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, Revised January 2026

For helpful discussions, we thank Alexander Cappelen, Charles Efferson, Ernst Fehr, Urs Fischbacher, Simon Gächter, Sanjeev Goyal, Mark Granovetter, John List, George Lowenstein, Marco Piovesan, Andrea Robbett, Danila Serra, Bertil Tungodden, John Wooders, Leeat Yariv, and Yves Zenou. We thank seminar participants at the University of Arizona, University of Bologna, University of Chicago, University of Cologne, Florida State University, Middlebury College, New York University, Norwegian Business School, University of Texas at Dallas, University of Verona, Yale University, and the University of Zurich. We also thank audiences at Asia Meetings of the Econometric Society (2026), Zurich Workshop in Economics & Psychology (2025), the Social Norms Workshop in Ascona, the Barcelona Summer Forum, the Dynamics of Social Change Workshop at NYU Abu Dhabi, the Norms and Behavioral Change Conference at the University of Pennsylvania, the French Experimental Economic Association meeting, the Maastricht Behavioral Economic Policy Symposium, the European Economic Association meeting (all in 2024), and the Economic Science Association World and Africa meetings (2023). The project was pre-registered at AEARCTR- 0010895. IRB approval has been obtained by the NYUAD (HRPP-2022-74) and UT Dallas (IRB-22-582) Institutional Review Boards. NN gratefully acknowledges financial support from Tamkeen under NYU Abu Dhabi Research Institute Award CG005. NN and SS gratefully acknowledge financial support from the National Science Foundation (grant #2242443). 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, Revised January 2026
JEL No. C83, C90, D63, D70

ABSTRACT

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

Moritz Janas
University of Gothenburg
moritz.janas@gu.se

Simon Siegenthaler
University of Texas at Dallas
simon.siegenthaler@utdallas.edu

Nikos Nikiforakis
New York University Abu Dhabi
Department of Economics
nikos.nikiforakis@nyu.edu

1 Introduction

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

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

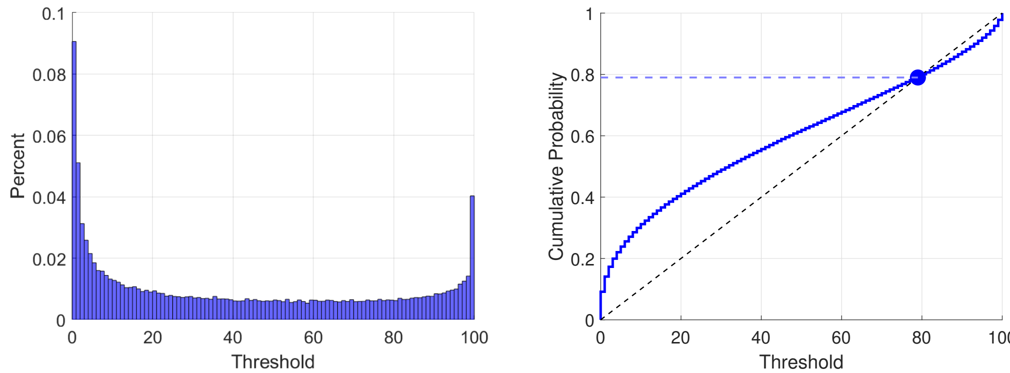
Our method builds on threshold models, which posit that individuals act when the number of others taking the same action crosses a personal threshold. These models—sometimes referred to as models of social influence (Young, 2009)—have been widely used to model how strategic complementarities in behavior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

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

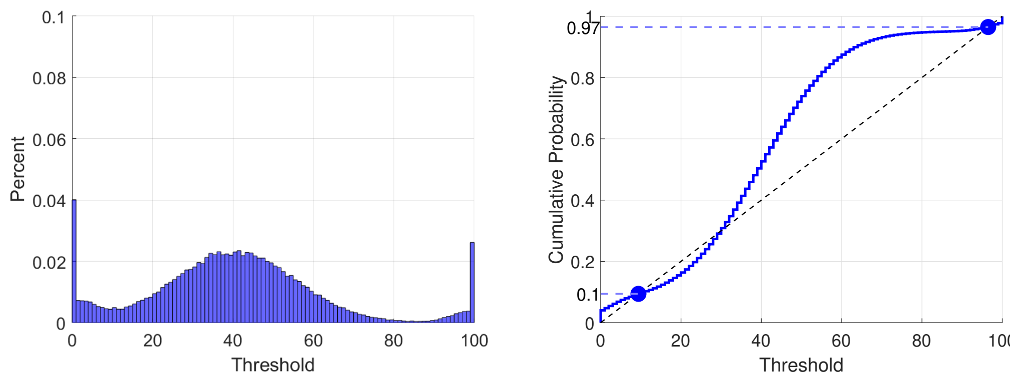
An example helps illustrate how threshold models work and sets the stage for our method. Suppose you are at a faculty meeting where a policy to adopt affirmative action (AA) in hiring is under discussion. Participants are asked to show support for the policy by raising their hands. You evaluate the policy's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

