A Model of Norm Change in Polarised Societies

Edward Brookman

Student ID: 20569248

Supervisors: Silvia Sonderegger

and Toomas Hinnosaar

Abstract

People are willing to abandon a social norm, as long as others do first. From this fact emerges the possibility of social tipping: a sudden, self-reinforcing shift in collective behaviour triggered by the actions of a few early-adopters. While this phenomenon has been shown to dismantle harmful norms, many real-world norms benefit one population group at the expense of another. Women benefit more than men from changing gender norms, just as Democrats and Republicans have conflicting preferences over various political norms. When willingness to change is grouped and polarised in this way, should we still expect to see tipping dynamics? This paper develops a game-theoretic threshold model with population groups that have opposing preferences over coordination equilibria. We find this conflict generates a bimodal distribution of adoption thresholds, giving rise to a false tipping point and an interior equilibrium of costly and persistent miscoordination. Consequently, we show that standard, well-meaning tipping interventions can backfire, leading the system into a 'polarisation trap'. We derive conditions under which such traps emerge and show how targeted interventions must account for group resistance to avoid these welfare-destroying outcomes.

Contents

1 Introduction **3**

2 Literature Review **4**
2.1 Social Norms for Economists 4
2.2 History of Threshold Models 5
2.3 Threshold Models Applied to Social Norms 6

3 The Model **8**
3.1 Agents and Groups 8
3.2 Actions, Payoffs, and Preferences 8
3.3 Agent Decision Problem and Individual Thresholds 10
3.4 Distribution of Thresholds 11

4 Equilibrium **11**
4.1 Stability of Equilibria 12
4.2 Existence of Equilibria 13

5 Consequence of Conflicting Group Interests **14**
5.1 Welfare in the Polarisation Trap 15
5.2 Conditions for Polarisation Trap 18

6 Social Change in Polarised Environments **21**
6.1 Ignition 21
6.2 Conversion 23

7 Interventions **23**
7.1 Minority Subsidy 24
7.2 Majority Subsidy 26
7.3 Mixed Subsidy 27
7.4 Optimal Policy 28

8 Conclusion **29**

Bibliography **30**

Appendix **33**

1 Introduction

Social norms are notoriously stable. Preferences for conformity, fear of social sanctions and coordination incentives all permit even harmful norms to persist. Yet, history has shown that norms *do* change, often rapidly and unexpectedly. In the late 1980s, 74% of people were reported to smoke indoors without asking permission. By the turn of the century, the arrival of anti-smoking norms had caused the same figure to plummet to just 10% (Nyborg & Rege, 2003). Today, the act is almost unthinkable. Similarly, public support for same-sex marriage in the US surged to a majority position in just fifteen years, despite being frowned upon for large swathes of western history (Rosenfeld, 2017).

Understanding and artificially replicating this rapid process of norm change has been a key problem in social science. Interventions are increasingly being proposed and tested in order to establish social norms that are beneficial to society (Berger, 2021; Constantino et al., 2022; Efferson et al., 2020; Otto et al., 2020; Pizziol & Tavoni, 2024). Particularly appealing to policy-makers is the concept of social tipping: once a system reaches a critical mass of individuals breaking the status quo, preferences for conformity rapidly cascade the remainder of the population. This phenomenon is a robust finding, having been identified theoretically (Baronchelli et al., 2006), observationally (Jones, 2009) and experimentally (Centola et al., 2018). By exploiting this mechanism, policy-makers can design small interventions that nudge a system beyond its tipping point and return population-wide behavioural change.

However, the excitement surrounding social tipping has been met with caveats. Recent experimental work has demonstrated that tipping interventions can fail when group identities become salient (Ehret et al., 2022). This finding challenges a key assumption underlying the standard model: that individual preferences for norm change are smoothly distributed throughout the population. In reality, many of today's most contentious social issues are characterized by distinct demographic groups with sharply different preferences over potential norms. Men and women exhibit different propensities to adopt gender equality norms; Republicans and Democrats have opposing views on gun-control norms. Indeed, recent empirical evidence finds that black females and white males have drastically different threshold distributions for supporting Affirmative Action (Janas et al., 2024). When societies are polarised along group lines, can social tipping still facilitate norm change, or do standard interventions risk creating new problems?

This paper constructs a game-theoretic threshold model populated by a minority and majority group. Groups have incentives to coordinate, but posses conflicting preferences over two potential norms. We show that the resulting distribution of adoption thresholds is bimodal. As a consequence, our model can create a 'false tipping point' that leads to a welfare-destroying interior equilibrium. In such cases, standard attempts to tip the system to an alternative norm prove not only inefficient, but harmful. We use the model

to derive the optimal policy under group polarisation, showing that effective interventions must target resistant group members to create bridges between opposing factions.

2 Literature Review

2.1 Social Norms for Economists

> "All such social norms or conventions can be regarded as equilibrium selection devices in the game of life that we play together." (Binmore, 2022)

Social norms pervade nearly all social interaction. Standing in line, picking up our litter and wearing ties are all things we do, not because we enjoy them, but because we care about conforming to the social norms of our society. As Binmore's quote suggests, social norms can be well-understood through the lens of game theory. Although the game of life offers many possible actions, we often benefit most when we all take the same action. Consider greeting someone for the first time in a business setting. There are many gestures we can initiate; a hug, a high-five, a kiss. Thankfully, we would all agree that extending our right hand is the correct action to take. Here, a social norm has selected the coordination equilibrium for us, allowing for quick, effective communication and preventing an in-the-moment fumble as both parties attempt to coordinate on a greeting.

Similarly, Douglas North conceptualised norms as *institutions* that structure social and economic interactions, often solving problems that formal contracts cannot (North, 1991). By creating shared expectations, they reduce uncertainty and transaction costs, allowing for smoother coordination in markets and organisations. However, norms possess power beyond simply selecting between customs. They also sustain cooperation in situations where narrow self-interest would otherwise lead to social dilemmas. The threat of social sanctions can effectively constrain self-regarding behaviour and enforce pro-social outcomes (Fehr & Fischbacher, 2004). Social sanctions include direct punishment (such as an angry confrontation) but also the retraction of benefits such as party invites or favours. Games of cooperation are thus *transformed* into coordination games by social norms, since no one person can deviate to uncooperative behaviour without incurring strong social costs.

While the equilibrium properties of social norms are well understood, their dynamic properties (transitions, emergence, etc.) remain less studied. It has been observed that social norms are not formed by a central authority, nor are any of the individual players aware that they are forming them (Young, 2015). Instead, they exist as a *group-level property* emerging from the interactions between many self-interested individuals. A popular analogy is the emergence of paths through a field (Ball, 2006). No single person decided where the path should go, but through the repeated choices of walkers, a stable

route emerges that future walkers feel a pressure to follow. While modelling the emergence of paths is relativity simple; situations governed by social norms are often more complex, with a wide range of individual motivations. To deal with this complexity, a simple decision rule is often assumed; the agent will abandon their current behaviour in favour of an alternative as soon as $X\%$ of the population has switched too. This logic forms the basis of a diverse range of models, collectively grouped under the term 'threshold models'.

Since classical game theory is silent on the topic of equilibrium selection, dynamics are often introduced by assuming that agents react *myopically* to their environment. They do not try to anticipate or comprehend the dynamic process that they are a part of, they simply take the current state of the world as given and make their decision accordingly. The remainder of this review details the history of threshold models, their empirical validation, and what we have learned from them thus far.

2.2 History of Threshold Models

Many economists have likely already been introduced to threshold models; Thomas Schelling's (1969) model of segregation uses the threshold logic explicitly. Individuals, represented as coloured cells on a grid, prefer that at least three of their eight neighbours are the same colour. In each round, they move only if this minimal threshold is not met. Over time, these seemingly tolerant, individual thresholds aggregate to produce a surprising emergent property: highly segregated neighbourhoods. Crucially, the macro-state does not follow intuitively from the micro-preferences. Observing the final, hyper-segregated grid, one would wrongly assume it was caused by individuals with a strong desire for segregation. Schelling (1978) constructed and explored many more of these models in his book 'Micromotives and Macrobehaviours'. One common theme across models was the existence of 'tipping point' dynamics; when individual decisions are interdependent, the actions of a few can create a chain reaction that rapidly alters the behaviour of the entire group. In game-theoretic terms, threshold models produce a *punctuated equilibrium* effect — where systems remain stable for long periods of time, before rapidly transitioning toward another equilibrium (Young, 2015).

Building on Schelling's logic, Granovetter (1978) formalized these ideas in his threshold models of collective behaviour. He used the example of a potential riot to model how participation unfolds. In his model, individuals will join a riot so long as a certain number of other agents have already joined, a number which varies for each person. Granovetter's key conclusion was that the final outcome is critically dependent on the distribution of these thresholds. He showed that if a crowd of 100 people had a perfectly uniform distribution of thresholds (0, 1, 2,...99), a complete "domino effect" would occur, resulting in a full-scale riot despite the average crowd member being not particularly disposed to

rioting. However, if that distribution was perturbed just slightly – for instance, by lacking a single person with a threshold of 1 – the cascade would break, and the riot would fizzle out with only one participant. Observing the collective outcome in this case would mislead the observer into thinking a riot was far from occurring. Later, Macy and Tsvetkova (2015) would refine the model by introducing stochastic thresholds, eliminating some of the implausible equilibria from the original work. Counter-intuitively, introducing realistic noise in this case makes system behaviour *more* predictable since the outcome is less reliant on small gaps in the threshold distribution.

Recognising the importance of the underlying distribution of thresholds, Yin (1998) derived equilibrium outcomes in a variety of distribution shapes. Particularly relevant to this paper, is the bimodal case. Yin found that bimodal societies are less susceptible to the sudden, unexpected revolutionary cascades seen in normal distributions and may settle into a state of partial mobilization. Yin's general conclusions about equilibria can also be seen in my model, as the conflicted group interests in my underlying coordination game produce a bimodal distribution of thresholds.

2.3 Threshold Models Applied to Social Norms

Since social norms can be thought of as equilibria of coordination games, and threshold models describe transitions between equilibria, it follows naturally that they should be applied to the more-modern topic of social norm change. The following papers do just this, as well as going further to test their predictions in the laboratory. Alongside the shift in application, the recent wave of threshold models have typically come from the field of economics as opposed to sociology. As a consequence of this, modern-day threshold models tend to be micro-founded, with an explicit, underlying stage game. Agent's thresholds are thus calculated from an expected utility maximising problem, as opposed to the author starting from an assumed distribution of thresholds (Ye & Zino, 2025).

To explore the potential of threshold models for norm change, Andreoni et al. (2021) developed a theoretical framework and a corresponding large-scale laboratory experiment designed to observe social tipping and norm change in a controlled setting. The authors model a society stuck in one coordination equilibrium attempting to transition to a welfare improving one. As per the threshold logic, individual's willingness to deviate from the norm depends on the proportion of others who have already done so. Individuals are heterogeneous, possessing different thresholds due to varying preferences over outcomes, risk tolerance, and dislike for conformity. From the payoff structure, they calculated the 'tipping threshold' – a measure of the neutral person's threshold for change. Through computer simulations, they found that societies below a critical tipping threshold (35%) had a near 100% probability of tipping to the welfare-improving norm. Conversely, in societies where the miscoordination penalty was too large (causing the tipping point to

be above 35%), the society remained locked in the sub-optimal norm.

