A Theory of Strategic Uncertainty and Cultural Diversity *

Willemien Kets† Alvaro Sandroni‡

June 18, 2020

Abstract

We identify a new mechanism through which cultural diversity affects economic outcomes, based on a model of culture as shared cognition. Under this view, cultural diversity matters because it increases strategic uncertainty. The model can help better understand a variety of disparate evidence, including why homogeneous societies can be more conformist, why diverse societies may get stuck in a low-trust trap, why companies with a strong culture may fail to adopt superior work practices, and why autocratic rulers in diverse societies may overinvest in state capacity.

*A version of this paper has been circulated under the title "Challenging Conformity: A Case for Diversity." Part of the material incorporated here was previously in a paper entitled "A belief-based theory of homophily" by the same authors (Kets and Sandroni, 2019). We are grateful to the Associate Editor and three anonymous referees for excellent suggestions. We thank George Akerlof, Robby Akerlof, Larbi Alaoui, Roland Bénabou, Jean-Paul Carvalho, Vincent Crawford, Yan Chen, Vessela Daskalova, Wouter Dessein, Dominik Duell, Georgy Egorov, Armin Falk, Karla Hoff, Matthew Jackson, Johannes Johnen, Wouter Kager, Rachel Kranton, Eliana La Ferrara, Bart Lipman, Hamish Low, George Mailath, Niko Matouschek, Meg Meyer, Rosemarie Nagel, Santiago Oliveros, Scott Page, Antonio Penta, Nicola Persico, Debraj Ray, Nick Robalino, Yuval Salant, Larry Samuelson, Paul Seabright, Rajiv Sethi, Eran Shmaya, Andy Skrzypacz, Jakub Steiner, Colin Stewart, Jeroen Swinkels, Tymon Tatur, Roberto Weber, Yiqing Xing, and numerous seminar audiences and conference participants for helpful comments and stimulating discussions. Jasmin Droege provided excellent research assistance.

†Department of Economics, University of Oxford. E-mail: willemien.kets@economics.ox.ac.uk.

‡Kellogg School of Management, Northwestern University. E-mail: sandroni@kellogg.northwestern.edu

1 Introduction

In this era of increasing economic integration and cultural mixing, the question of how cultural diversity affects economic outcomes is increasingly important. However, our current understanding of how cultural differences affect economic outcomes is still incomplete. While economic theory can account for the effects of cultural diversity when groups differ in factors that are directly payoff-relevant, such as preferences or skills,[1] there is mounting empirical evidence that people from different cultural backgrounds also differ in factors that are not directly payoff relevant, in particular *cognitive factors*: What they pay attention to, how they respond to contextual cues, and what is salient to them.[2] Despite this growing empirical interest in how culturally-determined differences in cognition affect economic outcomes, a unified theoretical framework is still lacking. However, without such a framework, it is difficult to predict how a change in diversity will affect economic outcomes and what the associated welfare implications are. The challenge for economic theory is therefore to develop a methodology that can model how culture shapes cognition and to demonstrate that it can deliver new economic insights.

The research program proposed in this paper is a systematic study of the economic effects of cultural diversity, and ultimately, a study of how culture can have important economic consequences by shaping people's cognition. To take a first step in this program, we formalize the idea of culture as shared cognition. Under this view, a common culture reduces strategic uncertainty while cultural diversity increases it. To model this, we build on research in psychology to develop a formal model of how people reason about others. The reasoning process selects an equilibrium that depends on both standard economic incentives and sociocultural factors such as diversity. We use this framework to derive new testable hypotheses and novel welfare implications and show that the model can help better understand a variety of disparate evidence. Our results demonstrate the importance of understanding the precise mechanisms that drive the effects of culture: Whether the effects of diversity are driven by differences in cognition or in payoff-relevant factors matters for predictions.

To study the effects of culture on cognition, we keep the baseline model deliberately simple. In our model, players belong to one of two (cultural) groups. A society in which all players belong to the same cultural group is culturally homogeneous; otherwise, it is culturally diverse. We focus on games with strategic complementarities. Because these games generally have multiple equilibria, the payoff structure of the game does not completely pin down behavior. Thus, players face considerable *strategic uncertainty*, that is, uncertainty about other players' actions. Traditional game theory assumes away strategic uncertainty by positing that players select one of the Nash equilibria. But such an approach fails to explain why some societies coordinate on better equilibria than others and which policies, if any, can lead societies to coordinate on better outcomes. We therefore depart from traditional game theory by explicitly

[1] See Alesina and La Ferrara (2005) for an excellent survey. Also see Section 6.

[2] For an excellent overview in the context of development, economic history, institutional economics, organizational economics, and behavioral economics, see World Bank (2015), Greif and Mokyr (2017), North (2005) and Kaplan and Henderson (2005), and Demeritt and Hoff (2018), respectively.

modeling how players reason about others. Our starting point is the observation by Schelling (1960) that in settings with strategic uncertainty, "[a player's] objective is to make contact with the other player through some imaginative process of introspection" (p. 96). To reach such a "meeting of the minds," players can use *theory of mind*. Theory of mind is a central concept in psychology. It refers to the cognitive ability to take another person's perspective.[3] Perspective-taking involves introspection: players put themselves into another person's shoes using their own subjective experience as a guide. That is, they observe their own mental state and project it onto others. This is a rapid and instinctive process referred to as first-person simulation (Goldman, 2006).[4] It is followed by a slower, more deliberative process whereby individuals reason about others' mental states using "folk psychology," i.e., a naive understanding of others' introspective process. This may lead them to adjust their initial belief (Gopnik and Wellman, 1994).

We model this by assuming that each players has a pre-reflective inclination (an "impulse") to take a certain action. Impulses do not directly affect payoffs and are privately observed. A player's first instinct is to follow his impulse. Upon introspection, he realizes that other players also have an impulse. Given this, it may not be optimal for him to follow his impulse. This may lead him to adjust his decision. Upon further reflection, he realizes that other players may likewise adjust their response, which may lead him to revise his choice. Players continue to reason in this way until no player wishes to revise his choice. The limit of this process defines the *introspective equilibrium*.

The key assumption is that people's impulses are influenced by their cultural background. This builds on the work in sociology (DiMaggio, 1997) and anthropology (D'Andrade, 1995) that shows that extended exposure to social and cultural patterns shapes people's cognitive frames, that is, what they pay attention to, how they structure their experience – in short, what is salient to them. This means that it is easier for people to put themselves into the shoes of people from the same cultural group: if an action is salient to a player (i.e., if he has an impulse to choose it), it is likely to be *culturally salient* in the sense that most players from his group will have the same impulse. On the other hand, people from different cultural groups are less likely to have been exposed to the same social and cultural patterns, and are therefore less likely to agree on what is salient. This makes it harder for players to take the perspective of people from other groups.[5] In that sense, *players face more strategic uncertainty in diverse societies than in homogeneous societies*.

So, culture enters into our model in a minimal way: it influences only the impulses that anchor the introspective process, and the sole difference between culturally homogeneous and

[3]Thus, theory of mind needs to be distinguished from emotional perspective-taking (empathy); the former is central to strategic reasoning, the latter may affect social preferences (Singer and Fehr, 2005).

[4]This idea also has a long history in philosophy. Locke (1690/1975) suggests that people have a faculty of "Perception of the Operation of our own Mind," and called introspection the "sixth sense." Mill (1872/1974) writes that understanding others' mental states first requires understanding "my own case."

[5]For evidence from psychology that people find it easier to take the perspective of members of their own group, see, e.g., Nelson and Baumgarte (2004), Williams et al. (2007), and Heinke and Louis (2009).

diverse societies is the degree of strategic uncertainty that players face. Nevertheless, diversity can have a profound influence on economic outcomes. Homogeneous and diverse societies may select different equilibria even when they are identical in all payoff-relevant aspects. Moreover, the equilibrium selection depends on the economic environment in a systematic way.

We first demonstrate this for settings where the incentive to choose an action varies smoothly with the share of players who choose it. When the payoff structure of the game provides little guidance, there can be *miscoordination*, i.e., players may fail to coordinate on one of the Nash equilibria of the game. In this case, culturally homogeneous societies outperform culturally diverse ones: Because it is easier for players to anticipate the instincts of members of their own group, the risk of miscoordination is minimized when the society is homogeneous. So, *when actions are nearly symmetric in terms of payoffs, cultural diversity is costly because it increases the risk of miscoordination.*

The situation is different when one of the actions stands out in terms of payoffs. In this case, diversity is beneficial. To see why, suppose that the Nash equilibria can be Pareto-ranked. Because there is little strategic uncertainty in homogeneous societies, players' behavior is strongly guided by which action they expect to be culturally salient. While this can facilitate coordination, it may also lead to *inefficient lock-in*.[6] That is, homogeneous societies may get locked into playing a culturally salient equilibrium even if all players would prefer (collectively) to switch to a different equilibrium. By contrast, because players' expectations are more likely to diverge in diverse societies, cultural salience plays a smaller role in these societies. As a result, choices are more strongly guided by payoff considerations, and this helps avoid inefficient lock-in. So, *when there is some asymmetry in terms of payoffs, cultural diversity is beneficial because it reduces inefficient lock-in.*

We next consider the case where the incentive to choose an action is a discontinuous function of the share of players choosing each action. We focus on models of *regime change*, where players want to overthrow a regime but benefit only if enough players attack the regime. These models are commonly used to model a wide variety of phenomena, including revolutions, bank runs, debt crises, and currency attacks (Morris and Shin, 2003). There is again tradeoff between cultural salience and payoff considerations, though the insights are subtly different. Strong regimes are more vulnerable to attacks in homogeneous societies while weak regimes are more fragile when the society is diverse. Intuitively, diverse societies are more fragmented than homogeneous societies. When ousting the regimes requires concerted action by a large number of players (i.e., the regime is strong), fragmentation makes it harder to coordinate a successful attack. But if the regime can be overthrown even if a small number of players attack (i.e, the regime is weak), then fragmentation makes the regime more vulnerable. This is because in fragmented societies, it is more likely that attacking is culturally salient for at least *some* groups even if it fails to be culturally salient for *all* of them. So, *diversity increases the likelihood of a successful attack when the regime is weak, but not when it is strong.*

In sum, cultural diversity can be an economic cost or benefit depending on the economic

[6]Inefficient lock-in is sometimes also referred to as *coordination failure* in the literature.

environment. On the one hand, a common culture enables effective coordination; on the other hand, shared beliefs may constrain players in their choices if these beliefs anchor players' expectations which then become self-fulfilling. This central insight is consistent with the view in sociology that culture both enables and constrains (Swidler, 1986) and the view in the organizational behavior literature that cultural diversity is a "double-edged sword" (Milliken and Martins, 1996). Our formal model goes beyond this by delivering testable hypotheses on the conditions under which each of these forces dominates. Our model predicts that whether the net impact of strategic uncertainty is positive or negative depends on the relative strength of cultural salience and payoff considerations.

We show that the basic tension between miscoordination and inefficient lock-in can help understand a range of economic phenomena and puzzles. For example, our model helps understand why homogeneous societies tend to be more conformist while diverse societies may be caught in a low-trust trap. It also sheds light on why organizations with a strong culture may not incentivize workers to choose more efficient work practices, why diverse societies may suffer more under autocratic regimes, and why weak regimes engage in nation-building while strong regimes create groups with distinct identities. In each application we consider, the aim is not to give a definite account of any one issue in particular, because a thorough treatment would warrant a paper of its own. Rather, the goal is to illustrate how *a single mechanism can help understand a variety of disparate empirical evidence.*

While our stylized model leaves out many elements that can be expected to be important in practice, our predictions are broadly in line with empirical evidence. For example, there is ample evidence that diverse societies have less trust and cooperation (Alesina and La Ferrara, 2002, 2005), that homogeneous societies exhibit excessive conformism (Bursztyn et al., 2018), that weak regimes are more likely to invest in nation-building (Alesina et al., 2018) while stronger regimes use divide-and-rule tactics to create distinct groups (Acemoglu et al., 2004), and that companies with a strong culture often fail to adjust to changing economic conditions (Kotter and Heskett, 1992). While some of these findings could be captured at least in part by other mechanisms (e.g., preference heterogeneity, social preferences), no existing model can accommodate them all. Our model instead provides a unified account, which, once properly extended, can also accommodate other applications. Another important point is that many of these alternative mechanisms require some form of equilibrium selection that is often left unmodeled. These approaches thus leave open the question of why some societies select better outcomes than others, and which policies can help a society move to a better equilibrium. By contrast, we explicitly model how economic and sociocultural factors shape equilibrium selection. So, while we do not expect our simple model to provide a definite account of the empirical phenomena that we consider, we believe it can offer a promising approach to better understand the effects of culture and diversity in a variety of applications.

The outline of this paper is as follows. After introducing the model in Section 2, Section 3 presents our main theoretical results and Section 4 considers applications. Section 5 discusses the key features of introspective equilibrium, and Section 6 discusses the related literature.

2 Model

2.1 Strategic complementarities

This section defines the class of games that we consider. There is a continuum $N = [0, 1]$ of players. Each player belongs to one of two *groups*, labeled A and B. Group membership is observable. The shares of players belonging to group A and B are α and β, respectively (i.e., $\alpha, \beta \geq 0, \alpha + \beta = 1$). Without loss of generality, we take $\beta \leq \frac{1}{2}$, so that A is the *majority group* and B is the *minority group*. Each player $j \in N$ chooses an action $s_j \in \{0, 1\}$. For ease of reference, we will often denote action 1 by H ("High") and action 0 by L ("Low"). A player's payoff depends on his own action and on the proportion of players choosing each action. Payoffs may also depend on an individual payoff parameter $\rho_j \in \mathbb{R}$. Specifically, if a player $j \in N$ chooses action s_j while the other players play according to the action profile s_{-j}, the player receives a payoff $u(s_j, m; \rho_j)$, where $m = m(s_{-j}) \in [0, 1]$ is the proportion of players $j' \neq j$ choosing the high action under s_{-j}.

We consider *monotone games with strategic complementarities* (Vives, 2005; Van Zandt and Vives, 2007). That is, the incentive to choose the high action is increasing in the proportion of players that choose the high action: for $\rho_j \in \mathbb{R}$ and $m, m' \in [0, 1]$ such that $m' \geq m$,

$$u(H, m'; \rho_j) - u(L, m'; \rho_j) \geq u(H, m; \rho_j) - u(L, m; \rho_j). \tag{2.1}$$

In addition, the incentive to choose the high action is monotone in ρ_j: for $m \in [0, 1]$, and ρ_j, ρ'_j such that $\rho'_j > \rho_j$,

$$u(H, m; \rho_j) - u(L, m; \rho_j) > u(H, m; \rho'_j) - u(L, m; \rho'_j). \tag{2.2}$$

Thus, the incentive to choose the high action is decreasing in the parameter ρ_j. We interpret ρ_j as the *risk* for player j associated with choosing the high action: a player has a greater incentive $u(H, m; \rho_j) - u(L, m; \rho_j)$ to choose the high action if his risk parameter ρ_j is small. We consider two cases for the risk parameters: (1) *identical preferences* (i.e., $\rho_j = \rho$ for all $j \in N$); and (2) *preference heterogeneity*: ρ_j is drawn from a continuous distribution $F(\rho_j)$ (i.i.d. across players). The parameters ρ_j are common knowledge.

Finally, we impose a separability assumption: the incentive $u(H, m; \rho_j) - u(L, m; \rho_j)$ to choose the high action is proportional to $g(m) - \rho_j$ where $g(m)$ increases with m. This ensures that an increase in the proportion m of players who choose the high action affects all players' incentives equally (independent of their parameter ρ_j). This is clearly without loss of generality if players have identical preferences.

Importantly, group membership does not affect payoffs in our model. For example, the payoff function $u(s_j, m; \rho_j)$ does not depend on the group that j belongs to and the distribution of risk parameters is the same across groups. Moreover, the payoff to a player depends only on the actions of other players, not on which group they belong to. And since there is no payoff uncertainty, it is also not the case that players have superior information about the preferences of members of their own group.

2.2 Introspection

This section defines the introspective process by which players resolve strategic uncertainty. The literature in psychology on theory of mind shows that people resolve strategic uncertainty by taking others' perspective (Apperly, 2012). A key component of theory of mind is an introspective process whereby people reflect on their own inclinations to form expectations about others. We model this as follows. Before choosing an action, each player j receives an *impulse* $I_j \in \{H, L\}$. Impulses are drawn from a distribution $\mu((I_j)_j)$ that may be shaped by players' culture, as we discuss in Section 2.3 below. Impulses are privately observed and do not directly affect payoffs. A player's instinctive reaction is to follow his initial impulse. That is, if the player's impulse is $I_j = s_j$, then his pre-reflective inclination is to choose action s_j. This defines the player's level-0 strategy σ_j^0 (i.e., $\sigma_j^0(I_j) = s_j$ whenever $I_j = s_j$). Upon introspection, the player realizes that other players likewise have an impulse. So, he forms a posterior belief $\mu(I_{-j} \mid I_j)$ about the other players' impulses. This allows the player to formulate a best response to the other players' level-0 strategy.[7] This defines his level-1 strategy σ^1. The reasoning does not stop here: For any level $k > 0$, the player's level-k strategy σ_j^k is a best-response to the level-$(k-1)$ strategy σ_{-j}^{k-1} of the other players. Players continue to reason in this way until they no longer wish to revise their choice. Accordingly, the behavior of player j is described by the limit $\sigma_j := \lim_{k\to\infty} \sigma_j^k$ (if it exists), so that $\sigma_j(I_j) \in \{H, L\}$ is j's equilibrium action if his impulse is I_j. The profile $\sigma = (\sigma_j)_j$ of limiting strategies is an *introspective equilibrium*.

We can relate introspective equilibrium to one of the classic equilibrium concepts in game theory:

Proposition 2.1. [**Common Knowledge of Rationality**] *Any introspective equilibrium is a correlated equilibrium.*

Together with the epistemic characterization of correlated equilibrium by Aumann (1987), Proposition 2.1 implies that introspective equilibrium is consistent with common knowledge of rationality. That is, even though players are assumed to be "folk game theorists" who do not engage in a full-fledged equilibrium analysis, they act as if they are rational, believe that the other players are rational and that they believe that others are rational, and so on. The intuition behind Proposition 2.1 is simple: At each level $k > 0$, players choose a best response to the level-$(k-1)$ strategy of the other players. In the limit $k \to \infty$, players thus choose a best response to strategies that are a best response to strategies that are a best response to.... These are precisely the actions that can be played in a correlated equilibrium. Proposition 2.1 holds very generally: As the proof shows, it holds for any game that satisfies some mild technical assumptions (including games without strategic complementarities).

We can also compare introspective equilibrium to level-k and cognitive hierarchy models (Nagel, 1995; Stahl and Wilson, 1995; Costa-Gomes et al., 2001; Camerer et al., 2004).[8] As

[7] If there are multiple best responses, an action is chosen using a fixed tie-breaking rule. The choice of tie-breaking rule does not affect our results.

[8] For comparisons to other concepts, see Section 5.

in these models, introspective players engage in an iterative reasoning process. However, there are two important differences. The first, relatively minor, difference is that our model abstracts from bounded rationality (i.e., we take $k \to \infty$). This is not for descriptive realism; rather, it is to emphasize that our results are not driven by bounded rationality.[9] Second, and more importantly, while the level-k literature has no room for culture, our model makes it possible to capture the effects of culture, as we discuss next.

2.3 Culture and strategic uncertainty

This section develops a simple model of how culture influences the introspective process. With probability p, action H is *culturally salient* for group G (denoted $\theta_G = H$) in the sense that a proportion $q \in (\frac{1}{2}, 1)$ of players in G have an impulse to choose H. With the remaining probability $1-p$, action L is culturally salient for G ($\theta_G = L$). For much of the paper, we take $p = \frac{1}{2}$ (without explicit mention), so as to abstract from systematic differences between actions. The parameter q is a measure of *culture strength*: If q is close to 1, almost all players in a group agree on which action is culturally salient; when q is close to $\frac{1}{2}$, players are almost equally divided on which action they expect to be salient for their group.

To model that players are more likely to agree on what is salient if they belong to the same group, we take cultural salience to be imperfectly correlated across groups. That is, the joint distribution of θ_A and θ_B (when $p = \frac{1}{2}$) is given by:

	$\theta_B = H$	$\theta_B = L$
$\theta_A = H$	$\frac{1}{4}(1+\eta)$	$\frac{1}{4}(1-\eta)$
$\theta_A = L$	$\frac{1}{4}(1-\eta)$	$\frac{1}{4}(1+\eta)$

where $\eta \in (0,1)$. Thus, the parameter $d := 1-\eta$ measures the *cultural distance* between groups: When d is close to 0, an action that is culturally salient for one group is highly likely to be culturally salient for the other group; when d is close to 1, whether an action is culturally salient for a group is almost completely uninformative of whether it is salient for the other group. Importantly, the impulse distribution depends on the population composition (i.e., α, β). This is because an action that is culturally salient for one group need not be salient for the other (i.e., $d > 0$). Thus, the correlation in impulses is maximized when all players belong to the same group ($\beta = 0$) and decreases with diversity (as measured by the minority share β).