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

Figure 1: Examples of Threshold Distributions and Equilibrium Predictions



A. Unique Equilibrium



B. Two Stable Equilibria

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

similar calculations. How many faculty members will raise their hands depends on the distribution of thresholds. Assume that the distribution of thresholds is the one shown in Figure 1A: 9% of participants have a threshold of 0, meaning they will raise their hand even if no one else does, while 4% have a threshold of 100, meaning they will never raise their hand even if everyone else does. The remaining 87% exhibit interdependent behavior, raising their hands only if enough others do. The stable equilibrium occurs at the point where the cumulative distribution function (CDF) intersects the 45-degree line from above—as long as the CDF lies above the line, the best-response dynamic implies that more individuals will raise their hands, encouraging others to follow. In this example, therefore, 78% of participants are predicted to raise their hands in support of the proposed policy.

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

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

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

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

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

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

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

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

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

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

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

2 Theoretical framework

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

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

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

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

model individual utility as:

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

where:

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

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

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

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

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

where

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

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

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

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

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

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

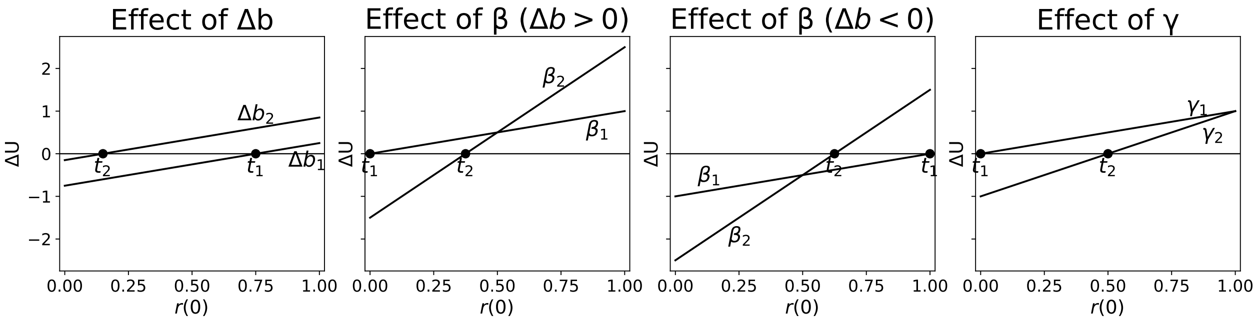
Changes in the social-alignment parameter β pull thresholds toward the midpoint. When individuals are more favorable to a ($t_i^{a,*} < 50\%$), higher β raises their threshold, $\frac{dt_i^{a,*}}{d\beta} > 0$; when they are less favorable ($t_i^{a,*} > 50\%$), higher β lowers it, $\frac{dt_i^{a,*}}{d\beta}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