To test their model's prediction, the authors introduced the "Social Tipping Game", which was designed to recreate the essential features of norm change. In multi-period interactions, societies of individuals played a coordination game with two options: Blue or Green. Initially, a Blue norm was established by inducing all participants to prefer Blue. This norm was reinforced by a penalty for miscoordination, with the cost being disproportionately high for the first individuals to deviate from the norm. The impetus for change was then introduced by having individuals' preferences gradually and stochastically switch from Blue to Green over time, a process that was public knowledge to all participants. As more individuals came to prefer Green, the Blue norm became detrimental to the group's welfare. In 96% of their experimental societies, their model predicted whether the participants would successfully establish the new Green norm, or retain the unanimously-harmful Blue norm. This paper is important in validating threshold models and providing an experimental game to test interventions.

In a similar experimental setup, Ehret et al. (2022) tested the effect of group identities on social tipping. In a neutral treatment, the success of using social tipping to establish a new norm was replicated. However, when the coordination options were given political labels ('Triumphant Biden' or 'Triumphant Trump'), tipping did not establish a new norm despite no change in the material incentives. Instead, the original norm was destroyed and replaced with persistent, costly miscoordination. This finding raises the central question: why do social tipping interventions, which are effective in standard settings, fail or even backfire in polarised ones, and what are the underlying dynamics that lead to a state of stable conflict instead of a new consensus? My paper attempts to answer these questions, extending the domain of the threshold model and deriving new insights about how to achieve social change in a polarised world.

Using threshold models, Efferson et al. (2020) and Efferson et al. (2024) theoretically test interventions to facilitate the transition between a harmful status quo norm and an alternative one. Policies involved making the alternative action more appealing for targeted individuals and were evaluated by the amount of endogenous change created beyond the direct effect of the intervention. The former paper, assuming a realistic scenario of heterogeneous responses to the intervention, established that the creation of spillovers is complex and hard to predict. Because individuals who are more resistant to change are also less likely to respond to the intervention, they found that it was most effective to target a representative subset of the population. In contrast, the latter paper's analytical model makes a simplifying assumption: that the intervention is perfectly and uniformly effective on all targeted individuals, regardless of their initial resistance. Under this condition, targeting the individuals most resistant to the new norm often produces the greatest overall shift in behaviour. The logic is that the perfectly effective intervention handles the most difficult segment of the population directly, leaving the most amenable

individuals as the non-targeted group.

Lastly, Janas et al. (2024) go a step further and actually elicit participants' individual thresholds for action. The authors directly measured these thresholds in a large, stratified US sample, using the politically salient topic of Affirmative Action. Their method involved participants stating the number of others who would need to support a change in a charitable donation (from an anti- to a pro-Affirmative Action organization, or vice-versa) before they would do so themselves. The study provides two crucial findings that bridge the gap between theory and empirical evidence. First, it offers powerful validation for the core assumption of the entire literature: a large majority of individuals across all demographic groups do, in fact, condition their choices on the actions of others, exhibiting interdependent, threshold-based behaviour. Second, the authors show that these elicited thresholds are predictable and systematically shaped by an individual's perceived benefits from the policy and the social pressure to conform. This work represents a significant methodological advance, providing a tool that can be used to measure the underlying distribution of preferences in a population, which in turn can be leveraged to better predict social outcomes.

3 The Model

This section constructs a model of norm adoption in a society comprised of two groups with conflicting interests. In each period, agents observe the aggregate behaviour from the previous period and choose an action to maximize their own utility in a coordination game.

3.1 Agents and Groups

The set of agents is a continuum, which we normalize to have measure one. Each agent i belongs to one of two observable, exogenously assigned groups: a **majority** group, denoted by uppercase M, or a **minority** group, denoted by lowercase m. We let $g_i \in \{M, m\}$ be the group identity of agent i. The population share of the Minority group is $\lambda \in (0, 1/2)$, and the share of the Majority group is $1 - \lambda$.

3.2 Actions, Payoffs, and Preferences

The Stage Game. In each discrete time period $t = 0, 1, 2, ...$ agents simultaneously choose one of two actions: adhere to the status quo (S) or adopt an alternative (A). After choosing their actions, agents are randomly matched into pairs. An agent receives a positive payoff only if their partner chose the same action; otherwise, the payoff is zero. The payoff matrix for an interaction between agent i and agent j is:

		Agent j	
		Status Quo	Alternative
Agent i	Status Quo	$U^S_{g_i}, U^S_{g_j}$	$0, 0$
	Alternative	$0, 0$	$U^A_{g_i}, U^A_{g_j}$

The coordination payoff an agent receives is dependant on their group. We are interested in societies where groups have **conflicted interests**. While both groups prefer to coordinate, they disagree on the ideal norm. The Majority group derives greater utility from coordinating on the status quo equilibrium, while the Minority group prefers the alternative. This is formalized by the following assumptions:

$$U^S_M \geq U^A_M > 0, \tag{1}$$

$$U^A_m \geq U^S_m > 0 \tag{2}$$

In subsequent sections, we normalise the payoff from each group's preferred equilibrium, U^S_M and U^A_m, to 1 for tractability. As such, I denote majority resistance as R where $R = 1 - U^A_M$. A value of $R = 0$ signifies that the majority is indifferent between the status quo and alternative norm. As R increases, the majority will increasingly suffer from and resist the alternative norm. We also restrict analysis to parameters which satisfy:

$$\lambda U^A_m + (1-\lambda)U^A_M > \lambda U^S_m + (1-\lambda)U^S_M \tag{3}$$

This ensures that population-wide coordination on the alternative equilibrium provides a greater population welfare than the status quo equilibrium. If this were not the case, there is little reason for a policy maker to intervene.

Agent Types and Utility. In addition to the stage game payoff, each agent has a privately known, idiosyncratic preference for the status quo. This is captured by a utility shock, η_i, which is realized with certainty if the agent chooses action S. This term can be positive or negative and captures an individual's unique political leanings, personal tastes, or moral convictions that are independent of social influence.

The shocks η_i are drawn independently from a normal distribution with a mean of zero and a group-specific variance: $\eta_i \sim N(0, \sigma^2_g)$. An agent's per period utility from choosing a_i is the sum of the stage game payoff and this idiosyncratic shock:

$$U_i(a_i, a_j) = u_i(a_i, a_j) + 1_{\{a_i = S\}} \cdot \eta_i \tag{4}$$

Where u_i is the coordination payoff from the table above.

3.3 Agent Decision Problem and Individual Thresholds

Let p_t represent the *actual* proportion of agents choosing A in period t. To make their decision, agents must form a belief about this proportion, which we denote by p'_t. Agents form **myopic expectations**; they observe the proportion of the population who chose A in the previous period (p_{t-1}), and believe that the same fraction will choose A in the current period ($p'_t = p_{t-1}$). This prescribes a relatively low degree of sophistication to agents – they do not predict or anticipate trends in behaviour. However, this assumption of non-strategic behaviour is reasonable within the model's framework. Because the agents form a continuum, each one correctly anticipates that their individual choice has a negligible influence on the aggregate outcome. This mirrors how people often reason about societal-level issues such as climate change, where one person's action feels insignificant, making a reactive decision rule the most plausible approach.

Given their belief, the expected utility for an agent i of group g_i for each action is:

$$E[U_i(S)] = (1 - p'_t)U^S_{g_i} + \eta_i \tag{5}$$

$$E[U_i(A)] = p'_t U^A_{g_i} \tag{6}$$

Agents maximise their expected utility, choosing A if $E[U_i(A)] > E[U_i(S)]$. This implies that for each agent, there is a critical proportion of alternative adopters that causes them to be indifferent between the two actions. We define this as their **individual threshold**, x_i. By setting the expected utilities equal, we can solve for this threshold:

$$x_i = \frac{U^S_{g_i} + \eta_i}{U^A_{g_i} + U^S_{g_i}} \tag{7}$$

Since the expected utilities are monotonic in p', the agent's optimal strategy is a simple threshold rule: choose A if enough people chose A in the previous period to meet or exceed their individual threshold ($p_{t-1} \geq x_i$), and choose S otherwise.

3.4 Distribution of Thresholds

Since the thresholds x_i are a linear transformation of the normally distributed shocks η_i, the thresholds themselves are normally distributed within each group. For any group g, the distribution of thresholds is $x_g \sim N(\mu_g, s_g^2)$, where the mean μ_g and standard deviation s_g are given by:

$$\mu_g = \frac{U_g^S}{U_g^A + U_g^S}, \quad s_g = \frac{\sigma_g}{U_g^A + U_g^S} \tag{8}$$

Let $F_g(x)$ be the cumulative distribution function (CDF) of thresholds for group g. The population-wide CDF of thresholds, $F(x)$, is the weighted average of the group CDFs:

$$F(x) = \lambda F_m(x) + (1 - \lambda) F_M(x) \tag{9}$$

4 Equilibrium

Figure 1 shows the population CDF of thresholds and the 45 degree line. It is an important graph in the analysis of threshold models since it can be used to visualise the dynamics of the system.

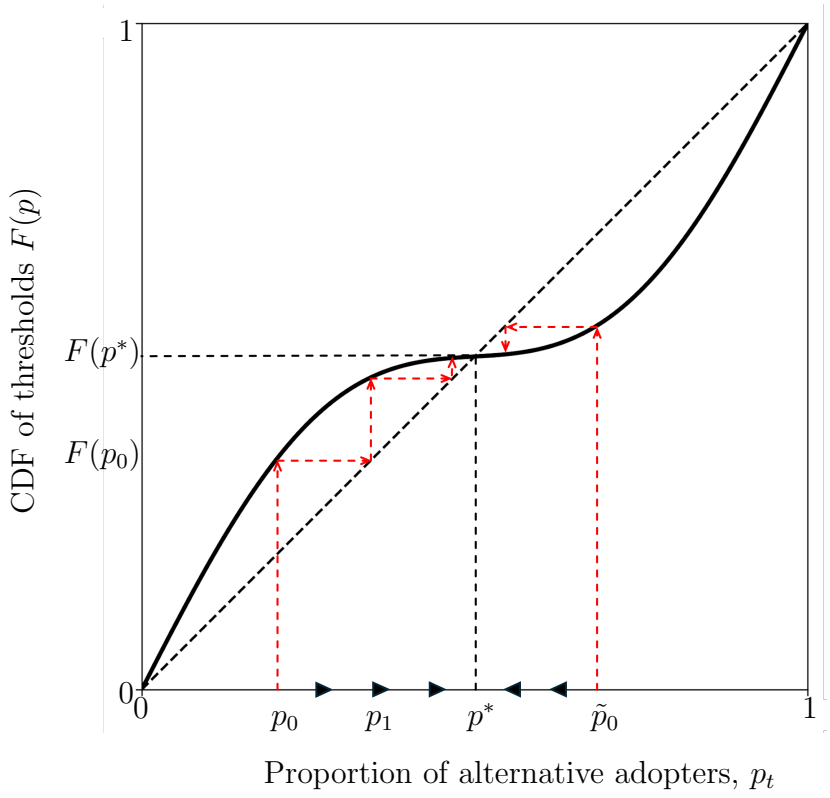


Figure 1: Thresholds (x axis) and their cumulative density (y axis). The graph can be used to analyse the dynamics and find the steady state of the system, given an initial p_0.

Consider an initial proportion of agents choosing the alternative action, p_0. In period 1, agents observe p_0 and consequently $F(p_0)$ number of agents have their threshold triggered. This gives next period's number of adopters, p_1, and so on. The evolution of p can be written as a simple difference equation:

$$p_{t+1} = F(p_t) \tag{10}$$

Whenever the system is at a point on the curve *above* the 45 degree line, the alternative action will gain adoption. Whenever, the system is *below* the 45 degree line, more agents will adopt the status quo. Because the CDF is non-decreasing, the sequence of p_t is monotonic and bounded by $[0, 1]$. As such, by the monotone convergence theorem, there can be no loops and the system is guaranteed to converge to a steady-state.