We next consider players' posterior beliefs. By observing their own impulses, players can make inferences about other players' impulses. Because a player with an impulse to choose action s expects more than half of players (of either group) to have an impulse to choose s, we say that a player with an impulse to choose s *expects action s to be culturally salient*. Importantly, because what is culturally salient for one group need not be salient for the other group ($d > 0$), players' impulses are more informative about their own group. To see this, note

[9] However, most of our results only require that players can reason up to $k = 2$, and all of our results go through qualitatively if all players reason up to some finite level k.

that a player with a given impulse expects a proportion $Q_{in} := q^2 + (1-q)^2$ of players in his group to have the same impulse while he expects a proportion $Q_{out} := d \cdot \frac{1}{2} + (1-d) \cdot Q_{in}$ of players from the other group to have the same impulse. Then, a player's impulse is more informative of the impulses of the members of his own group in the sense that $Q_{in} > Q_{out} > \frac{1}{2}$, and the difference is particularly pronounced when the cultural difference between groups is large (i.e., $Q_{in} - Q_{out}$ increases with d).[10] In this sense, *players face more strategic uncertainty when they interact with players from the other group*; see Online Appendix I for a formal statement.

3 Theoretical results

3.1 An illustrative example

We illustrate the key results with a simple example. Players' goal is to choose an action that is close to the economic fundamentals but they also want to match others' actions. Recalling that $H = 1$ and $L = 0$, the payoff to a player j who chooses action s_j is given by

$$-[(s_j - \bar{s})^2 + (s_j - \tau)^2], \tag{3.1}$$

where $\tau \in \mathbb{R}$ is a payoff state that represents the economic fundamentals and $\bar{s} = \bar{s}(s_j, s_{-j})$ is the average action, which is equal to the proportion m of players who choose the high action. This is a monotone game with strategic complementarities with risk parameter $\rho = 1 - \tau$. It is easy to check that H is a best response for a player whenever he expects at least a proportion $m = 1 - \tau$ of players to choose it. So, for any $\tau \in [0, 1]$, both all players choosing H and all players choosing L are Nash equilibria.

We study how behavior varies with economic fundamentals (i.e., τ) and diversity (i.e., β). First consider the case where no action is payoff salient in the sense that neither action has greater intrinsic appeal in terms of payoffs (i.e., τ is close to $\frac{1}{2}$). In this case, there is a unique introspective equilibrium, and in this introspective equilibrium, all players choose the action they expect to be culturally salient. The intuition is that, because no action stands out in terms of payoffs, players rely on cultural salience to guide their behavior. Players with an impulse to choose action s expect s to be culturally salient; consequently, they expect most players to have the same impulse. At level 1, therefore, their unique best response is to choose action s; and it follows from a simple inductive argument that the same is true at higher levels. So, cultural salience breaks the symmetry between the actions. However, there is *miscoordination*: Because players may disagree on which action is culturally salient ($q < 1$), some players choose the high action while others choose the low action. Because players are more likely to agree on which action is culturally salient if they belong to the same group, there is more miscoordination

[10] The group structure itself may also affect impulses. For example, if a man and a woman arrive at a door at the same time, then, at least in some societies, there is an expectation that the man holds the door and let the woman go first (cf. Akerlof and Kranton, 2000, 2010). We abstract away from this effect.

in diverse societies (i.e., $\beta > 0$). Hence, *when the payoff structure of the game provides little guidance, cultural diversity is costly as it increases the risk of miscoordination.*

We next consider the case where H is more attractive than L in terms of payoffs; say, $\tau = \frac{4}{5}$. In this case, there can be a tension between cultural salience and payoff salience. For example, even though the high action is payoff salient, a player may expect the low action to be culturally salient.[11] The key insight is that how this tension is resolved depends on the relative strength of cultural salience and payoff considerations. Moreover, the relative strength of each depends on diversity. First consider a homogeneous society ($\beta = 0$) with a strong culture (q close to 1). In this case, there is a unique introspective equilibrium, and in this introspective equilibrium, players choose the action they expect to be culturally salient. Intuitively, while one of the action stands out in terms of payoffs, cultural salience trumps payoff salience because the lack of strategic uncertainty implies that cultural salience has a large impact on players' expectations. In particular, a player with an impulse to choose L thinks it is highly likely that an overwhelming majority of players has an impulse to choose L (i.e., Q_{in} close to 1). So, by a similar argument as before, this player chooses L in introspective equilibrium. By contrast, in diverse societies (say, $\beta = \frac{1}{2}$), players coordinate on the payoff salient action in the unique introspective equilibrium. Intuitively, there is considerable strategic uncertainty in the sense that players' impulses contain little information about which action is culturally salient. In the extreme case that the cultural distance is large (d close to 1), a player's impulse is almost completely uninformative about the impulses of the other group (Q_{out} small). In that case, cultural salience is relatively weak and choices are guided by payoff considerations: A player with an impulse to choose L expects a significant share of players to have an impulse to choose H. Together with the payoff advantage of H, this implies that the player's unique best response at level 1 is to choose H. The same holds, a fortiori, for players with an impulse to choose H. So, in introspective equilibrium, all players choose the payoff salient action, regardless of which action they expect to be culturally salient. Diverse societies are thus more likely to avoid *inefficient lock-in*: for a range of payoff parameters, players in diverse societies choose the payoff salient action while players from homogeneous societies choose the action they expect to be culturally salient. Hence, *if there is a conflict between cultural salience and payoff considerations, cultural diversity is an economic benefit because it reduces the scope for inefficient lock-in.*

These observations have implications for welfare. We measure social welfare by the expected total payoff in introspective equilibrium, that is,

$$\begin{aligned}\widehat{W}(\tau;\beta) &:= -\tfrac{1}{2}\mathbb{E}_\beta\left[m\left((1-m)^2+(1-\tau)^2\right)+(1-m)\left(m^2+\tau^2\right)\right]\\ &= -\tfrac{1}{2}\mathbb{E}_\beta\left[m\,(1-m)+m\,(1-\tau)^2+(1-m)\,\tau^2\right],\end{aligned}$$

where m is the proportion of players who choose the high action in introspective equilibrium and the expectation is again taken over the impulse distribution (which is a function of β). So,

[11] While we do not model this explicitly, this could be because the society has a long history of playing L in this particular context (perhaps because the payoffs were different in the past) or in other, similar, situations.

social welfare is maximized if there is no miscoordination (i.e., $m = 0$ or $m = 1$) and if there is no inefficient lock-in (i.e., $m = 1$ if $\tau > \frac{1}{2}$; and $m = 0$ if $\tau < \frac{1}{2}$). The welfare implications of diversity are again driven by the tradeoff between miscoordination and inefficient lock-in: *if no action is payoff salient, cultural diversity increases the risk of miscoordination and cultural homogeneity is socially optimal. On the other hand, if one of the actions stands out in terms of payoffs, cultural diversity reduces the scope for inefficient lock-in and cultural diversity is socially optimal.* Hence, the same feature – strategic uncertainty – drives both the costs and benefits of cultural diversity, and whether cultural diversity is economically costly or beneficial depends on the economic environment.

The next section shows that the insights from this simple example apply generally, though the insights are somewhat richer in more general settings. For example, coordinating on the payoff salient action may not be socially optimal if there is a tension between the incentive to choose the high action and the welfare implications of doing so.[12] As another example, if one group is larger than the other ($\beta \in (0, \frac{1}{2})$), groups may face different incentives and this can lead to more complex patterns of behavior. And for games that have an additional parameters, such as games with a discontinuity in payoffs at a threshold T, the effects of diversity may depend on this additional parameter. Nevertheless, the key insights generalize. Moreover, the predictions are testable: While the model's point predictions depend on difficult-to-measure parameters like the cultural distance d and culture strength q, *the same qualitative comparative statics and welfare implications obtain for any combination of cultural factors.*

3.2 Equilibrium

This section studies the effects of diversity on equilibrium behavior. We measure *diversity* by the minority share β: Societies with a small minority (β close to 0) are nearly homogeneous while societies with a large minority (β close to $\frac{1}{2}$) are culturally diverse.[13] While many of our results hold more generally, we focus on *linear games* (i.e., $g(m) = m$) and *threshold games* (i.e., there is a threshold T such that $g(m) = 1$ if $m \geq T$ and $g(m) = 0$ otherwise) because these cover the main applications. If a game is linear and has heterogeneous preferences, then the distribution $F(\rho_j)$ of risk parameters is assumed to be unimodal and symmetric (e.g., normal or uniform). The game in Section 3.1 is an example of a linear game; see Section 4.4 for an application of a threshold game.

Proposition 3.1. [Existence & Uniqueness] *Every linear game or threshold game has an introspective equilibrium. It is essentially unique.*

[12] This is because payoff salience is defined in terms of players' incentives (i.e., $u(H, m; \rho_j) - u(L, m; \rho_j)$) while social welfare refers to players' total payoffs.

[13] Measuring diversity with the minority share simplifies the exposition while yielding the same comparative statics as more standard diversity indices. In particular, it gives the same comparative statics as the fractionalization index $\delta := 1 - \alpha^2 - \beta^2$ given that $\frac{d\delta}{d\beta} \geq 0$, with strict inequality for $\beta < \frac{1}{2}$.

Proposition 3.1 demonstrates that introspective equilibrium is well-suited to analyze the games that we consider. It demonstrates that introspective equilibrium delivers sharp predictions in the socioeconomic environments that we consider. By "essentially unique," we mean that introspective equilibrium uniquely pins down behavior except in knife-edge cases: If players have identical preferences, then the set of risk parameters for which the introspective equilibrium is unique has measure 1; and if there is preference heterogeneity, then introspective equilibrium uniquely determines behavior for a measure-1 set of risk parameters ρ_j. Uniqueness obtains even though games with strategic complementarities typically have many Nash (and correlated) equilibria (see Online Appendix II for a comparison). This will be critical for obtaining testable hypotheses and unambiguous welfare implications.

We next consider equilibrium behavior. The aim is not to derive detailed point predictions. Rather, we focus on obtaining simple comparative statics that deliver testable hypotheses for applications.[14,15] The following result considers linear games. To state the result, say that *players' decisions are driven by payoff considerations* if, in introspective equilibrium, all players choose the same action regardless of which action they expect to be culturally salient.

Proposition 3.2. [Sociocultural Factors: Linear Games] *In any linear game with identical preferences ($\rho_j = \rho$ for all $j \in N$),*

(a) If players' decisions are driven by payoff considerations in a homogeneous society ($\beta = 0$), then players' decisions are driven by payoff considerations in a diverse society ($\beta = \frac{1}{2}$).

(b) There is $\beta^* \in (0, \frac{1}{2})$ such that

(b1) For $\beta < \beta' < \beta^*$, if players' decisions are driven by payoff considerations when diversity is β, then players' decisions are driven by payoff considerations when diversity is β'.

(b2) For $\beta > \beta' > \beta^*$, if players' decisions are driven by payoff considerations when diversity is β, then players' decisions are driven by payoff considerations when diversity is β'.

(c) For $q, q' \in (0, \frac{1}{2})$ with $q' > q$, if players' decisions are driven by payoff considerations when the culture is strong (culture strength q'), then players' decisions are driven by payoff considerations when the culture is weak (culture strength q).

(d) For $d, d' \in (0, 1)$ with $d' > d$, if players' decisions are driven by payoff considerations when the cultural distance between groups is small (distance d), then players' decisions are driven by payoff considerations when the cultural distance is large (distance d').

[14] However, the proofs of Propositions 3.2–3.3 provide a full characterization of introspective equilibrium, which can be used to derive point predictions for settings where the relevant variables can be measured.

[15] Given our focus on the impact of diversity, we focus on the impact of sociocultural factors here. See Online Appendix III for more traditional comparative statics results on payoffs.

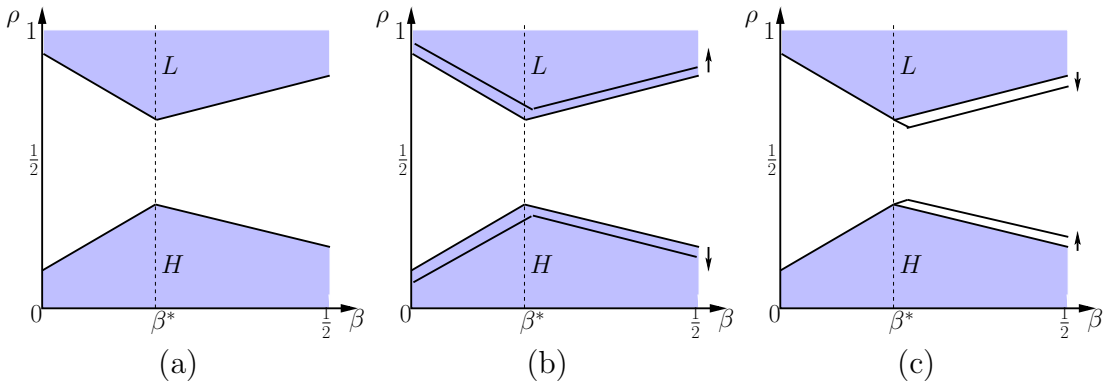

Figure 1: Introspective equilibrium or linear games: (a) Parameter combinations (β, ρ) for which players' decisions are driven by payoff considerations (shaded areas); (b) effect of an increase in culture strength q; (c) effect of an increase in cultural distance d.

In each case, the converse need not hold. These results extend to games with limited preference heterogeneity. For example, part (a) becomes: Fix a sequence $(F^n(\rho_j))_n$ of normal distributions such that for each n, $F^n(\rho_j)$ has mean ρ and variance $\tilde{\sigma}_n^2 > 0$, with $\tilde{\sigma}_n \to 0$. Then, for any $\varepsilon > 0$, for n sufficiently large, if the share of players whose action does not depend on which action they expect to be culturally salient in introspective equilibrium is at least $1 - \varepsilon$ under F^n in a homogeneous society ($\beta = 0$), then it is at least $1 - \varepsilon$ in a diverse society ($\beta = \frac{1}{2}$), but the converse need not hold. The other parts can be extended in a similar way.

Proposition 3.2(a) formalizes the insights from the example in Section 3.1 and shows that they generalize to any linear game. The intuition is the same as before: In homogeneous societies ($\beta = 0$), players' expectations about which action is culturally salient are largely aligned, so payoff considerations play a limited role; by contrast, in diverse societies ($\beta = \frac{1}{2}$), there is more strategic uncertainty and therefore more room for payoff considerations to affect behavior.

Proposition 3.2(b) provides further insights into the effects of diversity. The result is illustrated in Figure 1(a). The shaded areas in in Figure 1 represent the parameter combinations for which players' decisions are driven by payoff considerations, with players choosing H if the risk ρ of choosing the high action is low and L if ρ is high. While decisions are more strongly guided by payoff considerations in diverse societies ($\beta = \frac{1}{2}$) than in homogeneous societies ($\beta = 0$), the range of payoff parameters for which players' behavior is driven by payoff parameters is maximized at intermediate levels of diversity ($\beta = \beta^*$). Intuitively, when the society is nearly homogeneous, players from the minority group face significant strategic uncertainty and their actions tend to be guided by payoff considerations. However, if the minority is small (β close to 0), the actions of minority players have little impact on the incentives for the majority. So, majority players may choose the action they expect to be culturally salient even if one of the actions stand out in terms of payoffs. On the other hand, if the minority is large (i.e., β close to $\frac{1}{2}$), minority players face little strategic uncertainty. So, cultural salience is relatively strong and decisions may not be driven by payoff considerations unless one of the actions is

very attractive in terms of payoffs (i.e., ρ close to 0 or 1). Hence, *payoff considerations play the greatest role in shaping behavior when the minority has a critical mass* ($\beta = \beta^*$).

Parts (c) and(d) consider the effects of cultural factors. Proposition 3.2(c), which is illustrated in Figure 1(b), states that there is more scope for payoff considerations to drive behavior when the culture is weak. The intuition is similar to before: When the culture is weak, players face more strategic uncertainty, and this leaves more room for payoff considerations. In this sense, cultural diversity and a weak culture are substitutes. Proposition 3.2(d), which is illustrated in Figure 1(c), shows that increasing the cultural distance between groups has the opposite effect to strengthening the groups' culture:[16] When there is a larger cultural distance between groups, there is more strategic uncertainty and thus more scope for payoff considerations to shape behavior. Again, the intuition is that increasing strategic uncertainty reduces the power of cultural salience and gives more room to payoff considerations to drive behavior. Hence, *payoff considerations play a greater role in shaping behavior when the groups' culture is weak and the cultural distance between groups is large.* As Proposition 3.2 notes, these results extend to games with limited preference heterogeneity. This follows because, as we show, introspective equilibrium satisfies a continuity property: the introspective equilibrium of a game with limited preference heterogeneity is "close" to an introspective equilibrium of games with identical preferences (Lemma A.6).[17]

We next consider threshold games, i.e., there is a threshold T such that $g(m) = 1$ if $m \geq T$ and $g(m) = 0$ otherwise. The canonical threshold game has payoffs given by

$$\begin{aligned} u_j(H, s_{-j}) = & \begin{cases} B_j & \text{if } m \geq T; \\ -C_j & \text{otherwise;} \end{cases} \\ u_j(L, s_{-j}) = & \ 0; \end{aligned} \tag{3.2}$$

where $m = m(s_{-j})$ is the proportion of players who choose H under s_{-j}. A common interpretation of threshold games is that players can choose whether to attack a regime (play H) or not (play L); the regime falls if and only if the share of players attacking it exceeds the threshold T, which is a measure of the regime's strength. So, it is a best response for a player to attack if he assigns probability at least $\rho_j := \frac{C_j}{B_j + C_j}$ to the attack being successful (i.e., $m \geq T$). The following result shows that the effects of diversity are mediated by the regime's strength T:

Proposition 3.3. [Sociocultural Factors: Threshold Games] *In any threshold game with identical preferences:*

[16] As illustrated in Figure 1(c), the bounds on ρ vary with cultural distance d only if diversity β is sufficiently high while the bound on ρ varies with culture strength q for any β (Figure 1(b)). This is because when β is small, the bound on ρ is driven by the actions of majority players, and this effect is independent of d; see the proof of Proposition 3.2 for details.

[17] The comparative statics in Proposition 3.2 do not extend to settings where preferences are highly heterogeneous. In this case, the introspective equilibrium is independent of culture. This is similar to how the introduction of large payoff uncertainty or preference heterogeneity as in global games yields a unique (culture-independent) outcome (Vives, 2005, p. 444 and p. 472). However, the assumption that the risk parameters are normally distributed is much stronger than necessary; see the proof of Proposition 3.2 for details.

(a) The effects of diversity β depend on the threshold T:

(a1) For $T > \frac{1}{2}$, if players attack when the society is diverse ($\beta = \frac{1}{2}$), then they attack if the society is homogeneous ($\beta = 0$); moreover, attacks are more likely to be successful if the society is homogeneous.

(a2) For $T < \frac{1}{2}$, if players attack when the society is homogeneous ($\beta = 0$), then they attack if the society is diverse ($\beta = \frac{1}{2}$); moreover, attacks are more likely to be successful when the society is diverse.

(b) For $q, q' \in (0, \frac{1}{2})$ with $q' > q$, if players' decisions are driven by payoff considerations when the culture is strong (culture strength q'), then players' decisions are driven by payoff considerations when the culture is weak (culture strength q).

(c) The effects of cultural distance d depend on the threshold T. Fix $d, d' \in (0, 1)$ with $d' > d$. Then:

(c1) For $T > \frac{1}{2}$, if players attack when the cultural distance is large (i.e., d'), then they attack if cultural distance is small ($d = 0$).

(c2) For $T < \frac{1}{2}$, if players attack when the cultural distance is small (i.e., d), then they attack if the cultural distance is large (i.e., d').

In each case, the converse need not hold.