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

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

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

Figure 2: Effects of Model Parameters on Optimal Threshold



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

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

3 Experimental design

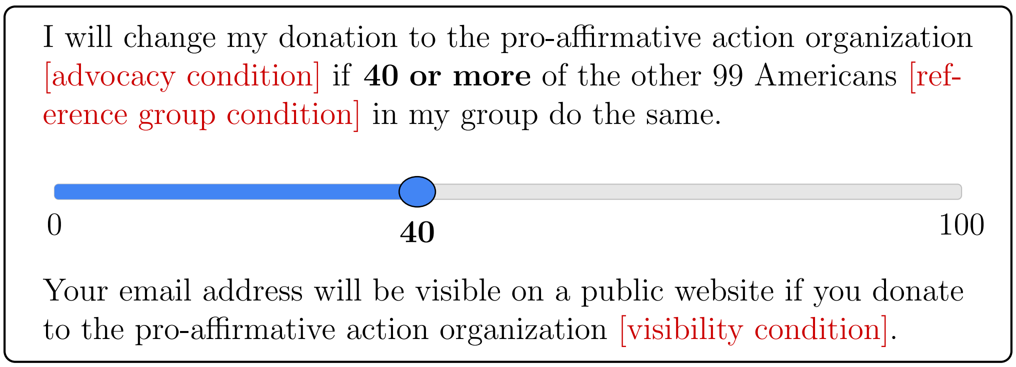
3.1 Decision context

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

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

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

Figure 3: Threshold Elicitation Interface



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

3.2 Experimental treatments and behavioral hypotheses

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

Table 1: Experimental Design

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

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

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

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

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

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

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

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

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

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

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

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

3.3 Sample and Procedures

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

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

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

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

Table 2: Sample

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

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

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

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

4 Empirical analysis

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

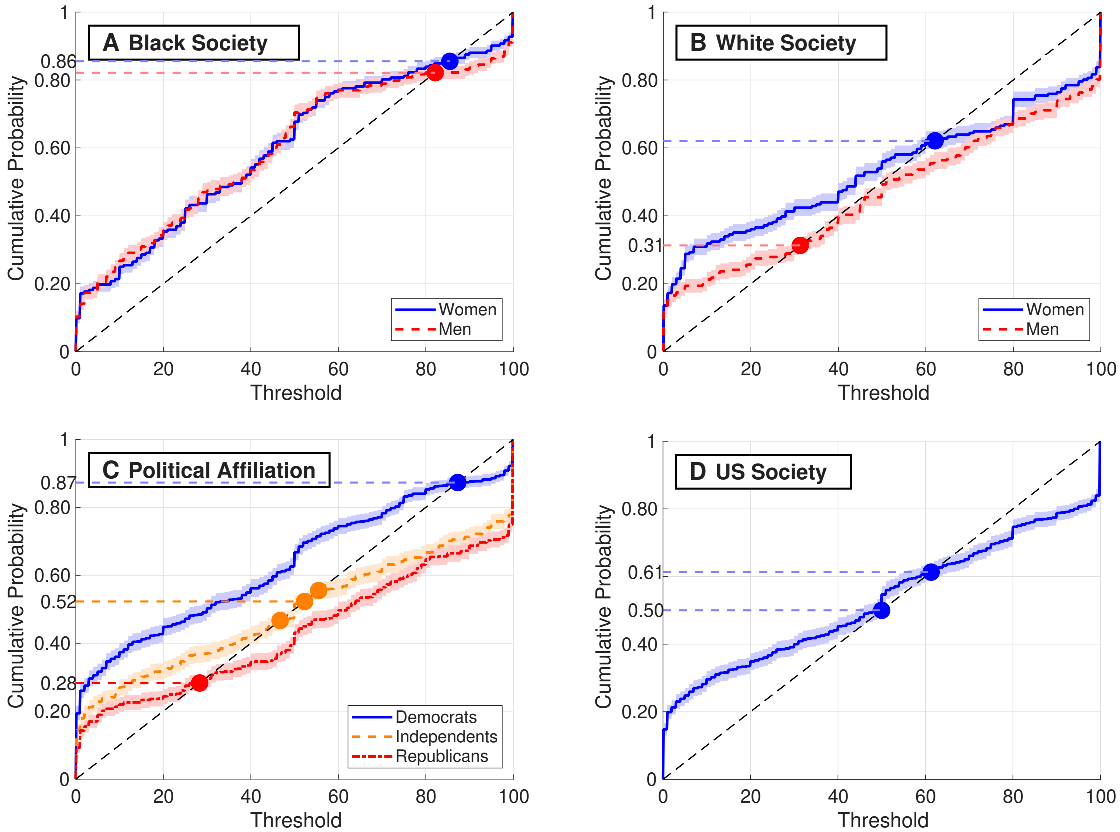
4.1 Overview of threshold distributions

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

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

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

Figure 4: Selected Threshold Distributions



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

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

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

4.2 Tests of behavioral hypotheses

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

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

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

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

Table 3: Perceived Benefits and Social Pressure Shift Threshold Levels

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

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

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

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

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

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

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

Table 4: Reference Groups and Conformity Shift Thresholds Inward

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

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

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

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

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

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

4.3 The role of forward-looking beliefs

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

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

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

Table 5: Leading Change—Structural Estimates of Model Parameters

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

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

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

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

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

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

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

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

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

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

5 Thresholds and institutional change

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

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

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

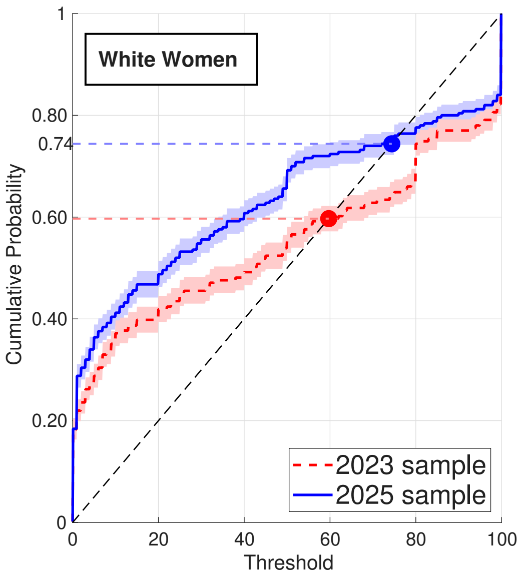
To formalize this intuition, recall that in our baseline model thresholds depend negatively on the perceived marginal benefit of action: as Δb_i increases, t_i decreases (Figure 2). We extend the model by allowing perceived benefits to depend on the prevailing institutional environment. Let S denote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