An **equilibrium** is therefore a proportion of alternative adopters $p^* \in [0, 1]$ such that if all agents expect a proportion p^* to choose the alternative action, the actual proportion of agents for whom it is optimal to choose the alternative action is p^*. Formally, it is a fixed point of the aggregate threshold CDF where:

$$p^* = F(p^*) \tag{11}$$

Graphically, an equilibrium occurs whenever the CDF crosses the 45 degree line.

4.1 Stability of Equilibria

While all fixed points are equilibria, we look to their Lyapunov stability to determine their long-run feasibility as a steady-state. An equilibrium p^* is deemed stable if small deviations from p^* lead back to p^*. Formally, this is true when $F'(p^*) < 1$. Graphically, an equilibrium is stable if the CDF crosses the 45 degree line from above. The interior equilibrium in figure 1 is Lyapunov stable, with both p_0 and $\tilde{p}_0$ leading to p^*.

An equilibrium is defined as unstable if small deviations from p^* lead away from p^* toward a different equilibrium. This is the case when the CDF cross the 45 degree line from below ($F'(p^*) > 1$). In fact, we tend not to think of these as equilibria at all. Instead, they are often referred to as 'tipping points'. This is because, an infinitesimal increase in p beyond the tipping point, will often result in a large cascade of adoption. Therefore, for a policy maker to transition the system between two stable equilibria, it is necessary only to push the system beyond the tipping point. After this, the endogenous feedback loop does the remainder of the work.

4.2 Existence of Equilibria

Proposition 1. *At least one equilibrium $p^* \in (0, 1)$ is guaranteed to exist.*

Proof. Define the function $g(p) = F(p) - p$ on the continuous and closed interval $[0, 1]$. An equilibrium is a root of this function. We evaluate the function at the interval's endpoints. At $p = 0$, we have $g(0) = F(0) > 0$, because the idiosyncratic shock η_i is drawn from a normal distribution with an infinite negative tail, ensuring a positive measure of agents have thresholds $x_i \leq 0$. Conversely, at $p = 1$, we have $g(1) = F(1) - 1 < 0$, as the distribution's infinite positive tail ensures a positive measure of agents have thresholds $x_i > 1$, meaning $F(1) < 1$. Given that $g(p)$ is continuous on $[0, 1]$ with $g(0) > 0$ and $g(1) < 0$, the Intermediate Value Theorem guarantees the existence of at least one $p^* \in (0, 1)$ such that $g(p^*) = 0$. This proves the existence of an equilibrium. □

Proposition 2. *A system governed by the aggregate threshold CDF $F(p)$, defined as a mixture of two normal distributions, can have **at most** five equilibria, of which at most three can be stable.*

The proof of this is provided in appendix A. To summarise, the system will have at least one stable equilibrium and at most three. Between each pair of stable equilibria must necessarily lie an unstable tipping point.

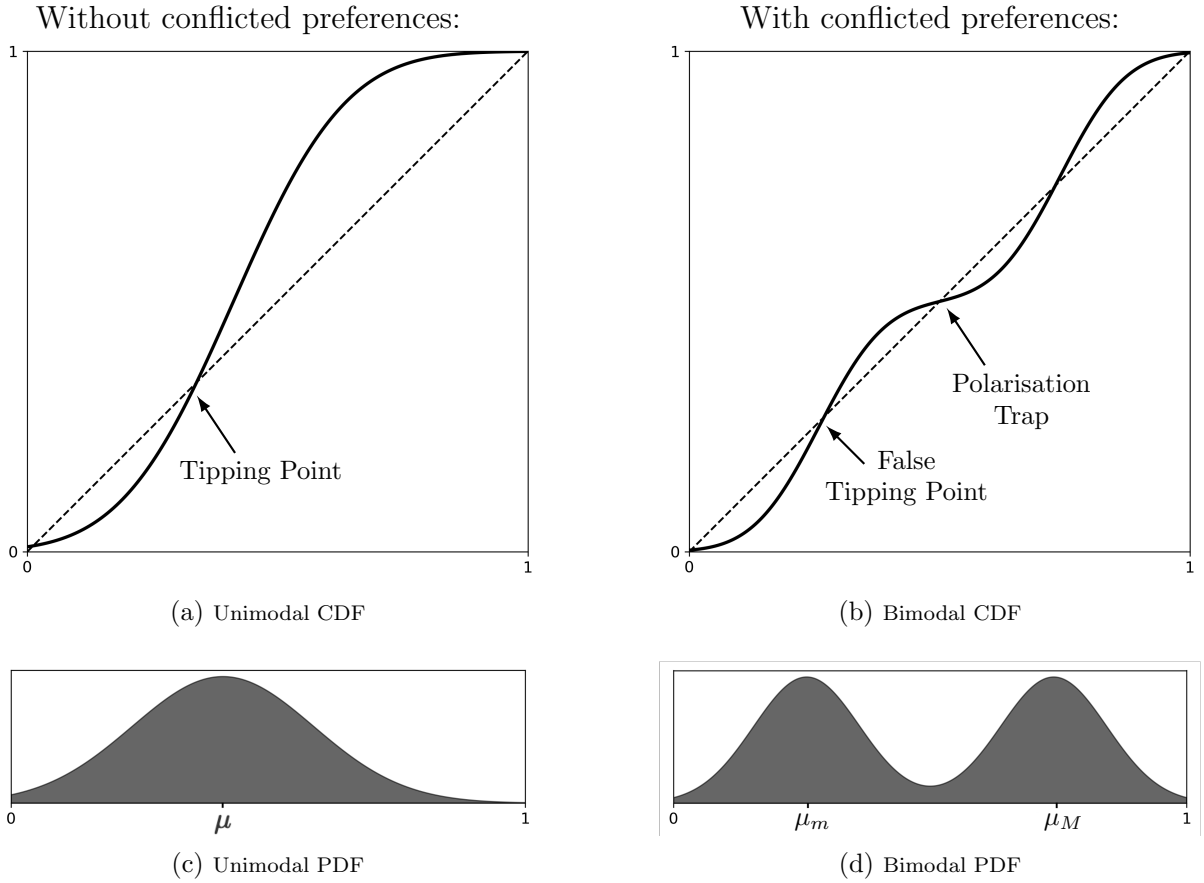
For the remainder of this paper, we refer to the stable, low-p equilibrium ($p \approx 0$) as the status quo norm, since almost-all agents are choosing the status quo action and almost-all agents are *expected* to select it in the next period. Similarly, we refer to the stable, high-p equilibrium ($p \approx 1$) as the alternative norm. Once established, the norms are enforced by the self-fulfilling expectations of the agents.

5 Consequence of Conflicting Group Interests

Under our assumptions of conflicting group interests, the minority group will, on average, have low thresholds for adopting the alternative norm and the majority group will, on average, be resistant:

$$\mu_m < \frac{1}{2} < \mu_M \tag{12}$$

Since the population distribution of thresholds is a mixture of two normal distributions with different means, the result is a bimodal distribution. Compared to the standard unimodal distribution, the bimodal case introduces the possibility of a third stable equilibrium:



(a) Unimodal CDF

(b) Bimodal CDF

(c) Unimodal PDF

(d) Bimodal PDF

Figure 2: **Left:** Unimodal distribution of thresholds creates two stable equilibria separated by a tipping point. **Right:** Bimodal distribution of thresholds creates three stable equilibria. The interior equilibrium is a state of costly disagreement.

The gap between the two modes in the PDF with conflicted interests creates a kink in the population CDF. When the kink crosses the 45 degree line, a stable interior equilibrium is created – which we refer to as the 'polarisation trap' herein. Compared to the standard unimodal case, the first tipping point is 'false' since reaching it cascades the system into the polarisation trap as opposed to the alternative norm.

Consider what a society looks like in the polarisation trap. The minority mass have had their thresholds triggered and are choosing the alternative action each period. The majority group are yet to have their thresholds triggered and are choosing the status quo action every period. Despite the underlying game being one of coordination, the society is trapped in a state of social gridlock, fraught with frequent miscoordination. This state is stable and will persist, since all agents interact with their *own* group sufficiently often to dissuade them from changing their behaviour.

To see an empirical example of this, we can look to the recent laboratory experiment in Ehret et al. (2022). In their study, they tested a society of participants engaged in a repeated, pure coordination game with two choices. When the two actions were labelled neutrally, a norm was quickly established, consistent with the predictions of the unimodal distribution from standard models. However, when labelling coordination choices with partisan images, they found that societies converged on a 'state of costly disagreement'.

The partisan labelling altered the perceived payoff structure for two demographic groups, Democrats and Republicans. In the language of our model, the psychological cost for a Democrat to choose a 'Triumphant Trump' image translates to a reduced utility for coordinating on the Trump norm, denoted as a low value of U^T_{dem}. Symmetrically, a Republican would have a low U^B_{rep} for a 'Triumphant Biden' norm. Our model predicts that this framing creates a separation between μ_{dem} and μ_{rep}, transforming the population threshold distribution from the single-peaked unimodal case into the two-peaked bimodal distribution. The valley between these two peaks in the threshold density gives rise to the third, stable interior equilibrium – the very state of costly and persistent disagreement observed in the experiment's findings.

To see a contemporary example, consider the use of individual gender-neutral pronouns. This practice emerged during the late 2010s and gained rapid adoption among progressive circles, driven by the LGBT community as early adopters. However, the practice failed to trigger the thresholds of right-wing individuals – 65% of Republicans report they are uncomfortable with gender-neutral pronouns, versus just 24% of Democrats (PPRI, 2023). Today, the practice exists in a polarisation trap, with many still using gender-neutral terms and many still refusing to use them. The result is miscoordination, social friction and organisations unsure as to what they should enforce.

5.1 Welfare in the Polarisation Trap

The welfare of the average individual in the population is always lower in the polarisation trap, than when there exists a widely-adopted status quo norm ($p = 0$) or alternative norm ($p = 1$).

