able to account for participants' networks, we asked:
Among the ten people you met most recently—outside your family—and with whom you exchanged opinions, how many do you think identify as [participant's racial/ethnic group]?

The responses indicate considerable segregation. Whites are 1.27 times more likely to engage in same-race interactions than they would in the representative network. Asians are 6.72 times more likely to interact with other Asians than they would in a US-representative network (smaller subgroups have a lower base probability of interacting with a same-race person in the representative network, which explains the higher number of Asians compared to Whites). The factors for Hispanics and Blacks are 3.06 and 4.73, respectively. Using an equivalent question, individuals report to be 1.37 times more likely to exchange opinions with others who have the same gender.

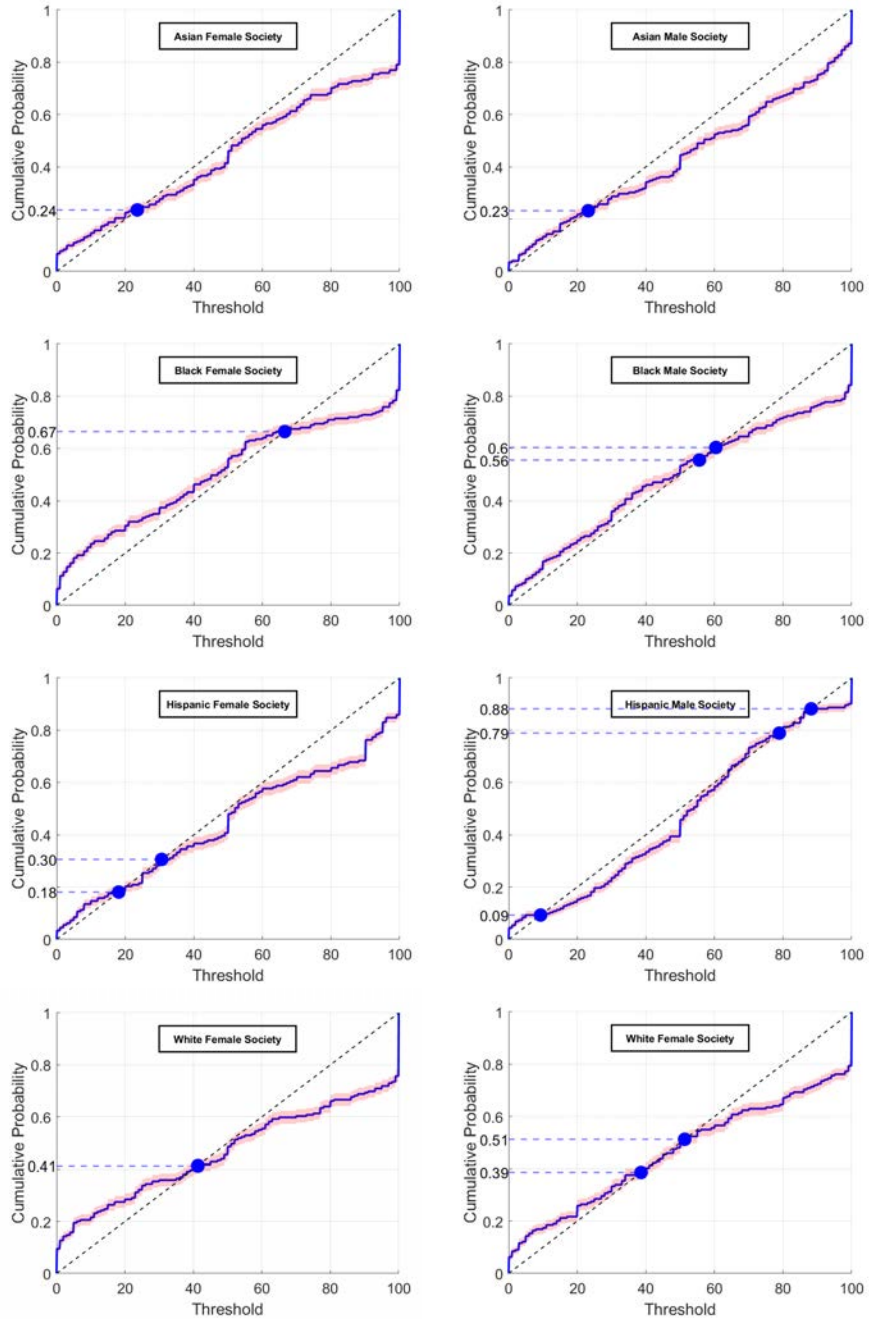
We use the elicited individual networks (i.e., the likelihood of interacting with others who share an individual's race/ethnicity/gender) to conduct network-weighted simulations. For instance, consider a Hispanic individual who reports that six out of ten of their recent interactions were same-race/ethnicity interactions, and three