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

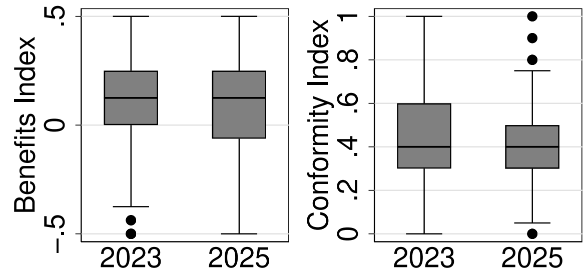
When formal institutional support weakens (a decline in S), concavity implies that $B_i'(S)$ increases, raising $\Delta b_i(S)$ and thereby lowering thresholds for pro-AA advocacy—even if absolute attitudes remain unchanged.

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

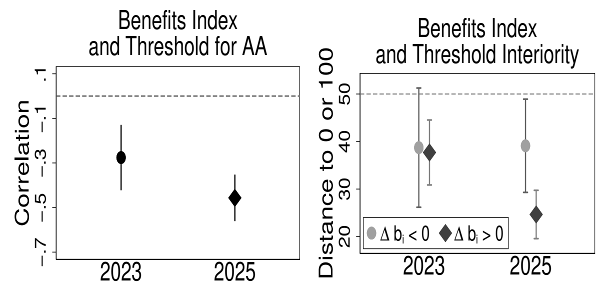
Figure 5: Changing Thresholds Despite Stable Preferences



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



B. Stable Preferences



C. Stronger Mapping

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

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

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

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

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

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

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

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

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

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

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

6 Using threshold data to anticipate changes in aggregate outcomes

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

Targeting near the equilibrium

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

Figure 6A provides an example by showing the distribution of t^{AA} after a targeted intervention that reduces thresholds among respondents identified as Republican or Strong Republican or among an equally-sized group of Democrats or Strong Democrats. In both cases, thresholds in the targeted group are lowered by 50 points, a magnitude that, under the structural estimates, corresponds to an increase in the benefits index to its maximum level.

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

Multiple equilibria and tipping points

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

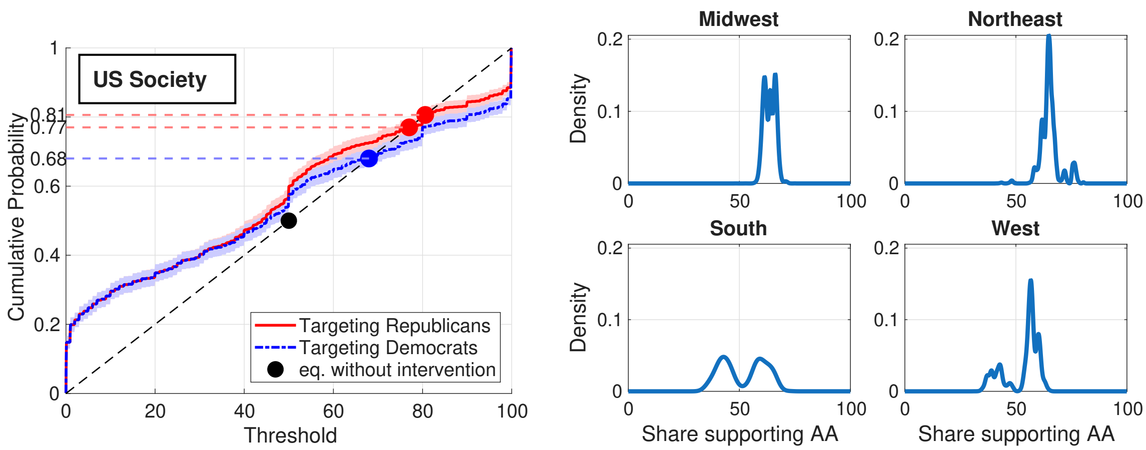
Figure 6b visualizes this logic using different U.S. regions. For each region, we simulate 10,000 groups, record all equilibrium outcomes, and plot the resulting density of equilibrium shares supporting affirmative action. The Northeast, for example, exhibits a single, tightly concentrated mass point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

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

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

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

Figure 6: Designing Effective Interventions



A. Targeted Intervention B. Empirical Equilibrium Structure

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

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

Social networks

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

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

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

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

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

7 Conclusion

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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