Given a proportion of the population choosing the alternative action, p, we can calculate the average utility for an agent in group g:

$$W_g(p) = \underbrace{\Phi(z_g(p)) \cdot pU_g^A}_{\text{Utility of } A \text{ choosers}} + \underbrace{(1 - \Phi(z_g(p))) \cdot ((1-p)U_g^S) + \overbrace{\sigma_g \phi(z_g(p))}^{E[\eta_i | x_i > p]}}_{\text{Utility of } S \text{ choosers}} \tag{13}$$

Where $z_g(p) = \frac{p - \mu_g}{s_g}$. $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF respectively. The average population utility is simply a weighted combination of the utility of minority and majority group members:

$$W(p) = \lambda W_m(p) + (1 - \lambda) W_M(p) \tag{14}$$

A utilitarian policymaker who cares equally about all members of society should try to maximise $W(p)$. The second derivative, $W''(p) > 0$, shows that this function is convex (appendix B). Figure 3 plots how population welfare changes as the alternative action increasingly gets adopted:

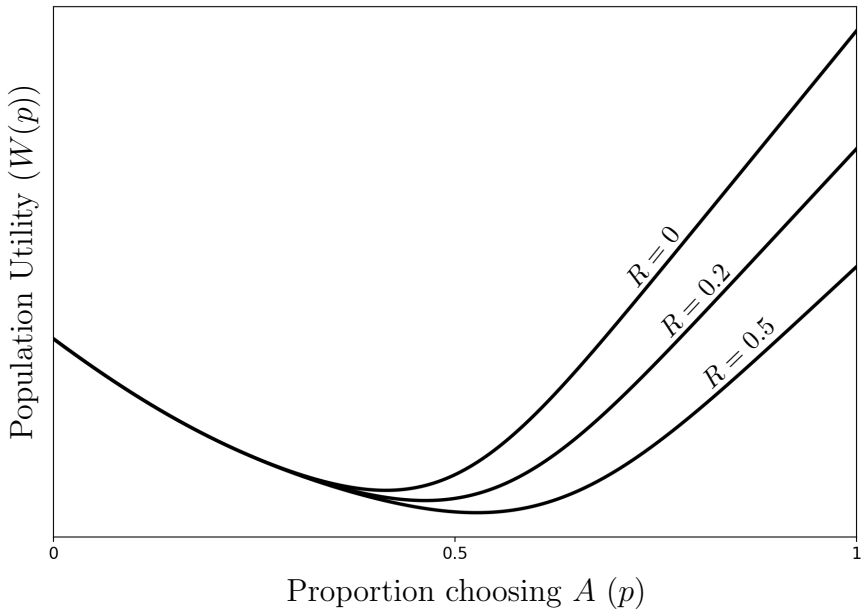


Figure 3: As more of the population chooses the alternative action, welfare initially declines as the status quo norm breaks down, before rising again as the alternative norm is established. Graph is shown for the parameter set: $\lambda = 0.4$, $\sigma_m = \sigma_M = 0.2$, $U_m^S = 0.2$.

Population utility is relatively high at full adoption of the status quo norm ($p = 0$) and the alternative norm ($p = 1$). At these endpoints, one of the groups is compromising but agents are coordinating with certainty. At intermediate values of p, population welfare declines as miscoordination becomes frequent.

To show this analytically, we make the following simplification: in the polarisation trap, all minority members are choosing A and all the majority members are choosing S. While this is not strictly true due to the long, infinite tails of η, it is generally true; recall that the trap forms in the gap between minority and majority thresholds where the

minority mass have had their thresholds triggered but the majority have not.

Given this, we say that $p = \lambda$ in the polarisation trap. To show that the trap reduces welfare, we prove that $W(\lambda) < W(0) < W(1)$. Firstly we look at the conditions for the minority group to prefer the status quo, that is, $W_m(\lambda) - W_m(0) < 0$:

$$\lambda U_m^A - U_m^S < 0 \tag{15}$$

In the trap, minority members receive their preferred outcome, but only when they are matched with another minority member (probability λ), otherwise they miscoordinate with a majority member and receive 0. In the status quo norm, they receive U_m^S with certainty. The inequality is not always true, but will be if the minority group is small enough, or the payoff increase is small enough to make the risk not worth it: $\lambda < U_m^S/U_m^A$.

The majority prefer the status quo norm when:

$$(1-\lambda)U_M^S - U_M^S < 0 \tag{16}$$

This is always satisfied, since the status quo norm returns their favoured payoff with certainty. The population welfare change from transitioning between a status quo norm and the polarisation trap ($W(\lambda) - W(0)$) is a population-weighted sum of conditions 15 and 16:

$$\underbrace{\lambda[\lambda U_m^A - U_m^S]}_{\text{Minority welfare change}} + \underbrace{(1-\lambda)[(1-\lambda)U_M^S - U_M^S]}_{\text{Majority welfare loss}} < 0 \tag{17}$$

The inequality highlights that the polarisation trap is harmful compared to a status quo norm whenever the utility lost by the majority dominates any potential utility gain by the minority. Applying the model's normalizations, $U_m^A = 1$ and $U_M^S = 1$, gives:

$$\lambda - \frac{1}{2} < U_m^S \tag{18}$$

The minority group has a population share $\lambda \in (0, 1/2)$ and U_m^S is strictly positive. Thus, the inequality holds for all valid parameter values.

Furthermore, we specified in restriction 3 that we are interested in cases where the alternative norm provides a welfare benefit over the status quo ($W(0) < W(1)$). As such, we can rank the population welfare in the three potential equilibria as: $W(\text{trap}) < W(0) < W(1)$.

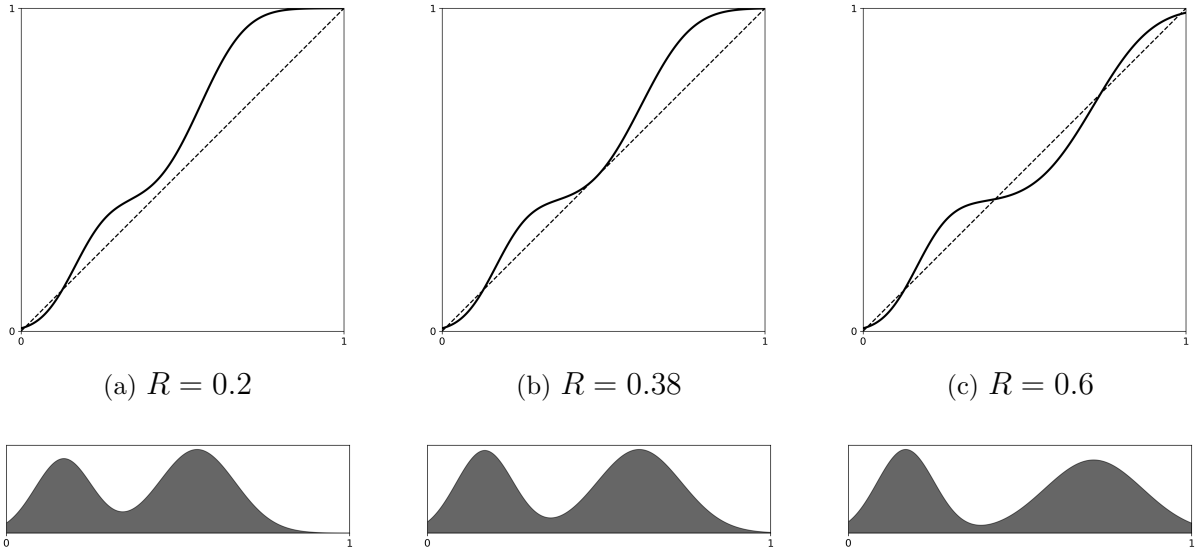
To test the robustness of this conclusion, I simulated the model under a wide variety of valid parameter combinations. Providing support to our simplifying assumption, 97.8% of minority members were choosing the alternative action in the polarisation trap, compared to only 0.7% of majority members. While it is true that the non-compliers in this case will have non-random η values, the results confirm that the error introduced by our assumption is negligible. Overall, in line with the prediction from our analytical result,

we found that the polarisation trap reduced welfare in 100% of simulations. All statistics can be found in appendix C.

5.2 Conditions for Polarisation Trap

Thus far, we have established that conflicting group interests can create a stable interior equilibrium, and that policymakers would be wise to avoid it. What conditions make the existence of such an equilibrium more likely? We identify two key parameters: Increased majority resistance (R) and decreased majority heterogeneity (σ_M).

The Polarisation Trap is created when the population CDF falls sufficiently to meet the 45 degree line:



(a) $R = 0.2$ (b) $R = 0.38$ (c) $R = 0.6$

Figure 4: As majority resistance increases from left to right, the second peak in the bimodal distribution is pulled away from the first. As the gap widens, the corresponding kink in the CDF decreases below the 45 degree line.

As majority resistance increases from (a) to (c), the two modes of the threshold distribution are pulled apart and a polarisation trap forms in the valley between them. At the exact point of creation (middle column), two conditions are met simultaneously: the equilibrium condition ($F(p^*) = p^*$) and the tangent condition ($f(p^*) = 1$).

This gives us a system of two equations with two unknowns, p^* and R (which enters through μ_M):

$$p^* = \lambda\Phi\left(\frac{p^* - \mu_m}{s_m}\right) + (1-\lambda)\Phi\left(\frac{p^* - \mu_M}{s_M}\right) \tag{19}$$

$$1 = \lambda\frac{1}{s_m}\phi\left(\frac{p^* - \mu_m}{s_m}\right) + (1-\lambda)\frac{1}{s_M}\phi\left(\frac{p^* - \mu_M}{s_M}\right) \tag{20}$$

The solution to this system returns the value of R at which the Polarisation Trap

is first created, which we denote $\hat{R}$. For all levels of majority resistance beyond this boundary ($R > \hat{R}$), the trap is guaranteed to exist, as an increase in R monotonically decreases the cumulative adoption probability $F(p)$, pulling the curve further below the 45-degree line (see Appendix D for proof).

Proposition 3. *The critical level of majority resistance required to create a polarisation trap, $\hat{R}$, is an increasing function of the majority group's preference heterogeneity, σ_M. That is, $\frac{d\hat{R}}{d\sigma_M} > 0$.*

Proof. A miscoordination trap emerges at a critical point $(p^*, \hat{R})$ where the aggregate threshold CDF, $F(p)$, is tangent to the 45-degree line. Applying the Implicit Function Theorem to the equilibrium condition at this critical point yields the relationship:

$$\frac{d\hat{R}}{d\sigma_M} = -\frac{\partial F/\partial \sigma_M}{\partial F/\partial \hat{R}}.$$

We determine the signs of the partial derivatives as follows:

$$\frac{\partial F}{\partial \hat{R}} = (1-\lambda)\phi(z_M)\frac{\partial z_M}{\partial \hat{R}} = (1-\lambda)\phi(z_M)\left(-\frac{p^*}{\sigma_M}\right) < 0 \tag{21}$$

$$\frac{\partial F}{\partial \sigma_M} = (1-\lambda)\phi(z_M)\frac{\partial z_M}{\partial \sigma_M} = (1-\lambda)\phi(z_M)\left(-\frac{p^*(2-\hat{R})-1}{\sigma_M^2}\right) > 0 \tag{22}$$

The derivative $\partial F/\partial \hat{R}$ is negative since p^*, σ_M, and $(1-\lambda)$ are positive. The derivative $\partial F/\partial \sigma_M$ is positive because the trap forms when $p^* < \mu_M$, which implies $p^*(2-\hat{R})-1 < 0$, making the term in the parentheses positive.

Combining these results confirms the relationship:

$$\frac{d\hat{R}}{d\sigma_M} > 0.$$

□

The result demonstrates that as the majority group becomes more diverse in their preferences, it takes a greater degree of inter-group conflict to create a polarisation trap. Intuitively, this captures a fundamental tension in social change: while we often think of diversity of opinion as an obstacle to consensus, our model reveals that heterogeneity within the majority can actually result in a more cohesive, unified society. When the majority holds diverse views (high σ_M), some members have relatively low thresholds for adopting the alternative norm, creating bridges between the minority and majority groups. These 'moderate' majority members can be convinced to switch even when average majority resistance is high, allowing social movements to maintain momentum and avoid getting stuck in a state of costly disagreement. Conversely, when the majority

is ideologically uniform (low σ_M), even modest resistance creates a sharp divide where the minority adopts the new norm but cannot convince a monolithic majority to follow. This result suggests that policymakers should be most cautious about tipping interventions not when societies are simply polarised, but when they are polarised and the opposing group is internally cohesive.