Figure D.2: Threshold for AA, distributions 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 D.3: Threshold against AA, distributions 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.

out of ten were with females. Then, we evaluate this person's threshold against the actions of a society consisting of 60 other Hispanics (42 males, 18 females) and 39 non-Hispanics (race/ethnicity-representatively drawn with weights according to the gender-network variable). We find that the network-weighted simulations result in similar but less support for AA compared with the representative society. Compared with societies fully segregated by REG group, using the elicited individual networks increases support for AA for White males and decreases support for AA for Asians and Blacks.

E Instructions

The exact wording of the survey 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 Inde-

pendent?
[*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*][25]
[*Open cell, 0-10*]

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

[25]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 default, **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 default 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 default 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 default 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 default.

ANTI Affirmative action

The computer program randomly selected the **ANTI-affirmative action organization** American Civil Rights Institute (ACRI). Thus, by default, 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 default.]

PRO Affirmative action

The computer program randomly selected the **PRO-affirmative action organization** American Association for Access, Equity and Diversity (AAAED). Thus, by default, 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[26]

Determining your donation

Please read carefully

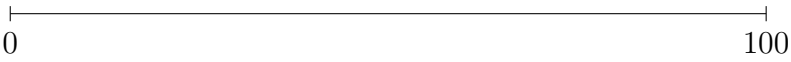
To determine whether you change your $1 donation (from the anti-affirmative action to the pro-affirmative action organization), you must **choose a number between 0 and 100**. The number indicates **how many others in your group have to change their donation to the pro-affirmative action organization** 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**[27]: your donation will change if 79 or more of the other people donate to the pro-affirmative action organization.
- ...and so on...

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

[27] 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.[28]

[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.*[29]

[28] 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."

[29] 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.*[30]

[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.*[31]

..... *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.

[30] Only shown to participants in the Public treatment.
[31] 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]

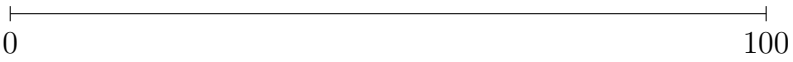
We will now place you in a group of **100 [men] [women] living in the U.S.** (you and 99 others). Group members represent the typical population of [men] [women] living in the U.S.
We will ask you one final time the question we asked you before, but this time, for the group of 100 [men] [women].

[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[32]

[32]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.



100 [men] [women]

Recall that the computer program randomly selected the anti-affirmative action organization as the default. 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.*[33]

[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].*

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

- *your email address will be posted on the website if you donate to the pro-affirmative action organization.*[34]

[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.*[35]

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*]

[34]Only shown to participants in the Public treatment.
[35]Only shown to participants in the Public treatment.

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

... 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 default-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 default-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 default-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 default-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 default-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?[36]

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

[*Open cell, 0-100*]

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 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?[37]
[*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

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

of affirmative action on the previous page?[38]
[*Open cell, 0-100*]

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?[39]
[*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[40]

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

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

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

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

AA enhances organizational performance in the long run.
[*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[41]

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

Regulations trigger a sense of resistance in me.

[41] 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".

[*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 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.[42]
[*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*]

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

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 the general US population.[43]
[*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.

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

[*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.