Proposition 3.3(a) shows that for threshold games, the effects of diversity depend on the threshold T: If the threshold is small ($T < \frac{1}{2}$), then players in a diverse society ($\beta = \frac{1}{2}$) choose H for a larger range of parameters and the share of players choosing H is more likely to exceed the threshold than in a homogeneous society ($\beta = 0$); but if the threshold is high ($T > \frac{1}{2}$), then the effects are reversed. Part (a1) states that *if the regime is strong ($T > \frac{1}{2}$), diversity reduces the likelihood of a successful attack*. To see the intuition, note that diverse societies are more fragmented than homogeneous societies: When attacking is culturally salient for one group, it need not be salient for the other ($d > 0$). The key point is that fragmentation reduces the likelihood of a successful attack when the regime is strong ($T > \frac{1}{2}$). This is because bringing down a strong regime requires concerted action by a large number of of players. This means that there is a risk of miscoordination, and this risk is greater if the society is diverse. To see this, suppose that the society is diverse ($\beta = \frac{1}{2}$) and that attacking carries some risk ($\rho > 1 - q$). Then, at level 1, attacking is a best response for a player if he expects attacking to be culturally salient for both groups (as $T > \frac{1}{2}$). But even if the player expects attacking to be culturally salient for his *own* group, he may be uncertain as to whether attacking is salient for the *other* group. This reduces his incentive to attack. By contrast, because players' expectations are largely aligned when all players belong to the same group ($\beta = 0$), players face little strategic uncertainty in homogeneous societies. Thus, a player who expect attacking to be culturally salient has a strong incentive to attack. Moreover, because the probability

that attacking is culturally salient for a single group is of course higher than the probability that attacking is culturally salient for two groups, the likelihood that the attack is successful is higher in diverse societies (keeping fixed the strategies, i.e., conditional on players in either society attacking whenever they expect it to be culturally salient for their group). Part (a2) shows that diversity has the opposite effect when the regime is weak: *If the regime is weak ($T < \frac{1}{2}$), diversity increases the likelihood of a successful attack.* Intuitively, because an attack is successful whenever at least one group revolts (as ($T < \frac{1}{2}$), it is sufficient if attacking is culturally salient for one group. Because the likelihood that attacking is culturally salient for at least one group is obviously higher than the likelihood that it is salient for any given group, an attack is more likely to succeed when the society is diverse. This further incentivizes players to attack. So, weak regimes are more vulnerable in diverse societies while strong regimes are more vulnerable if the society is homogeneous. Proposition 3.3(b)–(c) consider the effects of culture. Proposition 3.3(b) shows that when the culture is strong, players are more inclined to choose the action they expect to be culturally salient. The intuition is similar as for linear games: When the culture is strong, cultural salience trumps payoff considerations. Proposition 3.3(c) reinforces the insight from Proposition 3.2 that an increase in cultural distance has a similar effect as diversity. The intuition is the same as before: An increase in cultural distance leads to more strategic uncertainty, and this makes it more attractive to attack when the regime is weak ($T < \frac{1}{2}$) but less attractive when the regime is strong ($T > \frac{1}{2}$). So, again, *the same factor – strategic uncertainty – drives the costs and benefits of cultural diversity, and its net impact depends on the economic environment.*

3.3 Welfare

This section considers the welfare implications of diversity. We start with studying the optimal level of diversity, that is, the level of diversity that maximizes social welfare in introspective equilibrium. For ease of exposition, we focus on linear games with identical preferences (i.e., $g(m) = m$ and $\rho_j = \rho$ for all $j \in N$).[18] We fix a class of payoff functions $(u(\cdot;\rho))_\rho$ throughout and denote by $W(m;\rho)$ the total expected payoff if a proportion m of players chooses H and the risk parameter is ρ. Denote by $\widehat{W}(\rho;\beta)$ the expected social welfare in introspective equilibrium when the payoff function is $u(\cdot;\rho)$ and the level of diversity is β. Then, the *optimal level of diversity* $\bar{\beta}$ is the level of diversity that maximizes expected social welfare in introspective equilibrium (i.e., $\widehat{W}(\rho;\bar{\beta}) \geq \widehat{W}(\rho;\beta)$ for all β) and we say that *cultural diversity is socially optimal* if the optimal level of diversity is strictly positive (i.e., $\widehat{W}(\rho;\bar{\beta}) \geq \widehat{W}(\rho;0)$ for some $\bar{\beta} > 0$). The following result identifies the conditions under which cultural diversity is socially optimal.

[18] Similar results obtain for threshold games, though they are more difficult to state due to the additional parameter T. The results also extend to games with limited preference heterogeneity. However, when there is significant preference heterogeneity, the results may change, for two reasons. First, equilibrium behavior may depend less strongly on sociocultural factors (footnote 17). Second, the effects of diversity may become ambiguous if players have conflicting preferences.

Proposition 3.4. [Costs & Benefits of Diversity] *For linear games with identical preferences such that $W(m;\rho)$ is convex and quadratic in m and $W(1;\rho)-W(0;\rho)$ decreases with ρ, there exist $\underline{\rho},\bar{\rho}$, with $1-Q_{in}<\underline{\rho}<\frac{1}{2}<\bar{\rho}<Q_{in}$ such that cultural diversity is socially optimal if and only if $\rho<\underline{\rho}$ or $\rho>\bar{\rho}$.*

Proposition 3.4 shows that the optimal level of diversity is non-monotonic in economic incentives: culturally diverse societies have a higher aggregate payoff when one of the actions is payoff salient (i.e., ρ bounded away from $\frac{1}{2}$) while culturally homogeneous societies have a higher level of social welfare when the payoff structure provides little guidance (i.e., ρ close to $\frac{1}{2}$). While the conditions in Proposition 3.4 on the welfare function might seem restrictive, they are in fact satisfied by our applications (see Online Appendix IV).[19]

The intuition behind Proposition 3.4 is similar to that for the example in Section 3.1: The optimal level of diversity depends on the tradeoff between miscoordination and inefficient lock-in. When no action stands out in terms of payoffs, cultural diversity is costly because it increases miscoordination; but when one of the actions is more attractive in terms of payoffs, diversity improves welfare. This is driven by two mechanisms. First, in games where coordinating on the action that stands out in terms of payoffs is also socially optimal, as in the game in Section 3.1, diversity helps avoid inefficient lock-in. In other games, it may not be possible to avoid inefficient lock-in altogether (e.g,. Section 4.2 below). Nevertheless, cultural diversity can still improve welfare: Because minority players face more strategic uncertainty than majority players, their decisions are more likely to be driven by payoff considerations. This helps a diverse society coordinate on a single alternative, and this improves welfare precisely when the alternative is not too inferior (i.e., ρ sufficiently close to $\frac{1}{2}$). Hence, the tradeoff between miscoordination and inefficient lock-in is a key driver of the welfare implications of diversity also in this case.

Proposition 3.4 implies that diverse societies may have higher welfare than homogeneous ones if the level of diversity can be chosen optimally. However, it may not always be feasible to choose the optimal level of diversity; instead, a planner may be constrained to making small (local) changes. The following result shows that a small increase in diversity can be costly even in cases where diversity has economic benefits.

Proposition 3.5. [The Costs of Small Minorities] *For linear game with identical preferences such that $W(m;\rho)$ is convex and quadratic in m, $W(1;\rho)-W(0;\rho)$ decreases with ρ, and $\rho\in(1-Q_{in},Q_{in})$,[20] if a society is culturally homogeneous ($\beta=0$):*

(a) For $\rho\in(1-Q_{out},Q_{out})$, welfare decreases with diversity ($\frac{d\widehat{W}}{d\beta}|_{\beta=0}<0$);

(b) For $\rho\notin(1-Q_{out},Q_{out})$, welfare decreases with diversity ($\frac{d\widehat{W}}{d\beta}|_{\beta=0}<0$) if and only if the

[19] However, this does rely on certain payoff parameters being kept fixed as ρ is varied (Online Appendix IV). Without such restrictions, the welfare implications of diversity depend on details of the payoff function beyond ρ. See Kets et al. (2019) for further results on how welfare varies with a game's payoff parameters.

[20] If $\rho<1-Q_{in}$ or $\rho>Q_{in}$, introspective equilibrium is independent of diversity (Lemma A.5).

cost of inefficient lock-in is small and the culture is strong, i.e., $|\underline{m} - \frac{1}{2}| \leq \frac{1}{2}(2q-1)^2$, *where* $\underline{m}$ *is the value of* m *that minimizes* $W(m;\rho)$.

The intuition is for Proposition 3.5 is twofold. First, if no action stands out in terms of payoffs (ρ close to $\frac{1}{2}$), then diversity is always costly, because it increases miscoordination. If one of the actions stands out in terms of payoffs, diversity may improve performance but only if the resulting decrease in inefficient lock-in compensates for a potential increase in miscoordination. But, as noted earlier, a minority that does not have a critical mass ($\beta < \bar{\beta}$) will have a minimal impact on the incentives for the majority. So, a small increase in diversity increases miscoordination but has a limited impact on inefficient lock-in. The former, negative, effect of an increase in diversity dominates the latter, positive effect if the cost of inefficient lock-in is small (i.e., $|W(1;\rho) - W(0;\rho)|$ close to 0, or, equivalently, $\underline{m}$ close to $\frac{1}{2}$) and there is little miscoordination in homogeneous societies (i.e., the culture is strong). So, *the economic benefits of diversity may be nonmonotonic*: while a critical mass of minority players may help avoid inefficient lock-in, a token minority merely leads to more mishaps and miscoordination.

In the context of organizations, Propositions 3.4–3.5 sheds light on the economic effects of quota. If we interpret the expected total payoff $\widehat{W}$ as firm performance, Propositions 3.4–3.5 imply that small quota may reduce firm performance even if larger quotas improve performance. This helps better understand why, for culturally homogeneous organizations, introducing small quotas may reduce firm performance (see Joecks et al., 2013; Chapple and Humphrey, 2014; Liu et al., 2014, for empirical evidence). Proposition 3.5 also has implications for organizational design. To the extent that, for a given organization, the economic benefits of diversity are rooted in its effects on strategic uncertainty and the organization is nearly homogeneous, the organization may benefit from concentrating employees from minority groups in a single division until the groups reach a critical mass.

4 Applications

This section considers the implications of our results in applications.

4.1 Excessive conformism

This section studies how diversity affects conformism. People often have a taste for agreeing with others; yet, conformism can lead a society to coordinate on an outcome that everyone dislikes. One potential explanation is that people falsify their preferences, i.e., they may publicly support an option that they privately oppose under (perceived) social pressure (Kuran, 1987a,b, 1997).[21] Following Kuran (1987b), we model this using a simple generalization of the model (3.1): Each player $j \in N$ wants to choose an action (e.g., support a policy) that is close

[21] Bernheim (1994) considers an alternative explanation. Since his model is nonstrategic in nature, it complements ours. Also related is the work on groupthink (Bénabou, 2013).

to his private preference but also want to match others' actions. That is, the payoff to a player j who chooses action s_j and has private preference $\tau_j \in \mathbb{R}$ is given by

$$-[(1-\lambda)(s_j - \bar{s})^2 + \lambda(s_j - \tau_j)^2],$$

where $\lambda \in (0,1)$, $\bar{s}$ is the average action, and τ_j is a private preference parameter, drawn from a unimodal and symmetric distribution with mean τ and variance $\sigma_\tau^2 > 0$. This is a linear game with risk parameters $\rho_j = \frac{1}{2} + \frac{\lambda}{1-\lambda}(\frac{1}{2} - \tau_j)$.

We study how the scope for conformism varies with the sociocultural environment. Say that there is *excessive conformism* if there is a strictly positive probability that most players choose an action that is inconsistent with their private preference (i.e., $s_j = H$ if $\tau_j > \frac{1}{2}$ and $s_j = L$ if $\tau_j < \frac{1}{2}$). Then, a direct implication of Proposition 3.2 is that homogeneous societies are more prone to excessive conformism than diverse societies, especially when the culture is strong. Intuitively, excessive conformism is a form of inefficient lock-in.[22] In homogeneous societies with a strong culture, people have good sense of the prevailing social climate, and this can lead them to support a position that contradicts their private preferences. By contrast, in diverse societies, players face more uncertainty about which alternative may be culturally salient. As a result, players' choices are more likely to be reflect their private preference.

This finding can help understand the striking observation that many empirical studies of preference falsification focus on homogeneous groups with a strong culture, such as Saudi men on the issue of female labor force participation (Bursztyn et al., 2018)[23] or students at Ivy League universities on affirmative action (Van Boven, 2000). Our model suggests that these empirical findings may not extend to more diverse communities.[24]

While intuitive, the prediction that homogeneous societies tend to be more conformist is difficult to obtain within the standard framework. This is because models of conformism often have multiple equilibria and the standard framework does not explicitly model which equilibrium is selected. This leaves open the question why some societies coordinate on more conformist equilibria than others, and which interventions, if any, can get societies to move to an equilibrium with little conformism.[25] These predictions are also difficult to obtain with models

[22] To see this formally, note that there is excessive conformism if and only if players decisions are *not* driven by payoff considerations.

[23] While Saudi Arabia has a large contingent of immigrant workers, communities are highly homogeneous as the society is segregated along religious and ethnic lines (Glasze and Alkhayyal, 2002).

[24] Our results also help understand why, in the few direct comparisons of homogeneous and diverse societies, preference falsification is more prevalent in homogeneous societies (see Breed and Ktsanes (1961) on racial segregation and Schultz and Neighbors (2007) on excessive alcohol use). Finally, it helps understand the empirical phenomenon that people have a greater tendency to express their true opinions in times of significant cultural change (Noelle-Neumann, 1974) while strong social structures tend to perpetuate misrepresentation of preferences (Sunstein, 2018). This is consistent with our model's prediction that, when the culture is weak (q close to $\frac{1}{2}$), choices tend to be driven by payoff considerations rather than social factors.

[25] Some authors assume that, if a change in fundamentals causes the current equilibrium to disappear, then the society moves to the closest equilibrium (e.g., Kuran, 1987a,b). A problem with this approach is that it presumes that past history fully determines future play. This may be problematic if the equilibrium closest to

where preferences vary with diversity or depend on whom players interact with. For example, a model in which players care only about matching the actions of members of their own group (e.g., because they experience social pressure mostly from their own group) would predict no difference between culturally homogeneous and diverse societies.[26] As another example, preference-based models that posit that people put less weight on others' actions in diverse societies (e.g., λ increases with β in the conformism model) can also not fully explain the data. This is because such models either predict the same outcome for homogeneous and diverse societies, or predict that homogeneous societies have multiple equilibria, thus not fully resolving the equilibrium selection problem. Finally, the prevalence of excessive conformity in homogeneous societies is also difficult to explain by incomplete information about payoffs. Under the arguably natural assumption that people who belong to homogeneous, tight-knit communities have more accurate information about the preferences of others in their community than people who live in heterogeneous communities (Grout et al., 2015), there would be *less* excessive conformism in homogeneous communities, not more.

4.2 Cooperation and trust

This section studies how diversity affects trust and cooperation. We consider a model where players are matched in pairs to play an (infinitely) repeated prisoner's dilemma game. The payoffs of the stage game are given by

	c	d
c	$u_{\mathrm{cc}}, u_{\mathrm{cc}}$	$u_{\mathrm{cd}}, u_{\mathrm{dc}}$
d	$u_{\mathrm{dc}}, u_{\mathrm{cd}}$	$u_{\mathrm{dd}}, u_{\mathrm{dd}}$

where $u_{\mathrm{dc}} > u_{\mathrm{cc}} > u_{\mathrm{dd}} > u_{\mathrm{cd}}$. Players aim to maximize their discounted sum of payoffs $\sum_{t=0}^{\infty} \delta^t u_{s_j^t s_{-j}^t}$, where $\delta \in (0,1)$ is the common discount factor and s_i^τ is the action of player i in period τ. Actions are perfectly observed.

To analyze this setting, we allow players' impulses evolve over time. The key assumption is that past actions may affect today's impulses. For example, if players cooperated yesterday, they may be more inclined to cooperate today. That is, at any time $t > 0$, the probability that each player has an impulse to cooperate at t depends on the actions taken in time $t-1$: Conditional on players having chosen $(s, s') \in \{\mathrm{c}, \mathrm{d}\} \times \{\mathrm{c}, \mathrm{d}\}$ at $t-1$, the probability that they have impulses $(I, I') \in \{\mathrm{c}, \mathrm{d}\} \times \{\mathrm{c}, \mathrm{d}\}$ in t is $\mu(I, I' \mid s, s')$. To focus on the effects of diversity, we consider the simplest possible model: At any time $t > 0$, players have an impulse to cooperate at time t if and only if both players cooperated in the previous period; otherwise, both players

the original one is far inferior in terms of payoffs to some other equilibrium. In such cases, one would expect that the equilibrium selection would be driven by a combination of economic incentives and culturally/historically-determined expectations, as is the case in our model.

[26] Intuitively, if players care only about the actions of members of their own group, a diverse society effectively consists of separate sub-societies, each of which is culturally homogeneous.

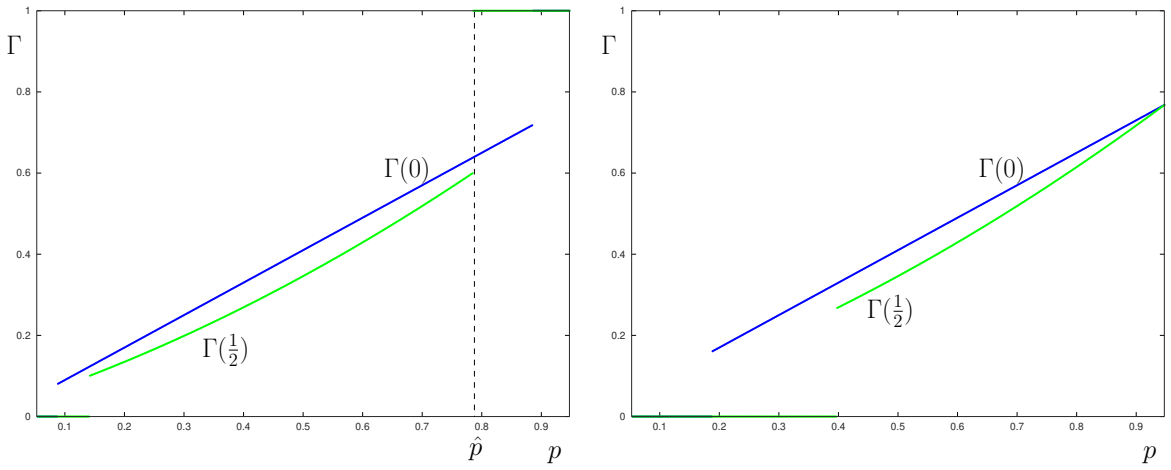

Figure 2: The probability Γ that two players cooperate, for homogeneous ($\beta = 0$) and diverse societies ($\beta = \frac{1}{2}$) for a small cost of being exploited (left panel: $u_{\text{cd}} = 0$) and a high cost of being exploited (right panel: $u_{\text{cd}} = -1$) keeping the other parameters fixed ($u_{\text{cc}} = 5, u_{\text{dd}} = 1, u_{\text{dc}} = 10$, $\delta = 0.6$, and $q = 0.9, \eta = 0.05$).

have an impulse to defect.[27] For the beginning of their relationship (i.e., $t = 0$), we assume that impulses are drawn from the distribution in Section 2.3 but with the slight generalization that the prior probability that cooperating is culturally salient for a group (i.e., $\theta_G = \text{c}$) can be any $p \in (0, 1)$. This specifies a level-0 (behavioral) strategy σ^0. At any level $k > 0$, the level-k strategy σ^k is a behavioral strategy that is a best response to the level-$(k-1)$ strategy σ^{k-1}. The introspective equilibrium of the repeated game is then the limit of the level-k strategies σ^k as $k \to \infty$.

The following result shows that if there is little initial trust (p small) or the cost of being exploited is high (u_{cd} small), then diverse societies have less cooperation than homogeneous societies. To state the result, let $\Gamma(\beta)$ be the cooperation rate in a society with diversity β, that is, the probability that a pair of players cooperates (i.e., both choose H).

Proposition 4.1. [Cooperation Deficit in Diverse Societies] *If there is mistrust or the cost of being exploited is high, then there is less cooperation in diverse societies: For every repeated game $(u_{cc}, u_{cd}, u_{dc}, u_{dd}, \delta)$, there is $\hat{p}$ such that $\Gamma(\frac{1}{2}) \leq \Gamma(0)$ for $p < \hat{p}$ and $\Gamma(\frac{1}{2}) \geq \Gamma(0)$ for $p > \hat{p}$. Moreover, $\hat{p}$ increases as u_{cd} falls, with $\lim_{u_{cd} \to -\infty} \hat{p} = 1$.*

The proof proceeds by showing that the introspective equilibria of the infinitely repeated game coincide with those for a linear game with risk parameter

$$\rho = \frac{(1-\delta)(u_{\text{dd}} - u_{\text{cd}})}{(1-\delta)(u_{\text{cc}} + u_{\text{dd}} - u_{\text{cd}} - u_{\text{dc}}) + \delta(u_{\text{cc}} - u_{\text{dd}})}.$$

This allows us to apply the techniques developed previously for static games to the repeated game.[28] Proposition 4.1 is illustrated in Figure 2. There are two effects. First, in diverse

[27] We focus on this level-0 strategy because it is reminiscent of the grim trigger benchmark. This ensures that the effects of introducing culture will be quite clear. See Kets (2019) for the general model.