As is common with complex systems, small changes in R around $\hat{R}$ produce large discontinuous effects on other outcomes, such as population utility:

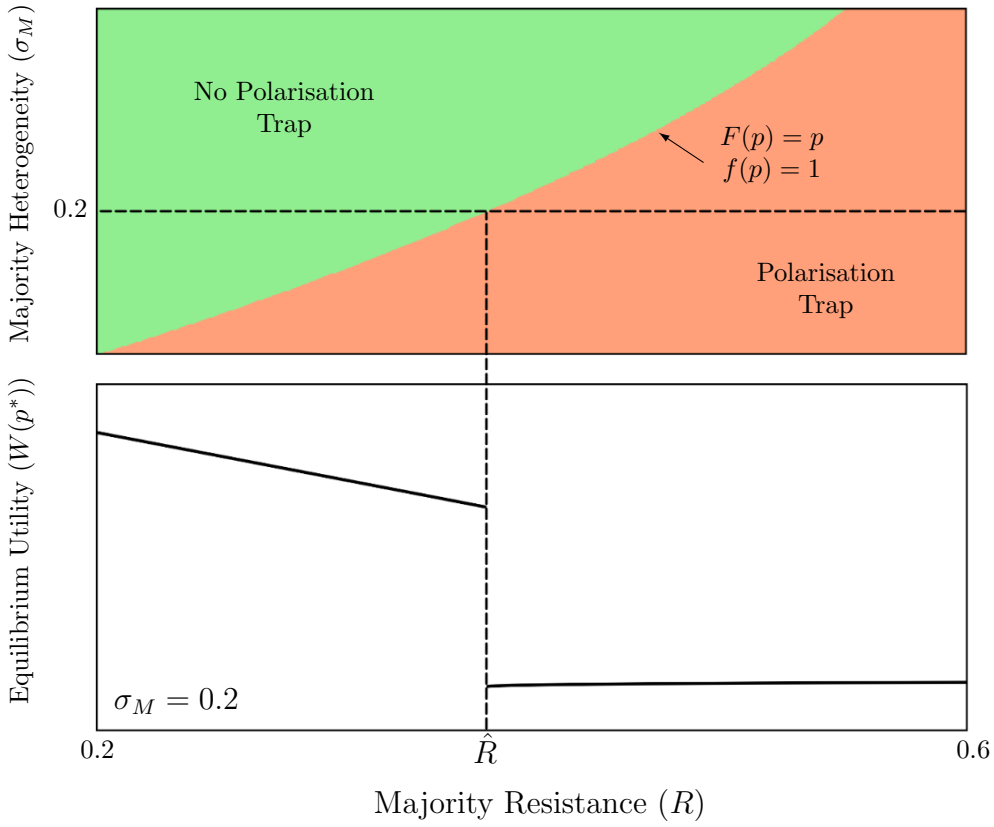


Figure 5: **Top:** Red area displays parameter combinations that create a Polarisation Trap. The trap emerges when majority resistance (R) is high and majority heterogeneity (σ_M) is low. **Bottom:** Cross sectional welfare analysis holding $\sigma_m = 0.2$ constant. As R increases past a critical threshold ($\hat{R}$), a Polarisation Trap is created, causing a discontinuous drop in population welfare for low initial adoption levels.

The top panel of Figure 5 shows that the polarisation trap is most likely to occur in the bottom-right corner. This is the scenario where the majority group is both highly resistant to change *and* very uniform in its opposition. The boundary between regimes demonstrates the relationship between R and σ_M from Proposition 3. They act as substitutes – the impact of a conflicted majority can be dampened by a heterogenous majority and vice versa.

The bottom panel demonstrates the discontinuous consequences of crossing the boundary. When majority resistance is low, the society successfully converges on the alternative norm. As the majority resistance (R) increases, the overall welfare of the society natu-

rally declines, driven by the majority dislike of the current, alternative norm. However, the moment it hits the critical point ($\hat{R}$) where the trap forms, there is a sudden, sharp drop in welfare. This happens because the society is no longer able to converge on a norm and settles in the newly formed polarisation trap instead.

6 Social Change in Polarised Environments

For a new social norm to replace an existing one in a polarised society, a movement must overcome two sequential and functionally distinct barriers. The first barrier is Ignition: generating sufficient initial momentum to disrupt the stability of the prevailing status quo. This is accomplished by a cohort of minority members who adopt the alternative behaviour unconditionally.

Once the status quo's hold is broken, the movement confronts the second barrier: Conversion. This is the challenge of expanding adoption from the initial core to the resistant majority, which is necessary to establish a new social consensus. The consequences of failing to surmount these barriers differ significantly. A failure of Ignition preserves the existing norm. However, a failure of Conversion after a successful Ignition leads the system to a stable interior equilibrium of costly miscoordination – the polarisation trap. This two-stage framework is therefore foundational for analysing the effectiveness of policy interventions designed to facilitate norm change.

6.1 Ignition

Ignition asks the question: are there sufficient radical, early alternative adopters to escape the pull of the status quo? If not, a movement will never begin and the problem of conversion is moot. The importance of ignition has been recognised by many studies that emphasise the building of a 'committed minority' as a method of tipping (Centola et al., 2018; Ye & Zino, 2025).

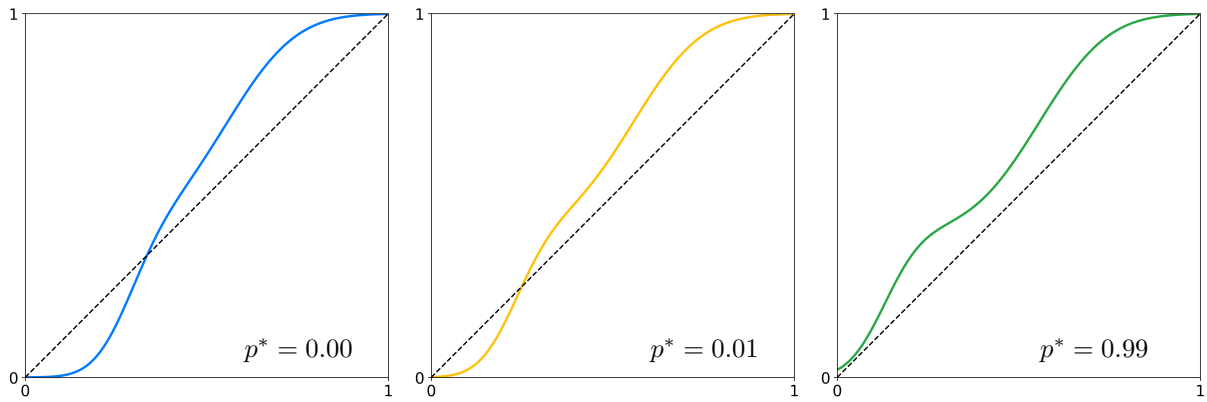
An agent is an **unconditional alternative adopter** if they choose A regardless of the expected proportion p. This occurs if their threshold $x_i \leq 0$. From the definition of x_i, this condition holds if $U^S_{g_i} + \eta_i \leq 0$. Since $U^S_{g_i} > 0$, this requires a sufficiently strong idiosyncratic distaste for the status quo (a large negative η_i). The proportion of unconditional alternative adopters is thus:

$$Pr(\eta \leq -U^S_g) = \Phi\left(\frac{-U^S_g}{\sigma_g}\right) \tag{23}$$

Where $\Phi(\cdot)$ is the standard normal CDF. Firstly, the equation shows that a diverse group ($\uparrow \sigma_g$) helps with ignition due to the creation of individuals with very low (and very high) thresholds. Secondly, more unconditional alternative adopters will be created

when dissatisfaction with the status quo increases ($\downarrow U_g^S$).

Consider a society in which the minority group are living under an entrenched status quo norm that does not serve their interests. As they become increasingly oppressed and conditions worsen, their guaranteed payoff from 'conforming to the system' decreases. Consequently, more and more radical minority members are being created. Once a critical mass has been created, a revolution will start:



(a) $U_m^S = 0.4$ (b) $U_m^S = 0.3$ (c) $U_m^S = 0.15$

Figure 6: **(a)** – **(b):** U_m^S decreases, creating unconditional alternative adopters but the status quo equilibrium is maintained. **(c):** Sufficient unconditional agents are created and a successful revolution is undertaken, destroying the status quo equilibrium.

Figure 6 shows the effect of worsening the status quo for the minority. In panel (c), the minority reaches their 'breaking point', in which enough radical A adopters have been created to initiate a successful transition to the alternative norm.

Consider a well-meaning policymaker that observes the society in panel (a). In an attempt to compensate the oppressed minority, they may design a policy ($\kappa > 0$) that makes the status quo norm more tolerable for them: $U_m^S \rightarrow U_m^S + \kappa$. However, it is important to consider that this will reduce the proportion of unconditional alternative adopters and thus reduce their ignition capabilities:

$$Pr(\eta \leq -U_g^S) \quad > \quad Pr(\eta \leq -(U_g^S + \kappa)) \tag{24}$$

The concept has also been used intentionally by rulers throughout history in order to prevent revolutions. In fact, the world's first welfare state was a concession implemented in 1880's Germany in order to placate the urban working class who were increasingly oppressed due to rapid industrialization (Gregory & Stuart, 2004). Here, Bismarck understood that providing a welfare state would increase $U^S_{\text{working class}}$, thus reducing their ignition capabilities and and decreasing the likelihood of a successful socialist movement.

6.2 Conversion

Conversion involves overcoming the resistance of the majority to achieve a full-scale shift to the alternative norm. After a movement has started (Ignition), it must generate sufficient momentum to persuade the majority group to adopt the new behaviour.

Mathematically, this challenge corresponds to lowering the system's second tipping point, which we denote $p^*_{u,2}$. This tipping point represents the critical mass of adopters required to trigger a cascade within the majority group. To analyse this, we perform comparative statics specifically on this equilibrium, which is identified as the unstable equilibrium ($F'(p^*_{u,2}) > 1$) located between the mean thresholds of the two groups ($\mu_M > p^*_{u,2} > \mu_m$). A parameter aids conversion if it reduces the value of $p^*_{u,2}$. For brevity, all derivatives are relegated to appendix E.

The analysis of the model's parameters reveals three primary pathways to successful conversion. Firstly, having a smaller majority to overcome ($\uparrow \lambda$), makes reaching the second tipping point an easier task. Secondly, conversion is aided by increasing their payoff from the alternative norm (U^A_m). If they do not see the alternative norm as particularly harmful to them, they will be easier to convert. Lastly, increasing their internal diversity (σ_M), creates crucial "bridge-builder" individuals that will be convinced to join the movement by the minority mass and in turn trigger the thresholds of their fellow majority members.

By the time a movement reaches the conversion stage, its success is primarily driven by the characteristics of the majority group it seeks to persuade. The direct effect of minority-specific parameters, such as their internal diversity (σ_m), becomes negligible at this point, as the minority's role in initiating the cascade is largely complete.

7 Interventions

Consider a society that, through history or otherwise, has an entrenched status quo norm ($p_0 = 0$). As policymakers, we would like to design an intervention to dislodge the status quo norm, ignite social change and transition to a new welfare-improving alternative norm.

To explore the effect of a general intervention, we simulate a targeted subsidy where a policymaker transfers a utility payment, s, to targeted individuals *if* they choose the alternative action. This intervention directly alters the expected utility calculation of a subsidized agent i and thus lowers their individual adoption threshold, x_i. The new threshold for a subsidized agent, $x_{i,post}$, is given by:

$$x_{i,post} = \frac{U^S_{g_i} + \eta_i - s}{U^A_{g_i} + U^S_{g_i}} \tag{25}$$

While the subsidy acts as a material incentive to choose the alternative action, the conclusions apply to any policy that lowers targeted individual thresholds. For example, s could be reinterpreted as a 'moral subsidy' that, through an informational campaign, causes the individual to receive a higher psychological payoff from 'doing the right thing'.