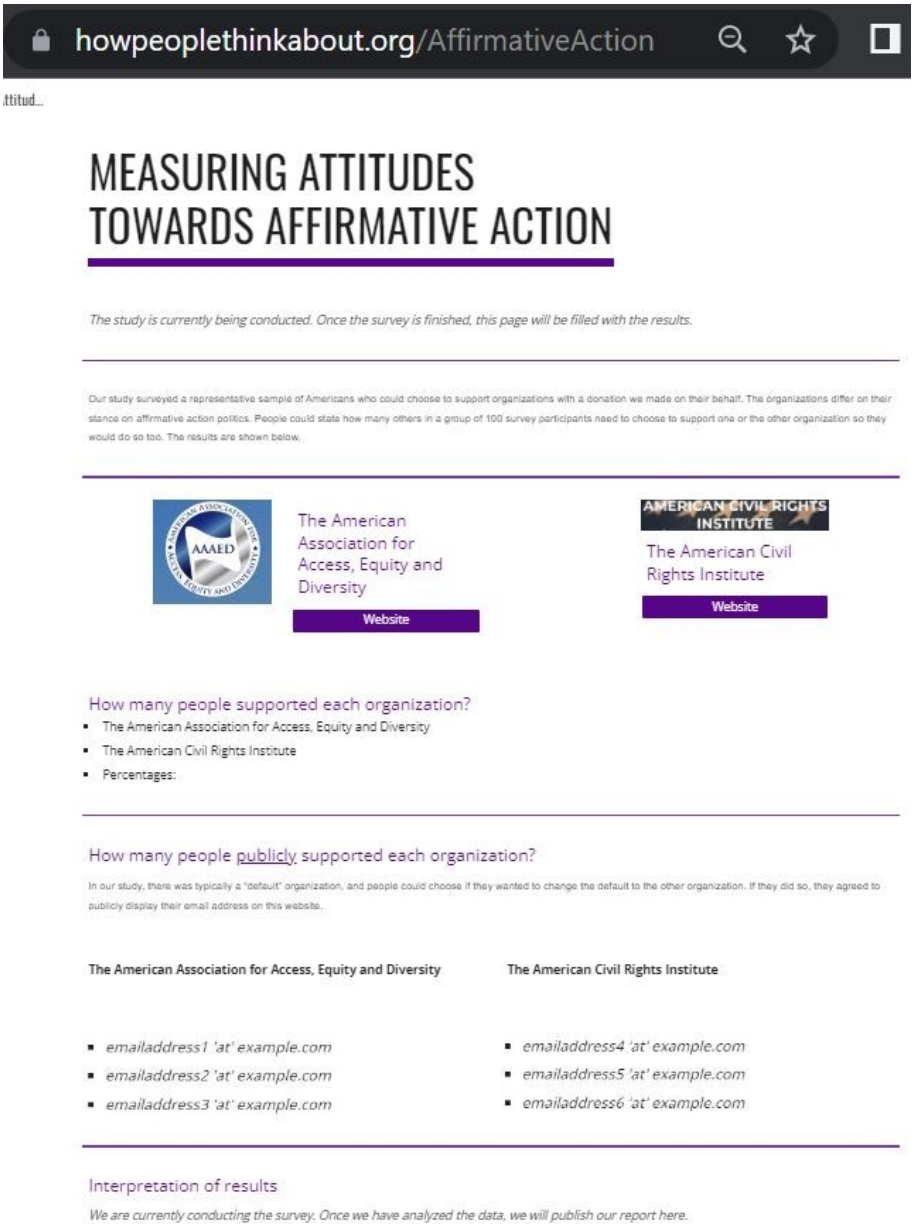
Please enter below **the same email address** to which we sent you the invitation to this survey.
Email:
[*open field*; *I prefer not to answer. I understand that this makes me ineligible for the bonus earnings (Amazon vouchers)*]

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.

F Website in Public Treatment

Figure F.1: 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 default in the Public treatment. In line with IRB requirements, the email addresses were removed after six months.