[28] The equivalence result with linear games presumes that players are matched uniformly at random and do not observe the group of the player they are matched with. This allows us to analyze the game in the same way

societies, players' decisions are more likely to be guided by payoff considerations than in homogeneous societies (Proposition 3.2). So, when the conditions for cooperation are unfavorable (in the sense that the cost of being exploited is high), players in diverse societies refrain from cooperating (i.e., defect in every period) for a larger range of parameters. This explains why the transition from a regime with no cooperation ($\Gamma(\beta) = 0$) to positive levels of cooperation ($\Gamma(\beta) > 0$) in Figure 2 occurs at a higher value for p for diverse societies. In words, diverse societies require a higher level of initial trust to escape a no-cooperation trap. Second, even when some players try to initiate cooperation (i.e., choose grim trigger), cooperation is more likely to succeed in homogeneous societies. This is because miscoordination can thwart cooperation: If, in a given pair of players, one player attempts to initiate cooperation at $t = 0$ (i.e., chooses grim trigger) while the other defects (chooses always defect), the pair will end up defecting in any period $t > 0$. Because there is more miscoordination in diverse societies ($d > 0$), this risk is greater in diverse societies. This explains why the cooperation rate is lower in diverse societies for intermediate levels of trust in Figure 2 ($0 < \Gamma(\frac{1}{2}) < \Gamma(1) < 1$).When the cost of being exploited increases (u_{cd} falls), the cost of miscoordination increases and the scope for cooperation in diverse societies further declines and may even disappear altogether ($\hat{p} \to 1$), as illustrated in the right panel of Figure 2. In fact, these results hold more broadly: the proof of Proposition 4.1 shows that the result extends to any change in payoff parameters that increases the risk parameter ρ (e.g., a fall in the discount factor δ or the benefits u_{cc} to cooperation).

So, *diverse societies have lower levels of cooperation when the economic conditions for cooperation are not very favorable (i.e., ρ high) and there is little initial trust (i.e., p close to 0).*

This result is in line with the empirical finding by Knack and Keefer (1997) that strategic uncertainty limits trust and cooperation in diverse societies. As they note, "individuals [in diverse societies] are less likely to share common backgrounds and mutual expectations about behavior, so it is more difficult to make self-enforcing agreements" (p. 1278). To the best of our knowledge, our model is the first to formalize this argument. The lack of trust and cooperation in diverse communities is often attributed to aversion to heterogeneity (Alesina and La Ferrara, 2002). Under this view, people tend to distrust those who are dissimilar from themselves. Our model can be viewed as providing foundations for this phenomenon. That is, rather than directly assuming that people distrust members of other groups, we show that aversion to heterogeneity can arise if people face more strategic uncertainty in diverse societies. In particular, because there is more miscoordination in diverse societies ($d > 0$), players' trust is more likely to be betrayed in these societies: players from diverse societies who attempt to initiate cooperation are more likely to encounter opponents who defect than players from homogeneous societies. This is true even if the prior probability that an individual defects is the same across groups, simply because people from the same group are more likely to agree on whether it is culturally salient to cooperate. This leads to novel implications: For example,

as a model with a continuum of players. However, it is not critical. Our results extend to settings where group membership is observable.

a reduction in discount factor (e.g., due to an increase in mobility) or an increase in the cost to being exploited has a larger negative impact in diverse communities because strategic uncertainty reinforces their deleterious impact.

These findings can also aid in designing policies to restore trust and cooperation. If, as our results suggest, aversion to heterogeneity is not necessarily a hard-wired distaste for interacting with others, then cooperation can be promoted by reducing strategic uncertainty (e.g., developing shared expectations about when to cooperate) or improving trust (i.e., increase p). At the same time, our model suggests that restoring trust may be nontrivial. If the level of initial trust decreases over time when past attempts to build cooperation have been unsuccessful, then our model suggests that diverse societies may become stuck in a low-trust trap while homogeneous societies may successfully build trust and improve cooperation. Thus, small initial differences between societies may increase over time, and without any interventions, distrust between groups may persist. This is in sharp contrast with the argument of the influential political scientist Robert Putnam (2007) that aversion to heterogeneity is likely to vanish over time as people become more familiar with people from other groups. Our model suggest that this is not necessarily the case: In diverse societies, people may be reluctant to initiate cooperation because they fear others may not reciprocate.[29] If that is the case, people from diverse societies will not experience the cooperative interactions with members of other groups that are necessary to improve trust, and diverse societies will remain stuck in a low-trust trap.

These predictions are difficult to obtain using existing models. A general difficulty is that many models of cooperation have multiple equilibria. Hence, these models cannot explain why some societies coordinate on more cooperative outcomes than others, and which interventions, if any, can move a society to a more cooperative outcome. This is the case, for instance, for the standard repeated-games framework or many models of group-dependent social preferences.[30] Another challenge is that many existing models abstract away from strategic uncertainty by positing that players coordinate on a Nash equilibrium, typically the cooperative one. These models are thus silent on why cooperation can be difficult to build. These models are therefore unable to explain recent empirical evidence that strategic uncertainty can be a major impediment to collusion (Byrne and De Roos, 2019) or that exogenous shocks to the level of strategic uncertainty can affect the scope for cooperation (Knittel and Stango, 2003). Finally, these models cannot explain the experimental finding that factors (such as the cost of being exploited) that do not affect whether cooperation can be sustained as a (Nash or subgame perfect) equilibrium may matter for cooperation (Dal Bó and Fréchette, 2018).[31]

[29] See Guiso et al. (2008) for a related argument. Guiso et al. do not consider the effects of diversity.

[30] The problem is that in many models, defection can always be supported in equilibrium even when cooperation is also an equilibrium, such as when players are sufficiently patient in the standard repeated games setting or when players get a positive payoff from cooperating with members of their own group, as in the model of Tabellini (2008)).

[31] Blonski et al. (2011) and Dal Bó and Fréchette (2011) explain this experimental finding by requiring that players cooperate only if it is risk dominant in the reduced game (i.e., $\rho < \frac{1}{2}$).

4.3 Organizational culture

This section studies how an organization's culture can impede the adoption of superior management practices. It has been well documented that competitively significant practices are slow to diffuse even in the absence of the usual informational frictions or incentive problems and that this can lead to persistent performance differences across organizations (Gibbons, 2010; Gibbons and Henderson, 2013). To study this, we consider a setting where a manager can incentivize agents to take a certain action but where incentive costs are influenced by the organization's culture. Following Holmstrom and Milgrom (1994), we allow work practices to be complements. Thus, work practices can be organized into "clusters," with each cluster consisting of work practices that complement each other. There are two clusters, denoted H and L. Because work practices are complements, employees have a dual objective: they respond to incentives but they also want to coordinate their choice with those of others. This basic tradeoff can be captured with the simple model from Section 3.1. Thus, the payoff to an employee j who chooses a work practice consistent with cluster $s \in \{H, L\}$ is

$$-[(1-\lambda)\,(s_j-\bar{s})^2+\lambda\,(s_j-\tau)^2],$$

where τ is an incentive payment. The manager chooses the incentive payment τ to maximize profits. Suppose that profits are maximized when employees choose practices consistent with the high cluster (i.e., all choose H). For example, H could consist of high productivity practices. Moreover, because work practices are complements, profits suffer if there is miscoordination (i.e., some employees choose practices from H, while others choose practices from L). So, the manager chooses the incentive payment τ to maximize

$$\Pi=-\mathbb{E}_\tau\big[\Lambda\,(1-m)^2+(1-\Lambda)\,m\,(1-m)\big]-c(\tau)$$

where $\Lambda \in (0,1)$ is the weight that the manager puts on employees choosing H (which may differ from λ), $c(\tau) \geq 0$ is the cost of providing incentives τ, and the expectation is taken over the share m of employees choosing H in introspective equilibrium (which depends on τ). For simplicity, we take the cost of τ to be linear. We assume that it is equally costly to incentivize H or L. Thus, $c(\tau) := |\tau - \frac{1}{2}|$ is linear and symmetric in $\tau = \frac{1}{2}$.

We consider the question of how an organization's culture affects the choice of incentives. To focus on the impact of the organization's culture, we take the organization to be culturally homogeneous ($\beta = 0$). The key parameters are then the culture's strength q and the prior probability $p \in (0,1)$ that the high cluster is culturally salient ($\theta = H$). If p is close to 1, then the high cluster is culturally dominant in the sense that H is likely to be culturally salient. So, if p is close to 1, the organization's culture is aligned with the manager's objectives; but if p is close to 0, then the organizational culture and the manager's objectives conflict. The following result shows how the cost of incentivizing all employees to choose practices from the high cluster vary with an organization's culture. To state the result, denote by $c_H(p,q)$ the minimum cost of incentives that ensure that all players choose H in introspective equilibrium

Proposition 4.2. [Organizational Culture & Incentive Costs]

(a) The incentive cost is minimized if the organization's culture and objectives are aligned (i.e., $c_H(p,q)$ decreases in p)

(b) If the organization's culture and objectives are aligned, the incentive cost is minimized if the culture is strong; otherwise, the cost is minimized if the culture is weak (i.e., there is p^* such that for $p \in (p^*,1)$, $c_H(p,q)$ decreases with q while for $p \in (0,p^*)$, $c_H(p,q)$ increases with q).

Proposition 4.2(a) shows that implementing practices that are not culturally dominant can be costly. This is true even though the manager's and employees' incentives are essentially aligned in the sense that all employees choosing H is a Nash equilibrium even if no incentives are provided (i.e., $m = 1$ is a Nash equilibrium for $\tau = \frac{1}{2}$). Proposition 4.2(a) has the important implication that the manager may choose *not* to implement H if this conflicts with the organization's culture (i.e., p close to 0). Proposition 4.2(b) says that if objectives and culture are aligned (p high) then it it is optimal to have a strong culture, but if there is a conflict between the two, then it is optimal to have a weak culture. Intuitively, if H is culturally dominant (i.e., p close to 1), then, in organizations with a strong culture (q close to 1), a large share of employees expects H to be culturally salient in the high probability event that H is culturally salient. Thus, strengthening the culture reduces the cost of implementing H. Conversely, if L is culturally dominant (i.e., p close to 0), in organizations with a weak culture, only a small share of employees expect L to be culturally salient in the high probability event that L is culturally salient. This makes it easier to implement H. Thus, if there is a conflict between the manager's objectives and the organization's culture, then the cost of implementing H is smaller when the culture is weaker.

These findings can help better understand why work practices may persist even if they have become dysfunctional. Proposition 4.2(a) shows that "culture costs" may distort manager's choices when there is strategic uncertainty: if the incentive cost is too high, then a manager may choose not to incentivize employees to choose H. This is akin to how agency costs may distort the principal's choice when there is asymmetric information. However, an important difference is that in our setting, a distortion in choices does not require a conflict of interest. Culture costs and their associated distortions can help better understand persistent performance differences across organizations: superior practices may be slow to diffuse if they conflict with the prevailing culture of organizations. The U.S. steel industry, which underwent significant changes during the 1980s and 1990s, provides an interesting case with which to confront the predictions of our model. The work practices in the steel industry are complements and can be divided into clusters (Ichniowski et al., 1995). Even though traditional work practices had long become obsolete and the benefits of switching to more modern work practices were substantial, many plants remained locked into inefficient practices even though the benefits were well known and investment costs were limited (Ichniowski et al., 1995, p. 48). In line with our results, Ichniowski et al. (1995, p. 52) note that switching to efficient practices often requires hiring

new workers. Likewise, Brynjolfsson and Milgrom (2013, pp. 13–14) observe that it may be optimal to build a new culture from scratch at "greenfield sites" or an isolated "skunkworks." This is broadly consistent with our predictions that changing work practices may require a change in culture (i.e., p) or weakening the culture (i.e., reduce q).

These findings can also help better understand the insight from the management literature that in volatile economic environments, companies with a weak culture outperform companies with a strong culture (Kotter and Heskett, 1992): When a change in economic conditions makes it optimal to change work practices, companies with a weak culture (q close to $\frac{1}{2}$) are able to make that switch by adjusting incentives while companies with a strong culture keep the incentives unchanged even if that means that workers choose inefficient practices.

While these results are intuitive, they are difficult to obtain with standard models. Some authors explain that performance differences between organizations may persist if there are multiple equilibria, with some organizations being locked into inefficient equilibria (Kreps, 1990, 1996) (also see Gibbons and Henderson, 2013). However, these theories cannot explain why organizations with a weak culture may avoid inefficient lock in.[32] These theories are also silent on what might induce an organization to change its practices (unless changes to the economic environment are so large that the set of equilibria changes). By contrast, within our framework, it is possible to show the intuitive result that changing practices becomes more attractive for a manager if the benefits of implementing superior practices increases (i.e., Λ increases) or if employees become more sensitive to incentives (λ increases). Our findings also complement those from the literature in organizational economics that focuses on challenges associated with coordination (Crémer, 1993; Prat, 2002; Dessein and Santos, 2006; Crémer et al., 2007; Dessein et al., 2015). This important literature generally abstracts from the coordination problems that drive our results by assuming that agents coordinate on the efficient outcome (subject to, e.g., informational constraints). By contrast, under our modeling approach, some organizations coordinate on more efficient practices than others.

4.4 Regime change

This section studies how diversity affects the scope for regime change. We consider a regime (autocratic ruler) that is overthrown when sufficiently many people revolt. Citizens who attack the regime (choose H) benefit if the regime is ousted but pay a cost if too few people revolt. Not attacking (choosing L) costs 0. Thus, this is a threshold game with payoffs given by

$$\begin{aligned} u(H, s_{-j}) &= \begin{cases} B & \text{if } m \geq T; \\ -C & \text{otherwise;} \end{cases} \\ u(L, s_{-j}) &= 0; \end{aligned}$$

where $m = m(s_{-j})$ is the proportion of players who choose H under s_{-j}; see, e.g., Persson and Tabellini (2009), Egorov et al. (2009), Bueno De Mesquita (2010), Fearon (2011), Little (2012),

[32] Similar comments apply to dynamic models that emphasize path dependence (e.g., Arthur, 1989).

Edmond (2013), and Boix and Svolik (2013) for related models of regime change in political science and economics.

We study policies that the regime can employ to reduce the threat of being ousted, with a focus on how their effectiveness is influenced by sociocultural factors. For expositional simplicity, we focus on (maximally) homogeneous ($\beta = 0$) and diverse societies ($\beta = \frac{1}{2}$) and assume that attacking and not attacking are equally likely to be culturally salient ($p = \frac{1}{2}$).

We start with measures that make the population more homogeneous (e.g., nation-building policies). The following result, which is a corollary of Proposition 3.3, shows that such measures are effective at reducing the threat to the regime if the regime is weak but not if it is strong.

Corollary 4.3. [Homogenization] *For weak regimes (i.e., $T < \frac{1}{2}$), the probability of a successful attack decreases when groups become more similar (i.e., d decreases); but for strong regimes ($T > \frac{1}{2}$), the reverse is true: the probability of a successful attack increases when groups become more similar.*

Corollary 4.3 shows that regimes benefit from increasing strategic uncertainty (increasing d) when they are strong but not when they are weak. The first part of Corollary 4.3 says that if the regime is weak ($T < \frac{1}{2}$), reducing the cultural distance between groups benefits the regime in that it reduces the likelihood of a successful attack. To see the intuition, note that the regime is ousted whenever it is culturally salient for one of the groups to attack (assuming T is not too large, i.e., $T < q^2 + (1-q)^2$). In the extreme case where groups are maximally different ($d \uparrow 1$), the event that attacking is culturally salient for one group ($\theta_G = H$) is nearly independent of the event that it is culturally salient for the other group ($\theta_{G'} = H$). Hence, the probability that a weak regime is ousted when groups are highly dissimilar is close to $\frac{3}{4}$. By contrast, in the other extreme case where groups are nearly identical ($d \downarrow 0$), the two events are almost perfectly correlated. Hence, the probability that the regime falls when groups are nearly identical is close to $\frac{1}{2}$. The second part of Corollary 4.3 states that if the regime is strong ($T > \frac{1}{2}$), it is more likely to stay in power when groups are dissimilar (d large). Intuitively, when the regime is strong, overthrowing the regime requires concerted action by players from both groups. Because the probability that attacking is culturally salient for both groups is close to $\frac{1}{2}$ when groups are almost identical ($d \downarrow 0$) but only close to $\frac{1}{4}$ when the groups are highly dissimilar ($d \uparrow 1$), a strong regime benefits from having groups that are dissimilar from each other.[33]

Corollary 4.3 has a number of interesting implications. First, it suggest that weak regimes are more likely to invest in homogenization policies (i.e., reduce d), such as the adoption of a state religion or a national language. This can help better understand the empirical finding that weak regimes are more likely to invest in nation-building through homogenization policies

[33]This intuition is related to the argument by Fearon (2011) that coordination may be more difficult if the signals that players receive about payoffs are noisy. However, an important difference (beyond the fact that we are concerned with strategic uncertainty as opposed to payoff uncertainty) is that in Fearon's setting, homogenization (i.e., reducing noise) always increases the likelihood of regime change, independent of regime strength.

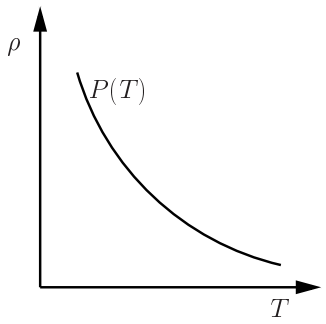

Figure 3: The set of feasible policies $(T, P(T))$ for given policy function $P(T)$.

(Alesina et al., 2018). On the other hand, strong regimes benefit from having distinctive groups, in line with the work of Acemoglu et al. (2004).[34] The current model unifies these findings, showing that each can be explained by how strategic uncertainty interacts with regime strength.

We next study how diversity affects a regime's choice to invest in state capacity (i.e., increase T). We focus on the case where the regime can increase its strength T at the expense of increasing the payoff B to a successful revolt. Intuitively, investing in state capacity requires diverting resources from productive uses (e.g., investment in public goods), which increases the gains from a successful revolt. Assuming that the cost C of an unsuccessful revolt is fixed,[35] the regime then faces a tradeoff between increasing T and decreasing $\rho = C/(B + C)$. That is, the regime can choose any combination of T and ρ on a downward-sloping function $P(T)$. This is illustrated in Figure 3: when the regime invests in state capacity (increase T), the quality of government deteriorates (B increases) and the incentive $1 - \rho$ to attack increases (ρ decreases). We refer to a pair $(T, P(T))$ as a (feasible) policy and to the function $P(T)$ as the policy function. Regimes want to stay in power: Given any pair of feasible policies $(T, P(T))$ and $(T', P(T'))$, the regime prefers the policy that minimizes the probability that it is overthrown. For ease of exposition, we also assume that if the probability of a successful attack is the same under two policies $(T, P(T))$ and $(T', P(T'))$, then it prefers the policy with the lowest investment in state capacity. The next result shows that a regime's incentive to invest in state capacity depends on whether the society is homogeneous or diverse:

Proposition 4.4. [Endogenous State Capacity] *Regimes in diverse societies tend to invest more in state capacity than in homogeneous societies and have a lower quality of government. That is, if $T(\beta)$ and $B(\beta)$ are the state capacity and quality of government chosen by a regime when diversity is β, then $T(\frac{1}{2}) \geq T(0)$ and $B(\frac{1}{2}) \geq B(0)$, with strict inequalities for some policy functions.*

Proposition 4.4 states that regimes invest more in state capacity when the society is diverse. The intuition builds on Proposition 3.3. The key insight is that, in diverse societies,

[34] In Acemoglu et al.'s (2004) work, these divide-and-rule strategies are used to buy off different groups. Our model implies that these strategies can be successful even when the regime cannot make (differential) transfers to groups.