To test a range of interventions, we define a flexible targeting strategy using two independent parameters: $\phi_m \in [0, 1]$, the proportion of the minority group offered the subsidy, and $\phi_M \in [0, 1]$, the proportion of the majority group offered the subsidy. In this section, we test the effectiveness of three archetypal strategies: a minority-only intervention ($\phi_M = 0, \phi_m = 1$), a majority-only intervention ($\phi_M = 1, \phi_m = 0$), and a balanced, mixed-targeting approach ($\phi_M = 0.5, \phi_m = 0.5$).

The effectiveness of the policy is evaluated by the net welfare surplus it generates. This is calculated by comparing the population welfare before and after the intervention, minus the cost of the intervention. First, we determine the stable equilibrium in the absence of any policy, p^*_{pre}, and calculate the corresponding population welfare, $W(p^*_{pre})$. Next, we introduce the subsidy and find the welfare in the new stable equilibrium, $W(p^*_{post})$. The total cost of the subsidy, $C(s)$, is the payment s multiplied by the number of subsidized agents who adopt A in the new equilibrium. The net welfare change, ΔW, is therefore:

$$\Delta W = [W(p^*_{post}) - W(p^*_{pre})] - C(s) \tag{26}$$

A positive ΔW indicates a successful intervention that has generated utility spillovers, while a negative value signifies that the policy has backfired, leading to a net decrease in societal welfare.

To establish a clear efficiency frontier, we derive the theoretical upper bound for any subsidy, denoted $\bar{s}$. This represents the absolute maximum amount a policymaker should spend. Any intervention requiring a subsidy $s > \bar{s}$ is guaranteed to be a net loss for society, even if it successfully achieves full social change $p^* = 0 \to p^* = 1$. The bound is found where the intervention's cost, $C(s)$, equals the gross welfare gain from a complete norm shift, $W(1) - W(0)$. Solving for $\bar{s}$ we find:

$$\bar{s}(\phi_M, \phi_m) = \frac{[\lambda + (1-\lambda)(1-R)] - [\lambda U^S_m + 1 - \lambda]}{\phi_m \lambda + \phi_M (1-\lambda)} \tag{27}$$

For each policy type, we present the results of the policy on a base parameter set. However, to ensure our conclusions generalize beyond this, we simulate the intervention across a wide range of possible parameter sets and report statistical regularities.

7.1 Minority Subsidy, ($\phi_M = 0, \phi_m = 1$)

Policymakers seeking to shift away from a detrimental status quo norm might consider offering a subsidy to the minority group to encourage the adoption of an alternative

norm. The logic is that for some individuals, the incentive will be sufficient to overcome the risk of miscoordination, thereby prompting them to become early adopters of the new norm. A critical mass of these early adopters could then propel the system past a tipping point, leading to the spontaneous adoption of the alternative norm by the entire population, without further direct intervention. This approach is particularly appealing because minority members already have a predisposition towards the alternative norm, making it less expensive to change their behaviour.

Figure 7 below visualises the net gain from the intervention for different subsidy sizes and degrees of polarisation:

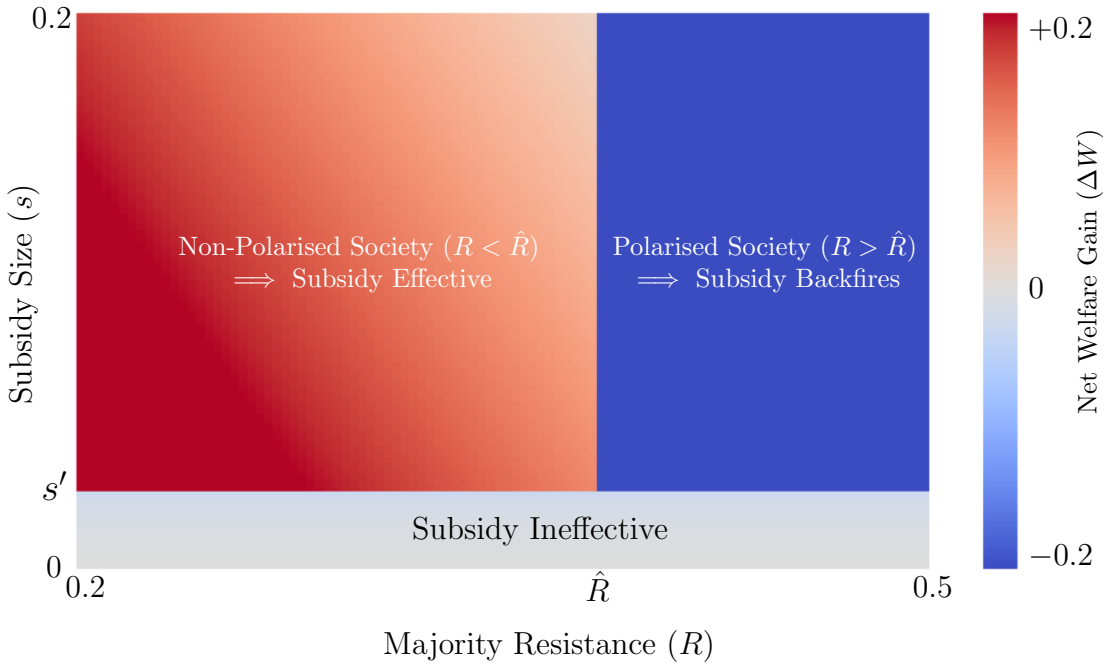


Figure 7: The welfare gain from a minority-targeted subsidy depends on the subsidy size and majority resistance. The subsidy is only impactful when it can overcome the tipping point ($s > s'$). Beyond this, the intervention yields a welfare surplus in non-polarised societies (red region) and a welfare loss in polarised societies (blue region). Parameter set: $\lambda = 0.4$, $U_m^S = 0.2$, $\sigma_m = 0.1$, $\sigma_M = 0.2$.

Results illustrate the critical importance of accounting for group-level preference differences when designing tipping interventions. The simulation reveals three distinct outcomes for a minority-targeted subsidy, with the result depending on the degree of societal polarisation (represented by the majority's resistance to the alternative, R). First, the grey region at the bottom shows that any subsidy below a certain size is simply ineffective; it fails to provide the necessary push to overcome initial inertia, regardless of the social structure. Second, in a non-polarised society where the majority group is not strongly resistant to the new norm (the red region), the subsidy works as intended. A sufficiently large incentive successfully nudges the system past its single tipping point, leading to a society-wide cascade and a net increase in social welfare. The endogenous spillovers created by the subsidy multiplies the direct effect of the intervention.

The most crucial finding, however, is demonstrated in the blue region. Here, where the two groups have sufficiently conflicted interests and the society is polarised, the very same intervention backfires. The subsidy is large enough to activate the minority and push the system over the first tipping point, but it lacks the momentum to overcome the resistance of the majority. Instead of reaching the welfare-improving alternative norm or falling back to a status quo norm, the society becomes stuck in the stable, welfare-destroying polarisation trap. This outcome demonstrates that a naive application of tipping interventions without an understanding of the underlying threshold distribution can be actively harmful, shattering a stable (if suboptimal) status quo and replacing it with persistent, costly miscoordination. Across a wide and systematically explored range of societal conditions, we find that the existence of a polarisation trap is the key predictor of whether a minority-targeted tipping intervention will backfire. When this trap is present, the policy destroys net welfare in 88% of cases by locking the society into a state of costly disagreement F.1. In the 12% of other parameter combinations, the system remained in the status quo.

7.2 Majority Subsidy, $(\phi_M = 1, \phi_m = 0)$

A policy targeted directly at the resistant majority is appealing. For example, to establish pro-immigration norms, we may design an intervention that encourages the majority demographic to be more open to cultural diversity in the country. This could be advantageous since, if the resistant group is first to change, the minority group will certainly join in and there will be no concern of a polarisation trap. Figure 8 shows the results:

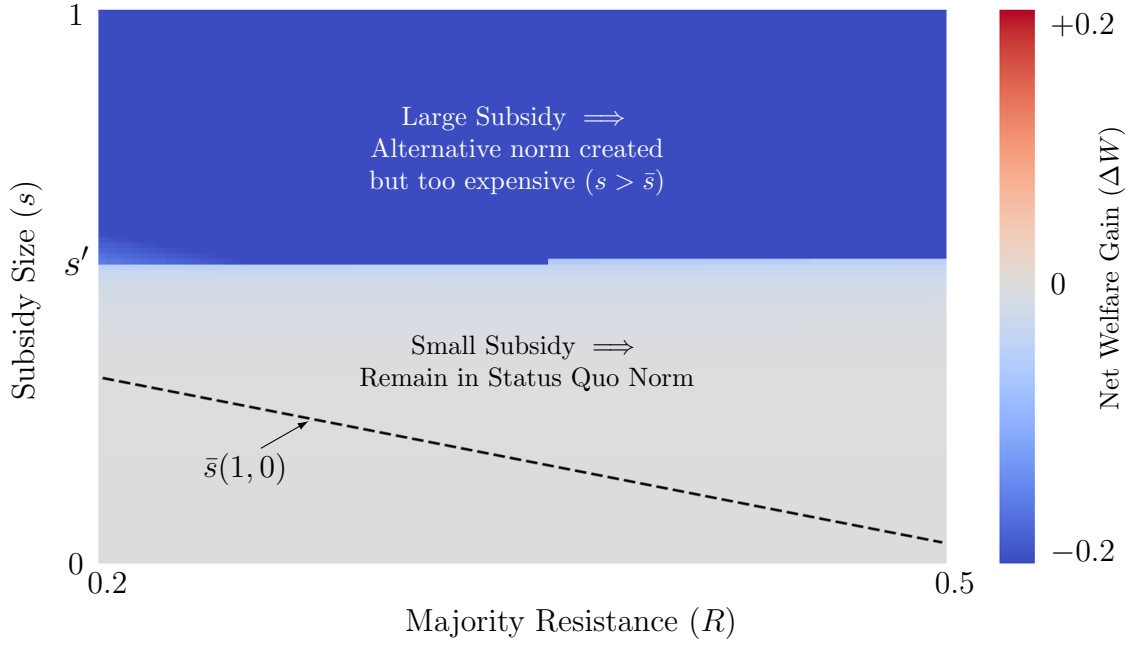


Figure 8: A majority targeted subsidy successfully creates the alternative norm but only for large subsidy sizes. Since the subsidy is large, the cost of the subsidy dominates the welfare gain from transitioning to the alternative norm. Parameter set: $\lambda = 0.4$, $U_m^S = 0.2$, $\sigma_m = 0.1$, $\sigma_M = 0.2$.

The majority-targeted subsidy succeeds at establishing the alternative norm but only for very large subsidy sizes. For this parameter set, the subsidy must be 12x larger in order to see an effect. This is because convincing the majority members to become early-adopters of a norm that is actively detrimental to them requires a huge financial incentive. By the time the subsidy is large enough to ignite change, the cost to the policymaker dominates any welfare gain from the norm transition: $s' > \bar{s} \ \ \forall \ R$. For this reason, we found that across the entire parameter space, a majority targeted subsidy produced a net welfare gain in only 2% of societies (Appendix F.2).