[35] This is not essential to our results. Our results extend if a regime can invest in repression (i.e., increase C) as long as there is a reduction of productive investments.

miscoordination makes mass uprisings and protests more difficult when the regime is strong, but not when it is weak. By contrast, in homogeneous societies, the effects of miscoordination are largely independent of regime strength. Proposition 4.4 suggests that diverse societies may invest in state capacity at the expense of productive investments. This can help understand why diverse societies are more likely to have poorer governance, less growth, and lower GDP per capita (e.g., Easterly and Levine, 1997; La Porta et al., 1999; Alesina et al., 1999, 2003). While we do not wish to suggest that our simple model can explain this complex empirical phenomenon, it opens up the question whether other factors can contribute to the relatively poor economic performance of diverse societies beyond those that are typically considered, such as excessive conflict due to preference heterogeneity (Alesina and Spolaore, 1997, 2003; Alesina et al., 2000) or wasteful spending to elicit the support of citizens (Padró i Miquel, 2007). Our model suggests that reducing conflicts or preference heterogeneity may not necessarily improve economic performance in diverse societies if there is significant strategic uncertainty.

Our model thus provides a unifying account of how diversity affects regime change and what this means for economic outcomes. In addition, our model also delivers novel predictions. This is because the costs and benefits of diversity stem from strategic uncertainty as opposed to preference heterogeneity, as in much of the literature. For example, while in Alesina et al. (2018), weak regimes introduce policies to make the preferences of the citizens more aligned with the regime's, our model instead suggests that regimes may choose to make preferences *less* aligned (i.e., increase B) if that helps stabilize the regime (increase T). Our model can help better understand empirical regularities that are difficult to capture with models where the costs and benefits of diversity are driven by preference heterogeneity. For example, our model sheds light on why autocratic regimes specifically censor social media posts that reduce strategic uncertainty (and even fabricate posts that increase strategic uncertainty) yet leave more payoff-relevant information (e.g., about grievances) untouched (King et al., 2013).[36]

While we focus on mass protests and revolutions for concreteness, our insights extend to other settings where interactions take a threshold form, such as bank runs (Goldstein and Pauzner, 2005), currency attacks (Morris and Shin, 1998), and debt crises (Morris and Shin, 2004). The techniques developed here may provide useful for these other settings as well. In particular, our framework makes it possible to endogenize the threshold T, which has proven challenging with existing methods (Angeletos et al., 2006). This opens up the possibility to study questions that range well beyond the issue of diversity, such as when it is optimal for a fixed-exchange rate regime to take costly actions to defend its currency or under which conditions a firm invests to enable it to meet short-term claims from its creditors to ensure its long-run survival.

[36]The risk to the regime of information that reduces strategic uncertainty is also highlighted by, e.g., Kuran (1989, 1991), Lohmann (1994), and Lorentzen (2013). However, in their models, the regime is essentially passive and equilibrium outcomes do not depend on strategic interactions between the regime and the citizens. Moreover, these models typically feature multiple equilibria, which makes difficult to derive testable comparative statics.

5 Discussion

Our results suggest that introspective equilibrium provides a unifying framework to help understand a variety of seemingly disparate empirical phenomena. However, the bar for introducing a new solution concept is – and should be – high. We therefore comment here on some of the key properties of our concept.

At the most general level, *introspective equilibrium provides a new way to deal with equilibrium multiplicity*. In our model, culture matters for equilibrium selection but is not decisive: *the outcome selected depends on the interplay of economic and sociocultural factors*:[37] While culture shapes the impulses that anchor the introspective process, the introspective process also interacts with the economic environment. This makes it possible to explain why some societies coordinate on inferior outcomes, and which interventions, if any, can help them move to a better outcome. This is in sharp contrast with the literature that views culture as an equilibrium selection device (Kreps, 1990; Myerson, 2004). While this alternative approach can help understand why societies that are fundamentally different in all payoff-relevant aspects often behave very differently, it is not well-suited to predict how economic outcomes vary with economic primitives (e.g., when the payoffs to an unplayed equilibrium improve), a question that is obviously of central importance for both comparative statics and policy analyses.[38]

Our results can also not be easily replicated by replacing impulses with random preference shocks. The impulses help select an equilibrium that depends on both economic and sociocultural factors. By contrast, when preference shocks are used to select an equilibrium (as in the global games approach), the equilibrium selection is independent of culture.[39] So, unlike our model, equilibrium selection methods based on preference shocks cannot explain why some societies select a better equilibrium than other societies that face the same economic environment. Another important difference is that a global games analysis abstracts from the cost of miscoordination. Hence, it cannot capture the central tradeoff between miscoordination and inefficient lock-in that drives our results. This also means that a global games analysis may not be suitable for analyzing the welfare consequences of moving to another equilibrium. To better understand this point, suppose that payoffs are such that a society coordinates on a Pareto-dominated Nash equilibrium (say, all choose L) and that the Pareto-optimal equilibrium becomes more attractive, so that some players choose the efficient action (i.e., choose H). Our methodology can be used to show that welfare may decrease or increase depending on whether

[37] The feature that outcomes depend on both payoffs and the broader environment is also familiar from learning models (e.g., Crawford, 1995). However, there are few papers that systematically explore how the interaction of payoffs and the broader social environment affect equilibrium selection, notable exceptions being Blume (1993, 1995) and Morris (2000), who focus on the role of network structure.

[38] In some cases where extensive data on the relevant culture is available, as in Akerlof (1976, 1980) and Greif (1994, 2006), this approach can shed important light on the economic implications of culture (i.e., a given equilibrium selection). However, even if we have good information on which equilibrium is selected by a given society at a given time, this approach cannot deliver comparative statics unless the set of equilibria changes.

[39] The same holds for the tracing procedure (Harsanyi and Selten, 1988) and many learning and evolutionary models (e.g., Kandori et al., 1993; Young, 1993).

the cost of miscoordination exceed the cost of inefficient lock-in (Kets et al., 2019). This means, for example, that policies that aim to stimulate investment but that fall short of attaining full investment (as in Morris and Yildiz, 2019) may make people worse off if investing is risky, a critical point that might be missed in a global games analysis, but that emerges naturally with introspective equilibrium. Finally, using preference shocks to select an equilibrium limits the questions that can be addressed. For example, it is difficult with a global games approach to endogenize a regime's decision of how to defend itself, unlike with our model (Proposition 4.4).[40]

One concern one might have about our approach is that impulses are generally unobservable, at least in applications.[41] While the unobservability of non-economic factors is a concern in economic studies of culture more generally (Guiso et al., 2006), we would argue that this concern is minimized in this case because we focus on comparative statics. A central feature of our approach is that *we obtain the same qualitative comparative statics on economic incentives and diversity* regardless *of the exact assumptions on impulses or culture*. Put another way, while the point predictions of our model may be difficult to test when sociocultural factors are unobservable, the model nevertheless provides testable comparative statics *even if* the details of a culture or impulses are unobservable. In this respect, we go a step further than the behavioral game theory literature that estimates the relevant behavioral parameters from data.[42] Instead of assuming that the relevant parameters can be estimated from data, we focus on predictions that are *independent* of the precise behavioral parameters.

Another potential concern is that cultural salience is independent of payoffs in our model. Separating cultural beliefs from economic incentives has the important advantage that it allows us to conduct comparative statics in two dimensions independently: We vary the payoffs in the game and then ask how diversity affects behavior and welfare for any given economic environment (i.e., payoffs). But while the assumption that impulses are driven by culture might not be unreasonable for decisions that have a strong cultural, moral, or ideological component, in other settings this assumption might perhaps be too strong. However, our results are robust to relaxing this assumption: our results go through as long as impulses are not *entirely* driven by payoffs.

Another point to note is that our approach is "detail-free" in that it is agnostic as to exact set of circumstances under which an individual has an inclination to take a given action. That is, while we view impulses as being driven by salience and other contextual factors, we do not take a stance on which action is salient in which particular context for which particular group.

[40] This is because, to obtain uniqueness using the global games approach, the regime's choices need to be uncertain whereas they are known in equilibrium (by definition) (Angeletos et al., 2006). One way around this is to study ex ante investments (i.e., before the regime knows its own capacity) (Morris and Shin, 1998, 2003). However, in at least some cases, it may be difficult for the regime to commit in this way.

[41] See Agranov et al. (2015), however, for an innovative protocol that makes it possible to observe subjects' instinctive reactions in experimental settings.

[42] Examples of behavioral parameters that are estimated from the data include the cursedness parameter in Eyster and Rabin (2005), the rationality parameter in quantal response equilibrium (McKelvey and Palfrey, 1995), or the fraction of level-k players in level-k models (Crawford et al., 2013).

Instead, we focus on how the aggregate patterns of behavior vary across different socioeconomic environments. While this means we lose some of the richness of more detailed models (e.g., Bordalo et al., 2013), it has the advantage that it allows us to provide a unified framework for analyzing seemingly disparate phenomena. A related point is that we do not endogenize the reasoning process as the outcome of a cost-benefit analysis (cf. Alaoui and Penta, 2016). We leave the important question of microfoundations for future work.

Finally, because introspective equilibrium is the outcome of a reasoning process, proving existence is more challenging than for other equilibrium concepts. The key difficulty is to show that the strategies converge. This requires us to develop new tools. For example, to prove existence in linear games with heterogeneous preferences, we use a novel monotonicity argument based on a change of variables (Lemma A.3). While our existence results cover a range of games of applied interest,[43] it is an open question to what extent they can be extended to other games. However, because the level-k strategies are well-defined at every finite level k, the methodology might still be useful to make predictions about the behavior of players who reason through only finitely many levels even for games for which convergence cannot be shown.

6 Related literature

In this section, we summarize related work on diversity not discussed elsewhere in the paper. Because the literature on diversity is too vast to survey here, we focus here on the theory literature in economics; see Alesina et al. (2016) and Laitin and Jeon (2015) for excellent surveys of the empirical literature and the literature outside of economics, respectively. Our work largely complements the existing literature on diversity. For example, an important literature shows that diversity can be costly if it is associated with preference heterogeneity and conflict (Esteban and Ray, 1994; Alesina and La Ferrara, 2005; Van den Steen, 2010) or when groups have group-dependent social preferences (Chen and Chen, 2011), while other prominent literatures show that diversity can be beneficial if diverse groups have access to more skills (Lazear, 1999a,b; Alesina et al., 2000; Hong and Page, 2001; Prat, 2002; Page, 2007) or if differences in opinions provide incentives to acquire costly information (Che and Kartik, 2009; Van den Steen, 2010). Relative to these literatures, we identify a new driver for the effects of diversity: In our model, a shared culture reduces strategic uncertainty. This implies that diversity matters even if groups are identical in all payoff-relevant respects. As we have shown in the context of applications, whether the effects of diversity are driven by differences in cognition or by factors that directly affect preferences matters for predictions. At a more fundamental level, our work complements the existing literature by focusing on culture rather than on identity (Akerlof and Kranton, 2000). This distinction has long been recognized in psychology, where it has been argued that shared cultural beliefs serve coordination and communication functions (Zou et al., 2009) while identity serves to differentiate oneself (often

[43] For other applications, see Akerlof and Holden (2017), Akerlof et al. (2017), Kets and Sandroni (2019), Kets et al. (2019), and Kets (2019).

positively) from other groups (Tajfel and Turner, 1986). This has implications for predictions. For example, Chen and Chen (2011) show that if people care about people with the same identity ("ingroup members") but not about outgroup members, then players are better able to avoid inefficient lock-in when they interact with ingroup members. By contrast, our model predicts that there is less inefficient lock-in in diverse societies.[44] Finally, a central feature of our approach is that the economic costs and benefits of diversity derive from a common mechanism. This allows us to provide a unified account of a variety of disparate evidence. It also has the important implication that the costs of diversity can generally not be separated from its benefits. This means that policies that affect the level of diversity in a particular context may have spillover effects to other settings.

Also related is the literature on cultural transmission and the persistence of diversity (Bisin and Verdier, 2000, 2001; Kuran and Sandholm, 2008). This literature studies the long-run evolution of culture and characterizes the conditions under which diversity can be sustained. By contrast, we largely abstract away from dynamic considerations and focus on the effect on economic outcomes of diversity for a given sociocultural (and economic) environment. The question of how cultural beliefs and economic behavior coevolve is a fascinating one that we leave for future research.

Appendix A Proofs

A.1 Proof of Proposition 2.1

We begin by defining the class of games. We generalize the model in Section 2. In particular, we do not require that the game has strategic complementarities here. We impose some minimal regularity conditions to avoid measurability problems. We prove the results for a simple generalization of the model in Section 2 (in particular, games with a continuum of players and finite strategy sets). However, the results can easily be extended to other classes of games with minor modifications. The key steps in the proofs are to show that (i) the level-k strategies are measurable; and (ii) players' expected utility converges when the level-k strategies converge. These results are straightforward to prove for many commonly studied games.

Basic definitions The set of players is $N = [0, 1]$. For simplicity, we assume that each player has the same finite set S of actions. Players belong to different cultural groups. That is, the set of players is partitioned into a finite set $\mathcal{G}$ of groups. Each group $G \in \mathcal{G}$ contains a continuum of (identical) players, and players know which group they belong to. Denote the proportion (measure) of players who belong to group G by $\alpha_G \in [0, 1]$ (so $\sum_G \alpha_G = 1$). The distribution of impulses is a direct generalization of that in Section 2.3 to multiple groups: A (random) state $\theta = (\theta_G)_{G \in \mathcal{G}}$ is drawn

[44] In a recent experimental paper, Le Coq et al. (2015) present evidence that subjects may have different beliefs depending on whether they interact with ingroup or outgroup members; in particular, subjects behave as if they face more strategic uncertainty when interacting with the outgroup, consistent with our model. However, their results are driven by a different mechanism than ours: their experimental treatment is not designed to have a consistent effect on beliefs. Without a systematic effect on beliefs, our mechanism cannot operate.

according to a common prior π, where for each group G, θ_G takes values in some finite set Θ_G. For each group G, the impulses of players from group G are independent conditional on the realization of θ_G. Then, for each group G, state θ_G, and action s, there is $p_{\theta_G}(s) \in [0,1]$ such that the realized proportion of players in G who have an impulse to play s is $p_{\theta_G}(s)$ (with probability 1). Players know their own impulse.

A player's payoff depends on his own action and the proportion of players choosing each action, i.e., for each player $j \in N$, $u_j = u_j(s_j, (m_s)_{s\in S})$ where m_s is the proportion of players $j' \neq j'$ choosing action $s \in S$. We consider two cases: (1) players have identical preferences (i.e., $u_j = u(\cdot)$ for all $j \in N$; or (2) players have heterogeneous preferences: each player has a payoff type, i.e., $u_j = u(s_j, (m_{s'})_{s'\in S}; \boldsymbol{u}_j)$, where $\boldsymbol{u}_j$ is drawn from a continuous distribution $\widetilde{F}(\boldsymbol{u}_j)$ on a subset $\mathcal{U}$ of a finite-dimensional Euclidean space, independently across players and independently of impulses and of the state θ. To unify notation, we write $u_j = u(\cdot; \boldsymbol{u}_j)$ for the payoff of each player; in case (1), it is understood that there is $\boldsymbol{u} \in \mathcal{U}$ such that $\boldsymbol{u_j} = \boldsymbol{u}$ for all $j \in N$. We take $u(s_j, (m_{s'})_{s'\in S}; \boldsymbol{u}_j)$ to be continuous in $(m_{s'})_{s'\in S}$ and $\boldsymbol{u}_j$ for all $s_j \in S$. Payoffs are commonly known.

Anonymous strategies Recall that a strategy σ_j for a player $j \in N$ maps an impulse $I \in S$ into an action $\sigma_j(I) \in S$. A collection of strategies, $\{\sigma_j : S \to S : j \in N\}$ is *anonymous* if a player's strategy does not depend on his player label, but only on his group and possibly payoff type (in case (2)). That is, if preferences are heterogeneous (case (2)), the collection $\{\sigma_j : S \to S : j \in N\}$ of strategies is anonymous if for every $\boldsymbol{u} \in \mathcal{U}$, and $G \in \mathcal{G}$, there is $\sigma_{G,\boldsymbol{u}} : S \to S$ such that for every player $j \in N$ with payoff type $\boldsymbol{u}$ who belongs to group G, $\sigma_j(I) = \sigma_{G,\boldsymbol{u}}(I)$ for $I \in S$. With some abuse of terminology, we refer to $\sigma_{G,\boldsymbol{u}}$ as an *(anonymous) strategy* and we write $\sigma := (\sigma_{G,\boldsymbol{u}})_{G\in\mathcal{G},\boldsymbol{u}\in\mathcal{U}}$ for the strategy profile $(\sigma_j)_{j\in N}$; we sometimes denote it by $\sigma : S \times \mathcal{U} \times \mathcal{G} \to S$. We will use the same notation for the case of identical preferences (case (1)); in this case, it is understood that $\mathcal{U} = \{\boldsymbol{u}\}$.

If players follow anonymous strategies, we can summarize each player's characteristics by a triple $(I, G, \boldsymbol{u})$ that specifies the player's impulse $I \in S$, group $G \in \mathcal{G}$, and payoff type $\boldsymbol{u} \in \mathcal{U}$. Denote the expected payoff of a player with an impulse $I \in S$ and payoff type $\boldsymbol{u}$ from group G when he takes action s and the other players follow an anonymous strategy profile σ by $U(s, (m_{s'}(\sigma))_{s'\in S}; I, G, \boldsymbol{u})$. We will refer to the triple $(I, G, \boldsymbol{u})$ as the player's *type*, noting that the payoff type $\boldsymbol{u}$ is the only characteristic that directly affects payoffs; the impulse I and group G affect the player's beliefs about other player's impulses.

For this general setting, some care needs to be taken in defining the level-k strategies. We require that, if at some level $k > 0$, players follow an anonymous strategy and there exists an anonymous strategy that is a best response to the level-k strategy, then the level-$(k+1)$ strategy is anonymous. Likewise, if at each level k, the level-k strategies are anonymous, then so is the associated introspective equilibrium. (For games with strategic complementarities, best responses are generically unique (Proposition A.2), so these are not significant issues here.)

We can now prove the following preliminary result:

Lemma A.1. *For each $k \geq 0$, the level-k strategies are anonymous and measurable.*

The proof is standard and therefore relegated to the online appendix. We are now ready to prove Proposition 2.1. We show that if the profile of level-k strategies converges, then its limit forms a correlated equilibrium. By Lemma A.1, the level-k strategies can be taken to be anonymous and

measurable for each $k \geq 0$. Hence, for each G and $\boldsymbol{u}$, we have a sequence of strategies $\{\sigma^k_{G,\boldsymbol{u}}\}_k$. Suppose that for each G and $\boldsymbol{u}$, the sequence $\{\sigma^k_{G,\boldsymbol{u}}\}_k$ converges to a strategy $\sigma_{G,\boldsymbol{u}} : S \to S$. If we write $\sigma = (\sigma_{G,\boldsymbol{u}})_{G\in\mathcal{G},\boldsymbol{u}\in\mathcal{U}}$ for the limiting strategy profile, then we need to show that for each $G \in \mathcal{G}$, $\boldsymbol{u} \in \mathcal{U}$, and $I \in S$,

$$\sigma_{G,\boldsymbol{u}}(I) \in \arg\max_{s^*\in S} U(s^*, (m_{s'}(\sigma))_{s'\in S}; I, G, \boldsymbol{u}). \tag{A.1}$$

To see this, note that the limiting strategy profile $\sigma : S \times \mathcal{U} \times \mathcal{G} \to S$ is jointly measurable (Aliprantis and Border, 2006, Lemma 4.29) so that for every s, I, G, the expected payoff $U(s, (m_{s'}(\sigma; I, G))_{s'\in S}; \boldsymbol{u})$ is well-defined. Moreover, by the dominated convergence theorem, $U(s, (m_{s'}(\sigma; I, G))_{s'\in S}; \boldsymbol{u})$ is the limit of the level-k expected payoff $U(s, (m_{s'}(\sigma^{k-1}))_{s'\in S}; I, G, \boldsymbol{u})$ as $k \to \infty$. The result then follows from a standard continuity argument. □

A.2 Proof of Propositions 3.1

Fix a monotone game with strategic complementarities with payoff functions $(u(\cdot;\rho_j))_{j\in N}$ and distribution $F(\rho_j)$ of risk parameters (where $F(\rho_j)$ may be degenerate, i.e., the game has identical preferences). Given a player $j \in N$ belongs to group G_j and has impulse I_j and risk parameter ρ_j, we refer to (I_j, G_j, ρ_j) as the player's *type*. For each type (I_j, G_j, ρ_j) and $k > 0$, assuming that the level-$(k-1)$ strategies $\sigma^{k-1} = (\sigma_j^{k-1})_{j\in N}$ are well-defined, define

$$\widehat{\Delta}^k(I_j, G_j, \rho_j) := \mathbb{E}_{\sigma^{k-1}}\Big[u(H, s_{-j}; \rho_j) - u(L, s_{-j}; \rho_j) \mid I_j, G_j\Big]$$

for the difference in expected payoff from choosing H versus L for a player from group G_j with impulse I_j and risk parameter ρ_j when other players play according to the level-$(k-1)$ strategy profile σ^{k-1}. So, H is a best response for type (I_j, G_j, ρ_j) if and only if $\widehat{\Delta}^k(I_j, G_j, \rho_j) \geq 0$, and it is the unique best response if the inequality is strict. Hence, by our separability assumption, H is a best response for type (I_j, G_j, ρ_j) if and only if $\mathbb{E}_{\sigma^{k-1}}[g(m)] \geq \rho_j$ (and it is the unique best response if the inequality is strict).

To prove Proposition 3.1, we show that the level-k strategies are switching strategies: For each level k, there exist cutoffs ρ^k_{IG} for $I \in \{H, L\}$ and $G \in \{A, B\}$ such that, at level k, a player from group G and with impulse I and risk parameter ρ_j chooses H if $\rho_j < \rho^k_{IG}$ and chooses L if $\rho_j > \rho^k_{IG}$ (if $\rho_j = \rho^k_{IG}$, then the player can choose either action). We then show that the level-k cutoff ρ^k_{IG} converge to a cutoff ρ_{IG} for each I, G as k goes to infinity, establishing Proposition 3.1. It turns out that the cutoffs can be ordered, which we note for future reference:

Proposition A.2. [Introspective Equilibrium] *For any linear or threshold game:*

(a) An introspective equilibrium exists and is essentially unique.

(b) There exist $\rho_{LA}, \rho_{LB}, \rho_{HB}, \rho_{HA}$ with $\rho_{LA} \leq \rho_{LB} \leq \rho_{LA} \leq \rho_{HA}$ such that, in introspective equilibrium,

(b1) If $\rho_j < \rho_{LA}$, then any player with risk parameter ρ_j chooses H;

(b2) If $\rho_j \in (\rho_{LA}, \rho_{LB})$, then a player with risk parameter ρ_j chooses H if he belongs to the minority group and chooses the action he expects to be culturally salient otherwise;

(b3) If $\rho_j \in (\rho_{LB}, \rho_{HB})$, then a player with risk parameter ρ_j chooses the action he expects to be culturally salient;

(b4) If $\rho_j \in (\rho_{HB}, \rho_{HA})$, then a player with risk parameter ρ_j chooses L if he belongs to the minority group and chooses the action he expects to be culturally salient otherwise;

(b5) If $\rho_j > \rho_{HA}$, then a player with risk parameter ρ_j chooses L.

Appendix A.2.1 proves Proposition A.2 for monotone games with strategic complementarities where players have identical preferences (including linear and threshold games). Appendix A.2.2 proves Proposition A.2 for linear games with heterogeneous preferences and $F(\rho_j)$ unimodal and symmetric, and Appendix A.2.3 proves the result for threshold games. This proves Proposition 3.1.

A.2.1 Identical preferences

We first prove Proposition A.2 for arbitrary monotone games with strategic complementarities with identical preferences (with common risk parameter ρ). That is, $g(m)$ can be any increasing function. We start with part (a). At level 0, all players follow their impulse. We claim that at level 1,

$$\widehat{\Delta}^1(H, A, \rho) \geq \widehat{\Delta}^1(H, B, \rho) \geq \widehat{\Delta}^1(L, B, \rho) \geq \widehat{\Delta}^1(L, A, \rho). \tag{A.2}$$

To see this, denote the posterior belief of a player with impulse I_j from group G_j over the proportion m^0 of players who choose H at level 0 by $\pi^0_{I_j G_j}$ and note that the posteriors can be ordered by first-order stochastic dominance

$$\pi^0_{HA} \succ_{FOSD} \pi^0_{HB} \succ_{FOSD} \pi^0_{LB} \succ_{FOSD} \pi^0_{LA}.$$

The claim now follows directly from the assumption that the game has strategic complementarities (Eq. (2.1)).

Now, by the monotonicity assumption (2.2), for each $I \in \{H, L\}$ and $G \in \{A, B\}$, there exists a unique ρ^1_{IG} such that $\widehat{\Delta}^1(I_j, G_j, \rho) \geq 0$ if and only if $\rho \leq \rho^1_{IG}$. By (A.2),

$$\rho^1_{HA} \geq \rho^1_{HB} \geq \rho^1_{LB} \geq \rho^1_{LA}.$$

At level $k > 1$, suppose that for each $I \in \{H, L\}$ and $G \in \{A, B\}$, there is a unique ρ^{k-1}_{IG} such that $\widehat{\Delta}^{k-1}(I, G, \rho) \geq 0$ if and only if $\rho \leq \rho^{k-1}_{IG}$, where the cutoffs ρ^{k-1}_{IG} satisfy

$$\rho^{k-1}_{HA} \geq \rho^{k-1}_{HB} \geq \rho^{k-1}_{LB} \geq \rho^{k-1}_{LA}.$$

Then, because first-order stochastic dominance is preserved under increasing transformations, the posterior beliefs π^{k-1}_{IG} of a player with impulse I from group G over the proportion m^{k-1} of players who choose H at level $k-1$ can again be ordered

$$\pi^{k-1}_{HA} \succ_{FOSD} \pi^{k-1}_{HB} \succ_{FOSD} \pi^{k-1}_{LB} \succ_{FOSD} \pi^{k-1}_{LA};$$

and it follows from strategic complementarities that

$$\widehat{\Delta}^k(H, A, \rho) \geq \widehat{\Delta}^k(H, B, \rho) \geq \widehat{\Delta}^k(L, B, \rho) \geq \widehat{\Delta}^k(L, A, \rho). \tag{A.3}$$

By a similar argument as before, for each $I \in \{H, L\}$ and $G \in \{A, B\}$, there is a unique ρ^k_{IG} such that $\widehat{\Delta}^k(I, G, \rho) \geq 0$ if and only if $\rho \leq \rho^k_{IG}$, where the cutoffs ρ^k_{IG} satisfy

$$\rho^k_{HA} \geq \rho^k_{HB} \geq \rho^k_{LB} \geq \rho^k_{LA}.$$

Before proving existence, we show that, if an introspective equilibrium exists, it is generically unique. Note that, at any level $k > 0$, the best response for a player with impulse I from group G is unique except when $\widehat{\Delta}^k(I, G, \rho) = 0$. The set of risk parameters for which this is the case (over all groups and impulses, and for any level $k \leq \infty$) is countable and thus has Lebesgue measure 0 in $\mathbb{R}$.

The proof that an introspective equilibrium exists follows from a standard monotonicity argument. We first consider generic games (i.e., games with unique best responses at every level). We need to show that the cutoffs converge, i.e., for each $I \in \{H, L\}$ and $G \in \{A, B\}$, there is ρ_{IG} such that $\lim_{k\to\infty} \rho^k_{IG} = \rho_{IG}$. Because there are finitely many types, the level-k strategies can be summarized by by a (finite-dimensional) vector that specifies the action that each type takes. For every $k \geq 0$, write σ^k_{IG} for the level-k action for type (I, G, ρ) and $\sigma^k = (\sigma^k_{HA}, \sigma^k_{HB}, \sigma^k_{LB}, \sigma^k_{LA})$ for the level-k strategy; note that σ^k_{IG} is unique for generic games (for given I, G).

Then, $\sigma^0 = (H, H, L, L)$, and by (A.2)–(A.3), the level-1 strategy profile σ^1 takes one of the following forms: (1) $\sigma^1 = (H, H, H, H)$; (2) $\sigma^1 = (H, H, H, L)$; (3) $\sigma^1 = (H, H, L, L)$; (4) $\sigma^1 = (H, L, L, L)$; (5) $\sigma^1 = (L, L, L, L)$. In case (1), the level-1 strategy profile is an introspective equilibrium: By strategic complementarities, if H is a best response for each type against the belief that players play according to $\sigma^0 = (H, H, L, L)$, then H is a best response for each type against the belief that players play according to $\sigma^1 = (H, H, H, H)$. By a similar argument, the level-1 strategy in case (5) is an introspective equilibrium. In case (3), we also have an introspective equilibrium as $\sigma^1 = \sigma^0$. It remains to consider cases (2) and (4). In case (2), by strategic complementarities, either $\sigma^2 = \sigma^1$ or $\sigma^2 = (H, H, H, H)$; in either case, $\sigma^3 = \sigma^2$, and we have an introspective equilibrium. Likewise, in case (4), by strategic complementarities, either $\sigma^2 = \sigma^1$ or $\sigma^2 = (L, L, L, L)$; in either case, $\sigma^3 = \sigma^2$, and we have an introspective equilibrium. Write $\sigma(\rho) = \{\sigma_{HA}, \sigma_{HB}, \sigma_{LB}, \sigma_{LA}\}$ for the introspective equilibrium of the game with risk parameter ρ (with fixed payoff function $u(\cdot, \rho)$. Thus, $\sigma(\rho) \in \{(H, H, H, H), (H, H, H, L), (H, H, L, L), (H, L, L, L), (L, L, L,)\}$. Then, by a simple inductive argument, for $\rho, \rho' \in \mathbb{R}$ with $\rho' > \rho$, if $\sigma(\rho') = H$, then $\sigma(\rho) = H$; and if $\sigma(\rho) = L$, then $\sigma(\rho') = L$. That is, the introspective equilibrium is monotonic in ρ. Thus, there exist cutoffs $\{\rho_{IG}\}_{I,G}$ with $\rho_{LA} \leq \rho_{LB} \leq \rho_{HB} \leq \rho_{HA}$ such that in introspective equilibrium, a player with impulse I from group G chooses H in introspective equilibrium if $\rho < \rho_{IG}$ and chooses L if $\rho > \rho_{IG}$, proving (b). (For generic ρ, one of these inequalities will be satisfied for every $I \in \{H, L\}$ and $G \in \{A, B\}$.)

The proof for nongeneric games is similar: given an arbitrary tie-breaking rule (which is independent of the level k) that specifies the action for a type if it is indifferent, the level-k strategies can be summarized by the action that each of the types takes. The rest of the proof is identical to that for the generic case and thus omitted. Notice that the introspective process converges at level 1 or 2 in this case. □

A.2.2 Linear games

We next prove Proposition A.2 for linear games. The proof for linear games with identical preference follows from the proof for games with identical preferences (Appendix A.2.1). So suppose that

there is preference heterogeneity, i.e., players' risk parameters are distributed according to a continuous distribution $F(\rho_j)$. In this case, there are infinitely many types (I, G, ρ_j) so the argument for the case of identical preferences does not apply.

We first prove (a). We show that at each level k, players follow a switching strategy: for all I, G, there is a cutoff ρ^k_{IG} such that type (I, G, ρ_j) chooses H at level k if $\rho_j < \rho^k_{IG}$ and L if $\rho_j > \rho^k_{IG}$. (If $\rho_j = \rho^k_{IG}$ then the type is indifferent and can choose either action; since the parameters ρ_j are continuously distributed, we do not need to specify its action.) We prove existence by showing that the cutoffs ρ^k_{IG} converge (for each I, G) as $k \to \infty$. To define the cutoffs, it will be convenient to introduce the notation $\tilde{x}$ for $1 - x$ for any variable x.

We are now ready to define the cutoffs. At level $k > 0$, H is a best response for type (I, G, ρ_j) if and only if its conditional expectation $\mathbb{E}^k[m \mid I, G]$ of the proportion of players choosing H at level $k-1$ is at least ρ_j. So, if we define $F(\infty) = \lim_{x\to\infty} F(x)$ and $F(-\infty) = \lim_{x\to-\infty} F(x)$ and set

$$\rho^0_{HA} := \infty; \qquad \rho^0_{LB} := -\infty; \tag{A.4}$$

$$\rho^0_{HB} := \infty; \qquad \rho^0_{LA} := -\infty; \tag{A.5}$$

it follows from a simple inductive argument that for any $k > 0$, players follow a switching strategy with cutoffs $\{\rho^k_{IG}\}_{I\in\{H,L\},G\in\{A,B\}}$ (i.e., type (I, G, ρ_j) chooses H if $\rho_j > \rho^k_{IG}$ and L if $\rho_j < \rho^k_{IG}$), where the cutoffs satisfy the following "law of motion":

$$\begin{aligned}
\rho^k_{HA} &= \tilde{\beta}Q_{in}F(\rho^{k-1}_{HA}) + \beta Q_{out}F(\rho^{k-1}_{HB}) + \beta\tilde{Q}_{out}F(\rho^{k-1}_{LB}) + \tilde{\beta}\tilde{Q}_{in}F(\rho^{k-1}_{LA});\\
\rho^k_{HB} &= \tilde{\beta}Q_{out}F(\rho^{k-1}_{HA}) + \beta Q_{in}F(\rho^{k-1}_{HB}) + \beta\tilde{Q}_{in}F(\rho^{k-1}_{LB}) + \tilde{\beta}\tilde{Q}_{out}F(\rho^{k-1}_{LA});\\
\rho^k_{LB} &= \tilde{\beta}\tilde{Q}_{out}F(\rho^{k-1}_{HA}) + \beta\tilde{Q}_{in}F(\rho^{k-1}_{HB}) + \beta Q_{in}F(\rho^{k-1}_{LB}) + \tilde{\beta}Q_{out}F(\rho^{k-1}_{LA});\\
\rho^k_{LA} &= \tilde{\beta}\tilde{Q}_{in}F(\rho^{k-1}_{HA}) + \beta\tilde{Q}_{out}F(\rho^{k-1}_{HB}) + \beta Q_{out}F(\rho^{k-1}_{LB}) + \tilde{\beta}Q_{in}F(\rho^{k-1}_{LA}).
\end{aligned}$$

Moreover, by a simple inductive argument, for every $k > 0$,

$$0 \le \rho^k_{LA} \le \rho^k_{LB} \le \rho^k_{HB} \le \rho^k_{HA} \le 1;$$

and

$$\rho^k_{HA} + \rho^k_{LA} = \rho^k_{HB} + \rho^k_{LB} = \tilde{\beta}\big[F(\rho^{k-1}_{HA}) + F(\rho^{k-1}_{LA})\big] + \beta\big[F(\rho^{k-1}_{HB}) + F(\rho^{k-1}_{LB})\big]. \tag{A.6}$$

That is, at level 0, all players choose the action they expect to be culturally salient ($F(\rho^0_{HG}) = 1$ and $F(\rho^0_{LG}) = 0$ for $G \in \{A, B\}$) and for $k > 0$, the cutoffs are a function of the proportion $F(\rho^{k-1}_{IG})$ of players choosing H at level $k-1$ given that they have impulse I and belong to group G, for $I \in \{H, L\}$ and $G \in \{A, B\}$.

To prove existence, we need to show that the cutoffs $\{\rho^k_{IG}\}_{k=0,1,2,\ldots}$ converge for every I and G as k goes to infinity. The standard approach for games with strategic complementarities is to show that the sequence of cutoffs is monotone. But while this can be done if the initial cutoffs can be chosen appropriately (e.g., Vives, 1990, Thm. 5.1), this approach does not work here because the initial values are fixed by the introspective process (viz., $F(\rho^0_{HG}) = 1$ and $F(\rho^0_{LG}) = 0$) and the resulting sequence need not be monotone. That is, the cutoffs $\rho^k_{LA}, \rho^k_{LB}, \rho^k_{HB}, \rho^k_{HA}$ can fluctuate with k. To overcome this difficulty, we employ a change of variables: we identify variables that pin down the cutoffs and whose evolution *is* monotone in k (see Lemma A.4 below).

We prove existence for the case that the mean μ of $F(\rho_j)$ is at most $\frac{1}{2}$; the proof for the case $\mu \ge \frac{1}{2}$ is analogous and can be found in the online appendix. We prove the result under slightly weaker

assumptions than in Proposition A.2 to highlight the key conditions driving the result: rather than assuming that $f(\rho_j)$ is unimodal and symmetric, we require that the density $f(\rho_j)$ satisfies

$$f(\tfrac{1}{2}+x) \geq f(\tfrac{1}{2}+y) \qquad \forall x, y \text{ s.t. } y \geq x \geq 0; \tag{A.7}$$

$$f(\tfrac{1}{2}-x) \geq f(\tfrac{1}{2}+x) \qquad \forall x \geq 0. \tag{A.8}$$

These conditions are clearly satisfied if $f(\rho_j)$ is unimodal and symmetric (with mean $\mu \leq \frac{1}{2}$): in that case, $f(\mu+x) = f(\mu-x)$ and $f(\rho_j)$ is decreasing on $[\mu, \infty)$ (decreasing on $(-\infty, \mu]$). However, other distributions also satisfy these conditions.[45]

It will be convenient to define, for each level $k > 0$,

$$\begin{aligned} \bar{\rho}^k &:= \tfrac{1}{2}\big[\rho^k_{HA} + \rho^k_{LA}\big] \\ &= \tfrac{1}{2}\tilde{\beta}\big[F(\rho^{k-1}_{HA}) + F(\rho^{k-1}_{LA})\big] + \tfrac{1}{2}\beta\big[F(\rho^{k-1}_{HB}) + F(\rho^{k-1}_{LB})\big]; \end{aligned}$$

where the last line uses (A.6). If we define $\bar{\rho}^0 := \frac{1}{2}$, then

$$\begin{aligned} \bar{\rho}^1 &= \tfrac{1}{2}\tilde{\beta} + \tfrac{1}{2}\beta = \tfrac{1}{2} \geq \bar{\rho}^0; \\ \rho^1_{LA} &= \tilde{\beta}\tilde{Q}_{in} + \beta\tilde{Q}_{out} \geq \rho^0_{LA}; \\ \rho^1_{LB} &= \tilde{\beta}\tilde{Q}_{out} + \beta\tilde{Q}_{in} \geq \rho^0_{LB}; \end{aligned}$$

where we have used (A.4)–(A.5). Also, we will use that ρ^k_{LA} and ρ^k_{LB} can be written as

$$\begin{aligned} \rho^k_{LA} &= \tilde{\beta}(Q_{in} - \tilde{Q}_{in})F(\rho^{k-1}_{LA}) + \tilde{\beta}\tilde{Q}_{in}\big[F(\rho^{k-1}_{HA}) + F(\rho^{k-1}_{LA})\big] + \\ &\quad \beta(Q_{out} - \tilde{Q}_{out})F(\rho^{k-1}_{LB}) + \beta\tilde{Q}_{out}\big[F(\rho^{k-1}_{HB}) + F(\rho^{k-1}_{LB})\big]; \\ \rho^k_{LB} &= \tilde{\beta}(Q_{out} - \tilde{Q}_{out})F(\rho^{k-1}_{LA}) + \tilde{\beta}\tilde{Q}_{out}\big[F(\rho^{k-1}_{HA}) + F(\rho^{k-1}_{LA})\big] + \\ &\quad \beta(Q_{in} - \tilde{Q}_{in})F(\rho^{k-1}_{LB}) + \beta\tilde{Q}_{in}\big[F(\rho^{k-1}_{HB}) + F(\rho^{k-1}_{LB})\big]. \end{aligned}$$

Because $Q_{in} > \tilde{Q}_{in} > 0$ and $Q_{out} > \tilde{Q}_{out} > 0$ (and using that $F(\rho_j)$ is increasing), ρ^k_{LA} and ρ^k_{LB} are increasing in ρ^{k-1}_{LA}, ρ^{k-1}_{LB}, $F(\rho^{k-1}_{HA}) + F(\rho^{k-1}_{LA})$, and $F(\rho^{k-1}_{HB}) + F(\rho^{k-1}_{LB})$. Moreover, because $\rho^k_{HA} = 2\bar{\rho}^k - \rho^k_{LA}$ and $\rho^k_{HB} = 2\bar{\rho}^k - \rho^k_{LB}$, the system is effectively three-dimensional: to prove that the cutoffs $\{\rho^k_{IG}\}_{k=0,1,\ldots}$ converge for every $I \in \{H, L\}$ and $G \in \{A, B\}$, it suffices to show that $\{\rho^k_{LA}\}_k$, $\{\rho^k_{LB}\}_k$, and $\{\bar{\rho}^k\}_k$ converge (as $k \to \infty$). This follows from the following two lemmas:

Lemma A.3. *Suppose $f(\rho_j)$ has mean $\mu \leq \frac{1}{2}$ and satisfies (A.7)–(A.8). For any group $G \in \{A, B\}$ and $k > 0$, if $\bar{\rho}^k \geq \bar{\rho}^{k-1} \geq \frac{1}{2}$ and $\rho^k_{LG} \geq \rho^{k-1}_{LG}$, then $F(\rho^k_{HG}) + F(\rho^k_{LG}) \geq F(\rho^{k-1}_{HG}) + F(\rho^{k-1}_{LG})$.*

Proof. For concreteness, take $G = A$. If $\rho^k_{HA} \geq \rho^{k-1}_{HA}$, then the result follows immediately from the fact that $F(\rho_j)$ is increasing. So suppose that $\rho^k_{HA} < \rho^{k-1}_{HA}$.