7.3 Mixed Subsidy ($\phi_M = 0.5, \phi_m = 0.5$)

Policymakers may consider an intervention that is targeted at a subsample of both groups. The logic is that the subsidized minority members become the initial push, and the subsidized majority members create the bridge between the two groups. Figure 9 shows the results on our base parameter set:

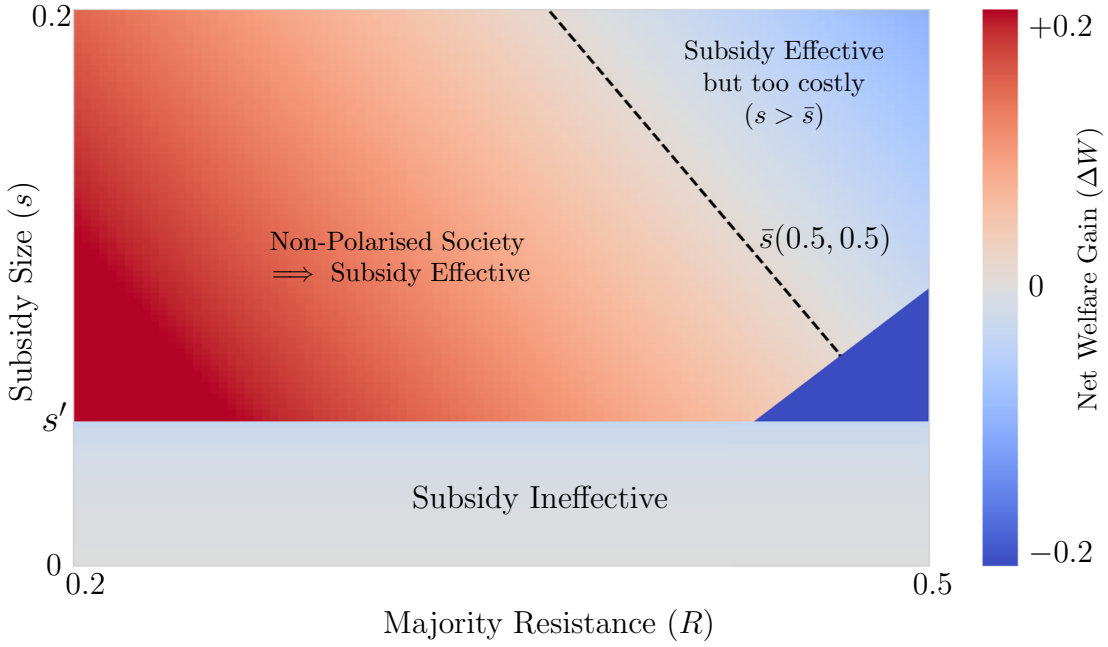


Figure 9: The subsidy is only impactful when it can overcome the tipping point ($s > s'$). Beyond this, the intervention yields a welfare surplus for many levels of polarisation (red region). The polarisation trap (deep-blue region) can be reduced by increasing subsidy size however the subsidy can become too costly to be worthwhile ($s > \bar{s}$) Parameter set: $\lambda = 0.4$, $U_m^S = 0.2$, $\sigma_m = 0.1$, $\sigma_M = 0.2$.

Compare this intervention to the minority-only subsidy in figure 7. Here, the dark-blue section resulting from a polarisation trap only appears for larger values of R. Furthermore, it is now a function of subsidy size, since this subsidy actively works to counteract the trap. Due to the newly-created moderate majority members, the area of the red section is larger, indicating that the subsidy is more effective for a wider range of polarisation levels. This allows the system to reach the welfare-improving alternative norm even in

polarised societies. That being said, there are caveats. Firstly, note that the subsidy size required for an effect (s') is larger than in figure 7. This is because our attention is being split between both groups, and less of the subsidy is going towards the initial minority push. Secondly, even though we successfully create the alternative norm for high levels of polarisation, if the subsidy to do so is too large ($s > \bar{s}$), the intervention is too costly.

Ultimately, we are presented with a trade off. For a given subsidy size, giving more to the minority increases the likelihood of seeing change (↑ Ignition) but giving more to the majority reduces the probability of creating a polarisation trap (↑ Conversion).

7.4 Optimal Policy

While the previous sections tested 'off-the-shelf' targeting strategies, we can conduct a two-dimensional policy optimization to find the optimal balance, given a fixed subsidy size, s. For each level of societal polarisation (R), we identify the single policy combination (ϕ_M, ϕ_m) that maximizes the net welfare gain for the entire population:

$$\begin{aligned} \max_{\phi_M, \phi_m} \quad & \left[W(p^*(s, \phi_M, \phi_m)) - W(p^*_{\text{pre}})\right] - C(s, \phi_M, \phi_m) \\ \text{s.t.} \quad & 0 \leq \phi_M \leq 1, \\ & 0 \leq \phi_m \leq 1 \end{aligned}$$

Figure 10 shows the optimal policy pair for different levels of polarisation:

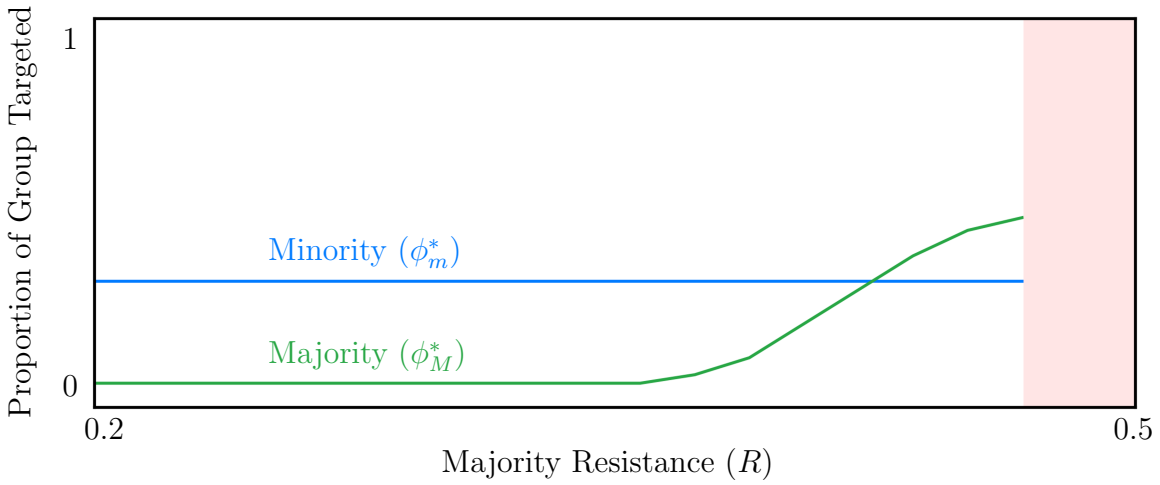


Figure 10: The graph displays the components of the policy that maximizes net welfare gain for each level of majority resistance (R). The optimal policy increasingly targets the majority as polarisation increases. The shaded red region indicates that polarisation is so severe that no active intervention produces a positive net welfare gain. Parameter set: $\lambda = 0.4$, $U^S_m = 0.2$, $\sigma_m = 0.1$, $\sigma_M = 0.2$, $s = 0.1$.

The structure of an effective policy is fundamentally dependent on the degree of societal polarisation. In societies with low resistance (R), the optimal strategy is intuitive: a subsidy targeted exclusively at the minority group is sufficient to ignite a welfare-

improving social cascade. Here, only ignition is of concern to the policymaker and this is addressed using the minority. However, as polarisation increases, conversion becomes a policy concern and a strategic shift is required. The optimal intervention now pivots towards a mixed-targeting approach, dedicating resources to a subset of the majority group alongside the minority. This is necessary to create moderate majority members and dispel the polarisation trap equilibrium. Beyond a certain critical point of polarisation, no active intervention is efficient; the cost required to bridge the social divide exceeds any potential welfare gain, making a policy of inaction the only rational choice.

8 Conclusion

The aim of this modelling exercise was to explore the dynamics of norm change in polarised environments. We found that, when groups disagree over their preferred social norm, the population will have a bimodal distribution of adoption thresholds. The gap between the two modes may give rise to a stable interior equilibrium where no norm is established and miscoordination is frequent. This interior 'polarisation trap' provides worse utility outcomes than full adoption of a social norm and is most likely to form when group preferences are highly conflicted or highly homogenous. When the trap is present in society, typical well-meaning social tipping interventions will backfire and destroy a functional norm. Instead, policymakers must dedicate some of their attention and resources toward targeting a subset of the resistant, majority group. The aim is not to create early adopters, but amenable majority members that 'bridge the gap' between minority and majority adoption. Alongside this, a subset of minority members should be targeted as the cheapest route to igniting social change.

While it is true that policymakers cannot observe the thresholds of their population, qualitative contextual work should be undertaken prior to an intervention in order to assess the likely degree of polarisation. When attempting to establish a new norm, how will different population groups react? If there exists a cohesive, resistant majority, care must be taken to facilitate change. Future work may wish to use the framework outlined in this paper to investigate another important question: given we initially begin in a polarisation trap, how do we escape and establish a universal norm?

Fundamentally, social norms evolve endogenously to solve coordination problems. This paper warns that attempts to culturally engineer a society without understanding the underlying group dynamics, risks replacing a stable norm with persistent and welfare-reducing conflict.

Bibliography

Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, *118*(16), e2014893118. https://doi.org/10.1073/pnas.2014893118

Ball, P. (2006). Critical Mass: How One Thing Leads to Another. Retrieved July 21, 2025, from https://philpapers.org/rec/BALCMH

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, *2006*(06), P06014. https://doi.org/10.1088/1742-5468/2006/06/P06014

Behboodian, J. (1970). On the Modes of a Mixture of Two Normal Distributions. *Technometrics*, *12*(1), 131–139. https://doi.org/10.2307/1267357

Berger, J. (2021). Social Tipping Interventions Can Promote the Diffusion or Decay of Sustainable Consumption Norms in the Field. Evidence from a Quasi-Experimental Intervention Study. *Sustainability*, *13*(6), 3529. https://doi.org/10.3390/su13063529

Binmore, K. (2022). THINKING ABOUT SOCIAL NORMS. *National Institute Economic Review*, *259*, 20–30. https://doi.org/10.1017/nie.2022.2

Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, *360*(6393), 1116–1119. https://doi.org/10.1126/science.aas8827

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S., & Weber, E. U. (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. https://doi.org/10.1177/15291006221105279

Efferson, C., Ehret, S., von Flüe, L., & Vogt, S. (2024). When norm change hurts. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *379*(1897), 20230039. https://doi.org/10.1098/rstb.2023.0039

Efferson, C., Vogt, S., & Fehr, E. (2020). The promise and the peril of using social influence to reverse harmful traditions. *Nature Human Behaviour*, *4*(1), 55–68. https://doi.org/10.1038/s41562-019-0768-2

Ehret, S., Constantino, S. M., Weber, E. U., Efferson, C., & Vogt, S. (2022). Group identities can undermine social tipping after intervention. *Nature Human Behaviour*, *6*(12), 1669–1679. https://doi.org/10.1038/s41562-022-01440-5

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, *83*(6), 1420–1443. Retrieved June 30, 2025, from https://www.jstor.org/stable/2778111

Gregory, P. R., & Stuart, R. C. (2004). *Comparing Economic Systems in the Twenty-first Century*. Houghton Mifflin.

Janas, M., Nikiforakis, N., & Siegenthaler, S. (2024). Eliciting Thresholds for Interdependent Behavior.