We first prove the result for $k > 1$. Define $\Delta^k := \rho^{k-1}_{HA} - \rho^k_{HA}$. Then,

$$\begin{aligned} 0 < \Delta^k &= (\rho^{k-1}_{HA} - \bar{\rho}^{k-1}) - (\rho^k_{HA} - \bar{\rho}^k) - (\bar{\rho}^k - \bar{\rho}^{k-1}) \\ &= (\bar{\rho}^{k-1} - \rho^{k-1}_{LA}) - (\bar{\rho}^k - \rho^k_{LA}) - (\bar{\rho}^k - \bar{\rho}^{k-1}) \\ &\leq \rho^k_{LA} - \rho^{k-1}_{LA}, \end{aligned}$$

[45] For example, any beta distribution whose parameters a, b satisfy either (i) $b \geq a > 0$ and $a + b \geq 2$; or (ii) $b \in [1, 2]$ and $a \in (0, 2 - b)$ also satisfies (A.7)–(A.8).

where the last line uses that $\bar{\rho}^k \geq \bar{\rho}^{k-1}$. This inequality says that, for a given group, the decrease in the cutoff for types with $I = H$ (when going from level $k-1$ to k) is smaller than the increase in the cutoff for types with $I = L$. We have

$$\begin{aligned}
&\left[F(\rho_{HA}^k) + F(\rho_{LA}^k)\right] - \left[F(\rho_{HA}^{k-1}) + F(\rho_{LA}^{k-1})\right] \\
&\geq \int_0^{\Delta^k} f(u + \rho_{LA}^{k-1})du + \int_0^{-\Delta^k} f(u + \rho_{HA}^{k-1})du \\
&= \int_0^{\Delta^k} \Big[f(u + \rho_{LA}^{k-1}) - f(-u + \rho_{HA}^{k-1})\Big]du \\
&= \int_0^{\Delta^k} \Big[f(\bar{\rho}^{k-1} - (\bar{\rho}^{k-1} - \rho_{LA}^{k-1} - u)) - f(\bar{\rho}^{k-1} + (\bar{\rho}^{k-1} - \rho_{LA}^{k-1} - u))\Big]du.
\end{aligned}$$

For all $u \in [0, \Delta^k]$, $\bar{\rho}^{k-1} - \rho_{LA}^{k-1} - u \geq \bar{\rho}^{k-1} - \rho_{LA}^{k-1} - \Delta^k = \rho_{HA}^k - \bar{\rho}^{k-1} \geq \rho_{HA}^k - \bar{\rho}^k \geq 0$. So, if we show that for $\bar{\rho} \geq \frac{1}{2}$ and $x \geq 0$, we have $f(\bar{\rho} - x) \geq f(\bar{\rho} + x)$, then the result follows. But this holds: First, if $x \leq \bar{\rho} - \frac{1}{2}$,

$$\begin{aligned}
f(\bar{\rho} + x) &= f(\tfrac{1}{2} + (\bar{\rho} - \tfrac{1}{2}) + x) \\
&\leq f(\tfrac{1}{2} + (\bar{\rho} - \tfrac{1}{2}) - x) \\
&= f(\bar{\rho} - x);
\end{aligned}$$

where the inequality uses (A.7). Second, if $x > \bar{\rho} - \frac{1}{2}$,

$$\begin{aligned}
f(\bar{\rho} + x) &= f(\tfrac{1}{2} + (\bar{\rho} - \tfrac{1}{2}) + x) \\
&\leq f(\tfrac{1}{2} - (\bar{\rho} - \tfrac{1}{2}) + x) \\
&\leq f(\tfrac{1}{2} + (\bar{\rho} - \tfrac{1}{2}) - x) \\
&= f(\bar{\rho} - x);
\end{aligned}$$

where the inequalities use (A.7) and (A.8), respectively.

Next, suppose $k = 1$. The result follows if we show that $F(\rho_{HA}^1) + F(\rho_{LA}^1) \geq F(\rho_{HA}^0) + F(\rho_{LA}^0)$. By (A.4)–(A.5) and using that $\bar{\rho}^1 \geq \frac{1}{2}$, it suffices to show that for $\bar{\rho} \geq \frac{1}{2}$ and $x \geq 0$, $F(\bar{\rho}+x) + F(\bar{\rho}-x) \geq 1$. But this holds:

$$\begin{aligned}
F(\bar{\rho} + x) + F(\bar{\rho} - x) &= 1 + F(\bar{\rho} - x) - (1 - F(\bar{\rho} + x)) \\
&= 1 + \int_0^{\infty} \Big[f(\bar{\rho} - (x + u)) - f(\bar{\rho} + (x + u))\Big]du \\
&\geq 1;
\end{aligned}$$

where the last line uses that $f(\bar{\rho} - y) \geq f(\bar{\rho} + y)$ for $y \geq 0$. □

Lemma A.4. *Suppose $f(\rho_j)$ has mean $\mu \leq \frac{1}{2}$ and satisfies (A.7)–(A.8). Then, for all $k > 0$, $\bar{\rho}^k \geq \bar{\rho}^{k-1} \geq \frac{1}{2}$, $\rho_{LA}^k \geq \rho_{LA}^{k-1}$ and $\rho_{LB}^k \geq \rho_{LB}^{k-1}$.*

Proof. The proof is by induction. As noted earlier, the result holds for $k = 1$. For $k > 1$, suppose that $\bar{\rho}^{k-1} \geq \bar{\rho}^{k-2} \geq \frac{1}{2}$, $\rho_{LA}^{k-1} \geq \rho_{LA}^{k-2}$ and $\rho_{LB}^{k-1} \geq \rho_{LB}^{k-1}$. Then, by Lemma A.3,

$$\begin{aligned}
F(\rho_{HA}^{k-1}) + F(\rho_{LA}^{k-1}) &\geq F(\rho_{HA}^{k-2}) + F(\rho_{LA}^{k-2}); \\
F(\rho_{HB}^{k-1}) + F(\rho_{LB}^{k-1}) &\geq F(\rho_{HB}^{k-2}) + F(\rho_{LB}^{k-2}).
\end{aligned}$$

Moreover, because $F(\rho_j)$ is increasing, $F(\rho_{LA}^{k-1}) \geq F(\rho_{LA}^{k-2})$ and $F(\rho_{LB}^{k-1}) \geq F(\rho_{LB}^{k-2})$. It then follows directly from the expressions for $\bar{\rho}^k$, ρ_{LA}^k, and $\bar{\rho}_{LB}^k$ that $\bar{\rho}^k \geq \bar{\rho}^{k-1}$, $\rho_{LA}^k \geq \rho_{LA}^{k-1}$, and $\bar{\rho}_{LB}^k \geq \bar{\rho}_{LB}^{k-1}$. □

It now follows immediately that the sequences $\{\rho^k_{LA}\}_k$, $\{\rho^k_{LB}\}_k$, and $\{\bar{\rho}^k\}_k$ converge: For each $k > 1$, $\bar{\rho}^k \leq 1$, $\rho^k_{LA} \leq 1$, and $\rho^k_{LB} \leq 1$. So, by Lemma A.4, each sequence is bounded and monotone. Hence, there exist $\bar{\rho} \in (\frac{1}{2}, 1)$, $\rho_{LA} \in (0, \bar{\rho}]$, and $\rho_{LB} \in (0, \bar{\rho}]$ such that $\bar{\rho}^k \uparrow \bar{\rho}$, $\rho^k_{LA} \uparrow \rho_{LA}$, and $\rho^k_{LB} \uparrow \rho_{LB}$. Because passing to the limit preserves inequalities, we have $\rho_{HA} \geq \rho_{HB} \geq \rho_{LB} \geq \rho_{LA}$, proving (b).

To see that introspective equilibrium is essentially unique, note that, because the risk parameters ρ_j are drawn from a continuous distribution $F(\rho_j)$ the set of types that are indifferent at any given level k is countable. This set has measure 0 (under $F(\rho_j)$). Hence, introspective equilibrium uniquely determines behavior except for a set of types of measure 0. □

A.2.3 Threshold games

Finally, we prove Proposition A.2 for threshold games. Again, the proof for threshold games with identical preference follows from the proof for games with identical preferences (Appendix A.2.1). So suppose that there is preference heterogeneity, i.e., players' risk parameters are distributed according to a continuous distribution $F(\rho_j)$.

We start with part (a). We first prove existence. We write $\mathbb{I}_E$ for the indicator function for the event E. That is, $\mathbb{I}_E = 1$ if E obtains, and $\mathbb{I}_E = 0$ otherwise. Also, we define $F(\infty) := \lim_{x\to\infty} F(x)$ and $F(-\infty) := \lim_{x\to-\infty} F(x)$, and we write $z := \frac{1}{2}(1+\eta)$ (and thus $\tilde{z} := 1 - z = \frac{1}{2}(1-\eta)$).

Again, the level-0 strategy is a switching strategy: At level 0, player of type (I_j, G_j, ρ_j) chooses H if $\rho_j < \rho^0_{I_jG_j}$ and chooses L if $\rho_j > \rho^0_{I_jG_j}$ (a type with $\rho_j = \rho^0_{I_jG_j}$ can choose either action). Then, we can summarize the level-0 by $(\rho^0_{HA}, \rho^0_{HB}, \rho^0_{LB}, \rho^0_{LA}) := (\infty, \infty, -\infty, -\infty)$. For $k > 0$, suppose that there exist cutoffs $(\rho^{k-1}_{HA}, \rho^{k-1}_{HB}, \rho^{k-1}_{LB}, \rho^{k-1}_{LA})$ with $\rho^0_{HA} \geq \rho^0_{HB} \geq \rho^0_{LB} \geq \rho^0_{LA}$, such that, at level $k-1$, every type (I, G, ρ_j) chooses H if $\rho_j < \rho^0_{IG}$ and chooses L if $\rho_j > \rho^0_{IG}$. (As before, if $\rho_j = \rho^0_{IG}$, the type may choose either action.) Then, at level k, H is the unique best response for type (I, G, ρ_j) if and only if

$$\rho_j < \mathbb{P}^k(m \geq T \mid I, G) =: \rho^k_{IG},$$

where $\mathbb{P}^k(m \geq T \mid I, G)$ is the conditional probability that a player from group G with impulse I assigns to the event that the proportion m of players who choose H is at least as high as the threshold (given the level-$(k-1)$ strategies given by the cutoffs $(\rho^{k-1}_{HA}, \rho^{k-1}_{HB}, \rho^{k-1}_{LB}, \rho^{k-1}_{LA})$). Likewise, L is the unique best response for type (I, G, ρ_j) if and only if $\rho_j > \rho^k_{IG}$. Consequently, at level k, players follow a switching strategy with cutoffs $(\rho^k_{HA}, \rho^k_{HB}, \rho^k_{LB}, \rho^k_{LA})$, where $\rho^k_{IG} = \mathbb{P}^k(m \geq T \mid I, G)$.

We can derive explicit expressions for the cutoffs. It will be helpful to introduce the notation $\boldsymbol{\pi}_{IG} = (\pi^{HH}_{IG}, \pi^{HL}_{IG}, \pi^{LH}_{IG}, \pi^{LL}_{IG})$ for the conditional beliefs of players over states, with $\pi^{\theta\theta'}_{IG}$ is the conditional probability that a player from group G with impulse I assigns to state $(\theta_A, \theta_B) = (\theta, \theta')$, i.e.,

$$\boldsymbol{\pi}_{HA} := (qz, q\tilde{z}, \tilde{q}\tilde{z}, \tilde{q}z); \qquad \boldsymbol{\pi}_{LB} := (\tilde{q}z, q\tilde{z}, \tilde{q}\tilde{z}, qz);$$
$$\boldsymbol{\pi}_{HB} := (qz, \tilde{q}\tilde{z}, q\tilde{z}, \tilde{q}z); \qquad \boldsymbol{\pi}_{LA} := (\tilde{q}z, \tilde{q}\tilde{z}, q\tilde{z}, qz).$$

Then, if we denote by $m^{k-1}_{\theta\theta'}$ the proportion of players who choose H if $(\theta_A, \theta_B) = (\theta, \theta')$, we have

$$\begin{aligned}
\rho^k_{HA} = \;& qz\mathbb{I}_{[m^{k-1}_{HH} \geq T]} + q\tilde{z}\mathbb{I}_{[m^{k-1}_{HL} \geq T]} + \tilde{q}\tilde{z}\mathbb{I}_{[m^{k-1}_{LH} \geq T]} + \tilde{q}z\mathbb{I}_{[m^{k-1}_{LL} \geq T]};\\
\rho^k_{HB} = \;& qz\mathbb{I}_{[m^{k-1}_{HH} \geq T]} + \tilde{q}\tilde{z}\mathbb{I}_{[m^{k-1}_{HL} \geq T]} + q\tilde{z}\mathbb{I}_{[m^{k-1}_{LH} \geq T]} + \tilde{q}z\mathbb{I}_{[m^{k-1}_{LL} \geq T]};\\
\rho^k_{LB} = \;& \tilde{q}z\mathbb{I}_{[m^{k-1}_{HH} \geq T]} + q\tilde{z}\mathbb{I}_{[m^{k-1}_{HL} \geq T]} + \tilde{q}\tilde{z}\mathbb{I}_{[m^{k-1}_{LH} \geq T]} + qz\mathbb{I}_{[m^{k-1}_{LL} \geq T]};\\
\rho^k_{LA} = \;& \tilde{q}z\mathbb{I}_{[m^{k-1}_{HH} \geq T]} + \tilde{q}\tilde{z}\mathbb{I}_{[m^{k-1}_{HL} \geq T]} + q\tilde{z}\mathbb{I}_{[m^{k-1}_{LH} \geq T]} + qz\mathbb{I}_{[m^{k-1}_{LL} \geq T]};
\end{aligned}$$

where

$$\begin{aligned}
m_{HH}^{k-1} &:= \tilde{\beta}qF(\rho_{HA}^{k-1}) + \beta qF(\rho_{HB}^{k-1}) + \beta\tilde{q}F(\rho_{LB}^{k-1}) + \tilde{\beta}\tilde{q}F(\rho_{LA}^{k-1});\\
m_{HL}^{k-1} &:= \tilde{\beta}qF(\rho_{HA}^{k-1}) + \beta\tilde{q}F(\rho_{HB}^{k-1}) + \beta qF(\rho_{LB}^{k-1}) + \tilde{\beta}\tilde{q}F(\rho_{LA}^{k-1});\\
m_{LH}^{k-1} &:= \tilde{\beta}\tilde{q}F(\rho_{HA}^{k-1}) + \beta qF(\rho_{HB}^{k-1}) + \beta\tilde{q}F(\rho_{LB}^{k-1}) + \tilde{\beta}qF(\rho_{LA}^{k-1});\\
m_{LL}^{k-1} &:= \tilde{\beta}\tilde{q}F(\rho_{HA}^{k-1}) + \beta\tilde{q}F(\rho_{HB}^{k-1}) + \beta qF(\rho_{LB}^{k-1}) + \tilde{\beta}qF(\rho_{LA}^{k-1}).
\end{aligned}$$

By the induction hypothesis, $m_{HH}^{k-1} \geq m_{HL}^{k-1} \geq m_{LH}^{k-1} \geq m_{LL}^{k-1}$, and (using that $\tilde{\beta} \geq \beta$ and $z > \tilde{z}$), $\rho_{HA}^k \geq \rho_{HB}^k \geq \rho_{LB}^k \geq \rho_{LA}^k$.

Establishing existence – i.e., showing that the cutoffs $\{(\rho_{HA}^k, \rho_{HB}^k, \rho_{LB}^k, \rho_{LA}^k)\}_k$ converge to some vector $(\rho_{HA}, \rho_{HB}, \rho_{LB}, \rho_{LA})$ as $k \to \infty$ – is facilitated by the fact that the indicator function can take on only two values, 0 and 1: At each level $k \geq 1$, $(\rho_{HA}^k, \rho_{HB}^k, \rho_{LB}^k, \rho_{LA}^k) \in \mathcal{R} := \{R_1, R_2, R_3, R_4, R_5\}$, where the five vectors

$$\begin{aligned}
R_1 &:= (0,0,0,0);\\
R_2 &:= (qz, qz, \tilde{q}z, \tilde{q}z);\\
R_3 &:= (q, qz + \tilde{q}\tilde{z}, \tilde{q}z + q\tilde{z}, \tilde{q});\\
R_4 &:= (1-\tilde{q}z, 1-\tilde{q}z, 1-qz, 1-qz);\\
R_5 &:= (1,1,1,1);
\end{aligned}$$

in $\mathcal{R}$ correspond to the configurations for $(\mathbb{I}_{[m_{HH}\geq T]}, \mathbb{I}_{[m_{HL}\geq T]}, \mathbb{I}_{[m_{LH}\geq T]}, \mathbb{I}_{[m_{LL}\geq T]}) \in \{0,1\}^4$ that are consistent with $m_{HH} \geq m_{HL} \geq m_{LH} \geq m_{LL}$. Importantly, the vectors in $\mathcal{R}$ can be ordered: $R_1 < R_2 < R_3 < R_4 < R_5$. It is now immediate that the process converges: write $\boldsymbol{\rho}^k := (\rho_{HA}^k, \rho_{HB}^k, \rho_{LB}^k, \rho_{LA}^k)$. Then, for $k \geq 1$, either (i) $\boldsymbol{\rho}^{k+1} \geq \boldsymbol{\rho}^k$; or (ii) $\boldsymbol{\rho}^{k+1} \leq \boldsymbol{\rho}^k$. If $\boldsymbol{\rho}^{k+1} \geq \boldsymbol{\rho}^k$, then, by strategic complementarities, $\boldsymbol{\rho}^{\ell+1} \geq \boldsymbol{\rho}^\ell$ for all $\ell \geq 1$; and if $\boldsymbol{\rho}^{k+1} \leq \boldsymbol{\rho}^k$, then $\boldsymbol{\rho}^{\ell+1} \leq \boldsymbol{\rho}^\ell$ for all $\ell \geq 1$. Moreover, for $k \geq 1$, we have $\boldsymbol{\rho}^k \in [0,1]^4$. So, we have a monotone sequence in a bounded space, which must converge. This proves existence. The proof that the introspective equilibrium is essentially unique again follows from the fact that the best response for a type with risk parameter ρ_j is unique for a set of risk parameters with measure 1 (under $F(\rho_j)$). The proof of (b) again follows by noting that $\rho_{HA} \geq \rho_{HB} \geq \rho_{LB} \geq \rho_{LA}$. □

A.3 Proof of Proposition 3.2

We first characterize the introspective equilibrium for linear games with identical preferences and derive the comparative statics for that case. We then show how the results extend to games with limited preference heterogeneity.

A.3.1 Identical preferences

We characterize the introspective equilibria of linear games with identical preferences. To state the result, define the cutoffs $\{\rho_{IG}^*\}_{I,G}$ by

$$\rho_{HA}^* := \max\Big\{(1-\beta)Q_{in}, (1-\beta)\,Q_{out} + \beta\,Q_{in}\Big\}; \qquad \rho_{HB}^* := (1-\beta)\,Q_{out} + \beta\,Q_{in};$$
$$\rho_{LA}^* := 1 - \rho_{HA}^*; \qquad \rho_{LB}^* := 1 - \rho_{HB}^*;$$

see Figure 4 for an illustration. Note that the cutoffs are symmetric around $\frac{1}{2}$. The following result uses these cutoffs characterize the introspective equilibrium for linear games with identical preferences.

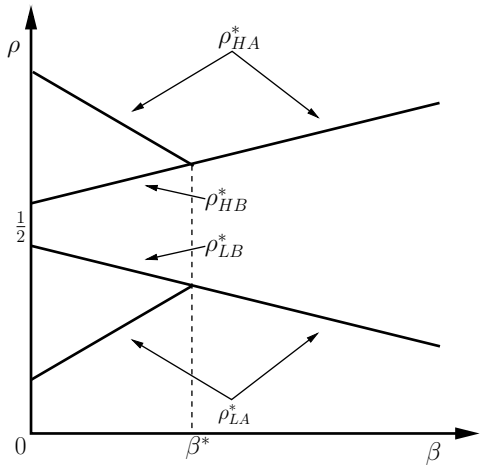

Figure 4: The cutoffs $\{\rho^*_{IG}\}_{I,G}$ as a function of diversity β.

Lemma A.5. [Linear Game with Identical Preferences: Equilibrium Characterization] *For any linear game with identical preferences ($\rho_j = \rho$ for all $j \in N$), across all tie-breaking rules,*

(a) If $\rho \leq \rho^*_{LA}$, then there is an introspective equilibrium in which all players choose H.