Jones, S. (2009). Dynamic Social Norms and the Unexpected Transformation of Women's Higher Education. *Social Science History*, *33*(3), 247–291. Retrieved August 19, 2025, from https://www.jstor.org/stable/40268002

Macy, M., & Tsvetkova, M. (2015). The Signal Importance of Noise. *Sociological Methods & Research*, *44*(2), 306–328. https://doi.org/10.1177/0049124113508093

North, D. C. (1991). Institutions. *The Journal of Economic Perspectives*, *5*(1), 97–112. Retrieved July 12, 2025, from https://www.jstor.org/stable/1942704

Nyborg, K., & Rege, M. (2003). On social norms: The evolution of considerate smoking behavior. *Journal of Economic Behavior & Organization*, *52*(3), 323–340. https://doi.org/10.1016/S0167-2681(03)00031-3

Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M., Doe, S. S. P., Lenferna, A., Morán, N., van Vuuren, D. P., & Schellnhuber, H. J. (2020). Social tipping dynamics for stabilizing Earth's climate by 2050. *Proceedings of the National Academy of Sciences*, *117*(5), 2354–2365. https://doi.org/10.1073/pnas.1900577117

Pizziol, V., & Tavoni, A. (2024). From niches to norms: The promise of social tipping interventions to scale climate action. *npj Climate Action*, *3*(1), 46. https://doi.org/10.1038/s44168-024-00131-3

PPRI. (2023, June). The Politics of Gender, Pronouns, and Public Education. Retrieved August 13, 2025, from https://prri.org/research/the-politics-of-gender-pronouns-and-public-education/

Rosenfeld, M. J. (2017). Moving a Mountain: The Extraordinary Trajectory of Same-Sex Marriage Approval in the United States. *Socius*, *3*, 2378023117727658. https://doi.org/10.1177/2378023117727658

Schelling, T. C. (1969). Models of Segregation. *The American Economic Review*, *59*(2), 488–493. Retrieved July 9, 2025, from https://www.jstor.org/stable/1823701

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

Ye, M., & Zino, L. (2025). Data-informed modeling of the formation, persistence, and evolution of social norms and conventions. https://doi.org/10.1007/978-3-030-93954-0_15-1

Yin, C.-c. (1998). Equilibria of collective action in different distributions of protest thresholds. *Public Choice*, *97*(4), 535–567. https://doi.org/10.1023/A:1004921725560

Young, H. P. (2015). The Evolution of Social Norms. *Annual Review of Economics*, *7*(Volume 7, 2015), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322

Appendix

A Proof of the Maximum Number of Equilibria

Proposition 4. *A system governed by the aggregate threshold CDF $F(p)$, defined as a mixture of two normal distributions, can have at most five equilibria. Of these, at most three can be stable and at most two can be unstable.*

Proof. Equilibria are fixed points $p^* \in [0,1]$ such that $p^* = F(p^*)$. These correspond to the roots of the function $G(p) = F(p) - p$. Let N be the number of distinct equilibria on $[0,1]$. By Rolle's Theorem, if $G(p)$ has N roots, its derivative, $G'(p)$, must have at least $N-1$ roots.

The derivative is $G'(p) = F'(p) - 1$, where $F'(p)$ is the aggregate threshold PDF, given by the mixture of two normal PDFs:

$$F'(p) = \frac{\lambda}{s_m}\phi\left(\frac{p-\mu_m}{s_m}\right) + \frac{1-\lambda}{s_M}\phi\left(\frac{p-\mu_M}{s_M}\right)$$

The roots of $G'(p)$ occur where $F'(p) = 1$. A mixture of two normal PDFs is known to have at most two local maxima (Behboodian, 1970). A function with at most two maxima can intersect a horizontal line (in this case, $y = 1$) at most four times. Thus, $G'(p)$ has at most four roots. This implies $N - 1 \leq 4$, so $N \leq 5$. The system can have at most five equilibria.

An equilibrium p^* is stable if $F'(p^*) < 1$ (i.e., $G'(p^*) < 0$) and unstable if $F'(p^*) > 1$ (i.e., $G'(p^*) > 0$). At the boundaries, we have $G(0) = F(0) > 0$ and $G(1) = F(1) - 1 < 0$. Since $G(p)$ is continuous and starts positive at $p = 0$, it must be decreasing at its first root p_1^*. Therefore, $G'(p_1^*) < 0$, and the first equilibrium is stable. For subsequent roots to exist, the function must change direction, meaning the sign of $G'(p^*)$ must alternate at each successive root. The sequence of stabilities is thus stable, unstable, stable, and so on. For the maximal case of five equilibria $(p_1^*, \ldots, p_5^*)$, the stabilities must be:

- p_1^*: Stable $\to$ p_2^*: Unstable $\to$ p_3^*: Stable $\to$ p_4^*: Unstable $\to$ p_5^*: Stable

This configuration allows for a maximum of three stable and two unstable equilibria. □

B Convexity of average utility function, $W(p)$

The second derivative of the individual group welfare functions is:

$$W_g''(p) = \underbrace{\left(\frac{U_g^A + U_g^S}{s_g}\right)}_{\text{Positive Constant}} \cdot \underbrace{\phi(z_g(p))}_{\text{Strictly Positive}} > 0 \tag{28}$$

Since the population average utility function is a weighted combination of the two groups, it is also strictly convex.

C Simulated Welfare in the Polarisation Trap

Table 1: Summary statistics from model simulation across a range of parameters

Statistic	**Value**
Average Equilibrium Position (p)	
Status Quo Norm (p_{eq1})	0.0119
Polarisation Trap (p_{eq2})	0.3883
Alternative Norm (p_{eq3})	0.9993
Average Equilibrium Welfare (W(p))	
Status Quo Norm ($W(p_{eq1})$)	0.6900
Alternative Norm ($W(p_{eq3})$)	0.7457
Polarisation Trap ($W(p_{eq2})$)	0.5295
Validation of Analytical Simplifications	
Avg. Absolute Difference $\lvert p_{eq2} - \lambda \rvert$	0.0109
Avg. Utility Missed by $E[\eta \mid S] = 0$ Assumption	0.0038
Group Composition at Trap Equilibrium (p_{eq2})	
% of Minority Group Choosing A	97.83%
% of Majority Group Choosing A	0.72%

D Proof of Polarisation Trap Existence $\forall R > \hat{R}$

Proposition 5. *Let $\hat{R}$ be the critical level of majority resistance at which a polarisation trap is created. For any $R > \hat{R}$, a stable interior equilibrium (a polarisation trap) is guaranteed to exist.*

Proof. At the critical point $R = \hat{R}$, the aggregate threshold CDF, $F(p, \hat{R})$, is tangent to the line $y = p$. A polarisation trap will persist for all $R > \hat{R}$ if the function $F(p, R)$ decreases as R increases. We prove this by showing that the partial derivative $\frac{\partial F}{\partial R}$ is strictly negative.

The parameter R only affects the majority group's component of the aggregate CDF, $F_M(p, R) = \Phi(z_M)$, where $\Phi(\cdot)$ is the standard normal CDF. Using the chain rule, we find the derivative:

$$\frac{\partial F}{\partial R} = (1-\lambda)\frac{\partial F_M}{\partial R} = (1-\lambda)\frac{d\Phi(z_M)}{dz_M}\frac{\partial z_M}{\partial R} = (1-\lambda)\phi(z_M)\frac{\partial z_M}{\partial R}$$

The standardized variable for the majority group is $z_M = \frac{p(2-R)-1}{\sigma_M}$. Its derivative with respect to R is:

$$\frac{\partial z_M}{\partial R} = -\frac{p}{\sigma_M}$$

Substituting this back into the expression for $\frac{\partial F}{\partial R}$ gives:

$$\frac{\partial F}{\partial R} = (1-\lambda) \cdot \phi(z_M) \cdot \left(-\frac{p}{\sigma_M}\right)$$

Since the weight $(1-\lambda)$, the standard normal PDF $\phi(z_M)$, the proportion of adopters p, and the heterogeneity σ_M are all positive for $p \in (0,1)$, the entire expression is strictly negative.

Because $\frac{\partial F}{\partial R} < 0$, the value of $F(p,R)$ decreases for any increase in R. Thus, once the curve $F(p)$ becomes tangent to $y = p$ at $\hat{R}$, it will dip further below for any $R > \hat{R}$, guaranteeing the trap's existence. □

E Comparative Statics for the Conversion Tipping Point

The following derivatives analyze the effect of key parameters on the second tipping point, $p^*_{u,2}$. The goal is to determine the sign of $\frac{dp^*_{u,2}}{d(\text{parameter})}$. The general formula is $\frac{\partial F/\partial(\text{parameter})}{1-F'(p^*_{u,2})}$. Since $p^*_{u,2}$ is an unstable equilibrium, the denominator is always negative. Therefore, the sign of the derivative is the opposite of the sign of the numerator.

- **Minority Size (λ):** The numerator is $\frac{\partial F}{\partial \lambda} = F_m(p) - F_M(p)$. At $p^*_{u,2}$, $F_m(p^*_{u,2})$ is high and $F_M(p^*_{u,2})$ is low, making the numerator positive. Thus, $\frac{dp^*_{u,2}}{d\lambda} < 0$. An increased minority size helps conversion.

- **Majority Alternative Payoff (U^A_m):** Increasing this payoff makes the majority less resistant and shifts their threshold distribution left, increasing $F_M(p)$ for all p. The numerator $\frac{\partial F}{\partial U^A_m}$ is therefore positive. Thus, $\frac{dp^*_{u,2}}{dU^A_m} < 0$. Reduced majority resistance helps conversion.

- **Majority Diversity (σ_M):** The numerator is $\frac{\partial F}{\partial \sigma_M} \propto -z_M\phi(z_M)$, where $z_M = (p-\mu_M)/s_M$. At the second tipping point, $p^*_{u,2} < \mu_M$, which makes z_M negative. The numerator is therefore positive. Thus, $\frac{dp^*_{u,2}}{d\sigma_M} < 0$. Increased majority diversity helps conversion by creating 'bridge-builder' adopters.

- **Minority Diversity (σ_m):** The numerator is $\frac{\partial F}{\partial \sigma_m} \propto -z_m\phi(z_m)$, where $z_m = (p-\mu_m)/s_m$. At $p^*_{u,2}$, we are far to the right of the minority mean, so z_m is large and positive. The numerator is therefore negative. Thus, $\frac{dp^*_{u,2}}{d\sigma_m} > 0$. However,

because $p^*_{u,2}$ is far in the tail of the minority's distribution, the magnitude of $\phi(z_m)$ is extremely small, making this effect negligible.

- **Minority Status Quo Payoff (U_m^S):** Increasing this payoff makes the minority more resistant, shifting their thresholds right and decreasing $F_m(p)$. The numerator $\frac{\partial F}{\partial U_m^S}$ is therefore negative. Thus, $\frac{dp^*_{u,2}}{dU_m^S} > 0$. Similar to σ_m, this effect is negligible as it occurs far in the tail of the minority's distribution.

F Intervention Simulation Robustness

F.1 Minority Targeted

Table 2: Minority Subsidy Outcome Distribution

Polarisation Trap	**Outcome Category (%)**		
Exists in Baseline?	Successful Tip	No Change	Backfire
No	13.1	53.6	33.3
Yes	0.0	12.5	87.5

This table shows the outcomes of the minority targeted intervention across 2000 simulated societies.

F.2 Majority Targeted

Table 3: Majority Subsidy Outcome Distribution

Outcome Category	**Percentage (%)**
Successful Tip	2.0
No Change	82.0
Backfire	16.0

This table shows the outcomes of the majority targeted intervention across 2000 simulated societies