(b) if $\rho \in [\rho^*_{LA}, \rho^*_{LB}]$, there is an introspective equilibrium in which majority players follow their impulse while minority players choose H.

(c) if $\rho \in [\rho^*_{LB}, \rho^*_{HB}]$, all players follow their impulse.

(d) if $\rho \in [\rho^*_{HB}, \rho^*_{HA}]$, then there is an introspective equilibrium in which majority players follow their impulse and minority players choose L.

(e) if $\rho \geq \rho^*_{HA}$, then there is an introspective equilibrium in which all players choose L.

There are no other introspective equilibria.

Before presenting the proof, we note that Lemma A.5 characterizes the introspective equilibria across all tie-breaking rules, i.e., it identifies the set of introspective equilibria that can occur for some tie-breaking rule. In particular, Lemma A.5 implies that, except in the nongeneric case $\rho = \rho^*_{IG}$ for some I and G, there is a unique introspective equilibrium.

Proof of Lemma A.5. At level 0, all players choose the action they expect to be culturally salient. It is easy to check that the level-1 conditional expectations are

$$\begin{aligned}
\mathbb{E}^1[m \mid I_j = H, G_j = A] &= \tilde{\beta} Q_{in} + \beta Q_{out} =: \rho^1_{HA};\\
\mathbb{E}^1[m \mid I_j = H, G_j = B] &= \tilde{\beta} Q_{out} + \beta Q_{in} =: \rho^1_{HB};\\
\mathbb{E}^1[m \mid I_j = L, G_j = B] &= \tilde{\beta} \tilde{Q}_{out} + \beta \tilde{Q}_{in} =: \rho^1_{LB};\\
\mathbb{E}^1[m \mid I_j = L, G_j = A] &= \tilde{\beta} \tilde{Q}_{in} + \beta \tilde{Q}_{out} =: \rho^1_{LA}.
\end{aligned}$$

As in the proof of Proposition 3.1 (Appendix A.2.1), it suffices to consider 5 cases, depending on how the risk parameter ρ compares to r^1_{IG} for each impulse I and group G. We start with discussing the generic case ($\rho \neq \rho^1_{IG}$ for $I = H, L$ and $G = A, B$). First suppose that $\rho^1_{IG} > \rho$ for all $G \in \{A, B\}$ and $I \in \{H, L\}$. Then, at level 1, the unique best response for any player is to choose H, regardless of his

impulse. By strategic complementarities (Eq. (2.1)), it follows that at all levels $k \geq 2$, the unique best response for players is to choose H (since for all $k > 1$, $\mathbb{E}^k[m \mid I, G] = 1 > \rho^1_{IG}$). Consequently, if the risk parameter is sufficiently small, all players choose H in introspective equilibrium. Second, suppose that $\rho^1_{IG} < \rho$ for all G and I. Then, by a similar argument, for any $k > 0$, all players choose L at level k. Consequently, if the risk parameter is sufficiently large, all players choose L in introspective equilibrium.

Third, if $\rho \in (\rho^1_{LB}, \rho^1_{HB})$, then the unique best response for each player at level 1 is to choose the action he expects to be culturally salient (i.e., $\sigma^1_j(I_j) = I_j$)). That is, the level-1 strategy coincides with the level-0 strategy, so that $\rho^1_{IG} = \rho^0_{IG}$ for all I, G. A simple inductive argument shows that, at any level $k > 0$, players choose the action they expect to be culturally salient. Consequently, in introspective equilibrium, all players choose the action they expect to be culturally salient.

Fourth, if $\rho \in (\rho^1_{HB}, \rho^1_{HA})$, then, at level 1, the unique best response for majority players is to choose the action they expect to be culturally salient while for minority players the unique best response is to choose L regardless of their impulse. By strategic complementarities, the players who chose L at level 1 also choose L at level 2. It thus remains to consider the incentives of majority players with an impulse to choose H. The unique best response for a majority player with impulse $I = H$ at level 2 is to choose H whenever $\mathbb{E}^2[m \mid I = H, G = A] > \rho$, where $\mathbb{E}^2[m \mid I = H, G = A] = \tilde{\beta} Q_{in} := \rho^2_{HA}$; by a simple inductive argument, at any level $k \geq 2$, majority players with impulse $I = H$ choose H and the other players choose L; hence, in introspective equilibrium, majority players choose the action they expect to be culturally salient while minority players choose L. If $\rho^2_{HA} < \rho$, on the other hand, then the unique best response for a majority player with impulse $I = H$ at level 2 is to choose L. By a simple inductive argument, all players choose L at any level $k \geq 2$; hence, in introspective equilibrium, all players choose L.

Fifth, if $\rho \in (\rho^1_{LA}, \rho^1_{LB})$, then, at level 1, the unique best response for majority players is to choose the action they expect to be culturally salient while for minority players the unique best response is to choose L regardless of their impulse. By strategic complementarities, the players who chose H at level 1 also choose H at level 2. It thus remains to consider the incentives of majority players with an impulse to choose L. The unique best response for a majority player with impulse $I = L$ at level 2 is to choose L whenever $\mathbb{E}^2[m \mid I = L, G = A] < \rho$, where $\mathbb{E}^2[m \mid I = L, G = A] = 1 - \tilde{\beta} Q_{in} := \rho^2_{LA}$; by a simple inductive argument, at any level $k \geq 2$, majority players with impulse $I = L$ choose L and the other players choose H; hence, in introspective equilibrium, majority players choose the action they expect to be culturally salient while minority players choose H. If $\rho^2_{LA} > \rho$, on the other hand, then the unique best response for a majority player with impulse $I = H$ at level 2 is to choose H. By a simple inductive argument, all players choose H at any level $k \geq 2$; hence, in introspective equilibrium, all players choose H.

In the nongeneric case ($\rho = \rho^1_{IG}$ or $\rho = \rho^2_{IG}$ for some $I = H, L$ and $G = A, B$), some types (I, G) are indifferent between H and L. In that case, the risk parameter sits at the boundary between two regimes (e.g., $\rho = \rho^1_{LB}$ is at the boundary between $\rho \leq \rho^1_{LB}$ and $\rho \geq \rho^1_{LB}$), and the introspective equilibria are the introspective equilibria for (the interior of) the two regimes (in the example, $\rho < \rho^1_{LB}$ and $\rho > \rho^1_{LB}$).

Hence, if we set

$$\rho^*_{HA} := \max\{\rho^2_{HA}, \rho_{HB}\}; \qquad \rho^*_{LB} := \rho^1_{LB};$$
$$\rho^*_{HB} := \rho^1_{HB}; \qquad \rho^*_{LA} := \min\{\rho^2_{LA}, \rho^1_{LB};$$

the proof is complete. □

We are now ready to prove Proposition 3.2 for games with identical preferences. By Lemma A.5, players' decisions are driven by payoff considerations whenever $\rho < \rho^*_{LA}$ or $\rho > \rho^*_{HB}$. Comparing Figures 1 and 4 (and using the expressions for ρ^*_{LA} and ρ^*_{HA}), we see that the shaded areas in Figure 1 correspond precisely to the areas for which $\rho < \rho^*_{LA}$ or $\rho > \rho^*_{HB}$. To prove Proposition 3.2(a), note that for $\beta = 0$, we need to consider only ρ^*_{LA} and ρ^*_{HA} (there is no minority), while for $\beta = \frac{1}{2}$, the cutoffs for group A and B coincide (i.e., $\rho^*_{IG} = \rho^*_{IA}$); hence, it suffices to consider the cutoffs ρ^*_{HA} and ρ^*_{LA} in both cases. If we write $\rho^*_{IA}(\beta)$ for the cutoff ρ^*_{IA} when diversity is β, we have $\rho_{LA}(0) > \rho_{LA}(\frac{1}{2}) > \rho_{HA}(\frac{1}{2}) > \rho_{HA}(0)$. This proves part (a). To prove part (b), note that the critical mass β^* is the level of diversity that solves $\rho^1_{LB} = \rho^2_{LA}$, or, equivalently (by symmetry), $\rho^1_{HB} = \rho^2_{HA}$. This yields

$$\beta^* = \frac{Q_{in} - Q_{out}}{2Q_{in} - Q_{out}} \in (0, \tfrac{1}{2}).$$

Moreover, ρ^*_{LA} is increasing in β for $\beta < \beta^*$ and decreasing in β for $\beta > \beta^*$; likewise, ρ^*_{HA} is decreasing in β for $\beta < \beta^*$ and increasing in β for $\beta > \beta^*$. This proves part (b). To prove (c), note that ρ^*_{LA} and ρ^*_{HA} are decreasing and increasing in q, respectively. Finally, prove (d), note that ρ^1_{LB} and ρ^1_{HB} are increasing and decreasing in d, respectively, while ρ^2_{LA} and ρ^2_{HA} are independent of d; hence, ρ^*_{LA} and ρ^*_{HA} are increasing and decreasing in d, respectively. □

A.3.2 Limited preference heterogeneity

We next extend the comparative statics for games with identical preferences to games with limited preference heterogeneity. Notice that that the specification of a game includes payoffs as well as the sociocultural parameters β, q, and d. The payoffs are defined by specifying a payoff function u and a distribution F of risk parameters. Hence, we write $\mathcal{G} = (u, F, \beta, q, d)$ for a game with heterogeneous preferences, where u is a (linear) payoff function, F is a (continuous, unimodal, and symmetric) distribution of risk parameters, β is the level of diversity, q is the culture strength, and d is the cultural distance between group; and we write $\mathcal{G} = (u, \rho, \beta, q, d)$ for a game with identical preferences, where ρ is the common risk parameter, and other terms are defined as before.

The following result shows that the introspective equilibrium of games with limited preference heterogeneity is "close" to the introspective equilibrium of the game with identical preferences. We prove the result for weaker conditions than in the main text (Proposition 3.2): Rather than requiring that the distributions $F^n(\rho_j)$ are normal, we require only that they have full support and that they satisfy the following condition: For every $\eta > 0$, $x > 0$, there is $\gamma > 0$ such that

$$\limsup_{n\to\infty} \frac{F^n(\rho - x)}{F^n(\rho - \gamma x)} < \eta. \tag{A.9}$$

As noted below, this condition is satisfied by the normal distribution but also holds more generally.

Lemma A.6. [Continuity] *Fix a risk parameter ρ, sociocultural parameters β, q, d (and thus Q_{in}), and a linear payoff function $(u(\cdot, \rho_j))_{\rho_j}$. Let $(F^n(\rho_j))_n$ be a sequence of (continuous, unimodal, and symmetric) distributions with full support, mean ρ, and variance $\hat{\sigma}_n^2 > 0$ that satisfy (A.9) such that $\hat{\sigma}_n \to 0$. Then, the introspective equilibrium σ^n of the game $\mathcal{G}^n = (u, F^n, \beta, q, d)$ converges to an introspective equilibrium σ_ρ for the corresponding game $\mathcal{G} = (u, \rho, \beta, q, d)$ with identical preferences. That is, for each state $(\theta_A, \theta_B) \in \{H, L\} \times \{H, L\}$, the proportion of players playing according to σ_ρ under σ^n goes to 1 as n grows large.*

The proof is relegated to the online appendix. The results for games with limited preference heterogeneity now follow directly.

As can be seen from the proof of Lemma A.6 in the online appendix, the condition that the distributions F^n have full support and satisfy (A.9) is used only in the proof for the nongeneric case that $\rho = \tilde{\beta} Q_{in}$ or $\rho = 1 - \tilde{\beta} Q_{in}$. To interpret (A.9), note that if the distributions F^n come from the same family of distributions with a standard form – i.e., there is a distribution function $\tilde{F}(x)$ such that $F^n(\rho_j) = \tilde{F}\left(\frac{\rho_j - \rho}{\hat{\sigma}_n}\right)$ –, then (A.9) reduces to

$$\limsup_{n \to \infty} \frac{\tilde{F}(-x)}{\tilde{F}(-\gamma x)} < \eta.$$

This version makes clear that the role of condition (A.9) is to restrict the tail behavior of the distributions F^n. It holds for "thin-tailed" distributions (e.g., the normal distribution $\tilde{F}(x) = \Phi(x)$) for all $\gamma > 0$ and any $\eta > 0$ (i.e., $\limsup_{n \to \infty} \tilde{F}(-x)/\tilde{F}(-\gamma x) = 0$ for all $\gamma > 0$) but also for some heavy-tailed distributions such as distributions that fall off like a power law. Finally, continuity results such as Lemma A.6 also hold for other distributions beyond those that satisfy (A.9) and the full support condition, including the uniform distribution; showing this, however, requires a different type of proof.

A.4 Proof of Proposition 3.3

The following result characterizes the introspective equilibrium of threshold games with identical preferences. To state the result, note that for games with identical preferences, the proportion of players choosing to attack at level 0 in state $(\theta_A, \theta_B) = (\theta, \theta')$ is $m^0_{\theta\theta'}$ is given by:

$$m^0_{HH} := q; \qquad m^0_{LH} := \tilde{\beta}\tilde{q} + \beta q;$$
$$m^0_{HL} := \tilde{\beta} q + \beta \tilde{q}; \qquad m^0_{LL} := \tilde{q};$$

where $m^0_{HH} \geq m^0_{HL} \geq \frac{1}{2} \geq m^0_{LH} \geq m^0_{LL}$. Also, recall that $z := \frac{1}{2}(1+\eta)$ (and thus $\tilde{z} := 1 - z = \frac{1}{2}(1-\eta)$), where $\eta = 1 - d$ decreases in the cultural distance d.

Lemma A.7. [Threshold Games with Identical Preferences] *For any threshold game with identical preferences ($\rho_j = \rho$ for all $j \in N$), across all tie-breaking rules,*

(a) If $T \leq m^0_{LL}$, then for $\rho \leq 1$, there is an introspective equilibrium in which all players choose H; for $\rho \geq 1$, there is an introspective equilibrium in which all players choose L.

(b) If $T \in (m^0_{LL}, m^0_{LH}]$, then

- *for $\rho \leq 1 - qz$, there is an introspective equilibrium in which all players choose H;*
- *for $\rho \in [1 - qz, 1 - \tilde{q}z]$, there is an introspective equilibrium in which each player chooses the action he expects to be culturally salient;*
- *for $\rho \geq 1 - \tilde{q}z$, there is an introspective equilibrium in which all players choose L.*

(c) If $T \in (m^0_{LH}, m^0_{HL}]$, then

- *for $\rho \leq 1 - q$, there is an introspective equilibrium in which all players choose H;*
- *for $\rho \in [1 - q, 1 - \tilde{q}\tilde{z} - qz]$, then for $1 - \tilde{\beta}q \leq T$, then there is an introspective equilibrium in which minority players choose H and majority players choose the action they expect to be culturally salient; for $1 - \tilde{\beta}q \geq T$, there is an introspective equilibrium in which all players choose H;*
- *for $\rho \in [1 - \tilde{q}\tilde{z} - qz, 1 - \tilde{q}z - q\tilde{z}]$, there is an introspective equilibrium in which each player chooses the action he expects to be culturally salient;*
- *for $\rho \in [1 - q\tilde{z} - \tilde{q}z, q]$, then for $\tilde{\beta}q \geq T$, there is an introspective equilibrium in which minority players choose L and majority players choose the action they expect to be culturally salient; for $\tilde{\beta}q \leq T$, there is an introspective equilibrium in which all players choose L;*
- *for $\rho \geq q$, there is an introspective equilibrium in which all players choose L.*

(d) If $T \in (m^0_{HL}, m^0_{HH}]$, then

- *for $\rho \leq \tilde{q}z$, there is an introspective equilibrium in which all players choose H;*
- *for $\rho \in [\tilde{q}z, qz]$, there is an introspective equilibrium in which each player chooses the action he expects to be culturally salient;*
- *for $\rho \geq qz$, there is an introspective equilibrium in which all players choose L.*

(e) If $T > m^0_{HH}$, then for $\rho \geq 0$, there is an introspective equilibrium in which all players choose L; for $\rho \leq 0$, there is an introspective equilibrium in which all players choose H.

There are no other introspective equilibria.

Before presenting the proof, we note that Lemma A.7 characterizes the introspective equilibria across all tie-breaking rules, i.e., it identifies the set of introspective equilibria that can occur for some tie-breaking rule. In particular, Lemma A.7 implies that, except in nongeneric cases, there is a unique introspective equilibrium.

Proof of Lemma A.7. Recall the notation $(\pi^{HH}_{IG}, \pi^{HL}_{IG}, \pi^{LH}_{IG}, \pi^{LL}_{IG})$ for the conditional beliefs of players over states, where $\pi^{\theta\theta'}_{IG}$ is the conditional probability that a player from group G with impulse I assigns to state $(\theta_A, \theta_B) = (\theta, \theta')$.

If $T \leq m^0_{LL}$, then, for any $\rho < 1$, the unique best response at level 1 is to choose H: at level 0, the proportion of players with an impulse to choose H is at least m^0_{LL} in any state (θ_A, θ_B); so, any attack will be successful. If $\rho > 1$, then players have a strictly dominant strategy to choose L. Clearly, at level 1, all players are playing best responses to others' (level-1) strategies, so this describes the introspective equilibrium.

Next suppose that $T \in (m^0_{LL}, m^0_{LH}]$. Then, at level 0, an attack is successful if and only if $(\theta_A, \theta_B) \neq (L, L)$ (i.e., if attacking is culturally salient for at least one group). If $\rho < 1 - \pi^{LL}_{LA}$, then the unique best response at level 1 is to choose H; if $\rho \in (1 - \pi^{LL}_{LA}, 1 - \pi^{LL}_{HA})$, then the unique best response at level 1 for players is to choose the action they expect to be culturally salient; and if $\rho > 1 - \pi^{LL}_{HA}$, then the unique best response at level 1 is to choose L. Again, at level 1, all players are playing best responses to others' (level-1) strategies, so this describes the introspective equilibrium.

We next consider $T \in (m^0_{LH}, m^0_{HL}]$. Then, at level 0, an attack is successful if and only if $\theta_A = H$ (i.e., if attacking is culturally salient for group A). If $\rho < 1 - \pi^{LL}_{LA} - \pi^{LH}_{LA}$, then the unique best response at level 1 is to choose H; if $\rho \in (1 - \pi^{LL}_{LB} - \pi^{LH}_{LB}, 1 - \pi^{LL}_{HB} - \pi^{LH}_{HB})$, then the unique best response at level 1 for players is to choose the action they expect to be culturally salient; and if $\rho > 1 - \pi^{LL}_{HA} - \pi^{LH}_{HA}$, then the unique best response at level 1 is to choose :. In each of these cases, all players are playing best responses to others' (level-1) strategies, so the introspective equilibrium coincides with the level-1 strategies. It remains to consider the cases $\rho \in (1 - \pi^{LL}_{LA} - \pi^{LH}_{LA}, 1 - \pi^{LL}_{LB} - \pi^{LH}_{LB})$ and $\rho \in (1 - \pi^{LL}_{HB} - \pi^{LH}_{HB}, 1 - \pi^{LL}_{HA} - \pi^{LH}_{HA})$. In the first case ($\rho \in (1 - \pi^{LL}_{LA} - \pi^{LH}_{LA}, 1 - \pi^{LL}_{LB} - \pi^{LH}_{LB})$), the unique best response for majority players is to choose the action they expect to be culturally salient, while the unique best response for minority players is to choose H. This need not be an introspective equilibrium, so we need to consider the level-2 strategies. At level 2, all players play a best response against their belief except majority players with an impulse to choose L, i.e., type (L, A). At level 2, H is the unique best response for (L, A) if and only if $\tilde{q}\mathbb{I}_{[1-\tilde{\beta}\tilde{q} \geq T]} + q\mathbb{I}_{[1-\tilde{\beta}q \geq T]} > \rho$, which holds if and only if $1 - \tilde{\beta}q \geq T$. Hence, at level 2, either all players choose H (if T is sufficiently small), or minority players choose H while majority players choose the action they expect to be culturally salient; in either case, all types play a best response against others' level-2 strategies, so we have an introspective equilibrium. The proof for the second case ($\rho \in (1 - \pi^{LL}_{HB} - \pi^{LH}_{HB}, 1 - \pi^{LL}_{HA} - \pi^{LH}_{HA})$) is similar and thus omitted.

Next suppose that $T \in (m^0_{HL}, m^0_{HH}]$. Then, at level 0, an attack is successful if and only if $(\theta_A, \theta_B) = (H, H)$ (i.e., if attacking is culturally salient for both groups). If $\rho < \pi^{HH}_{LA}$, then the unique best response at level 1 is to choose H; if $\rho \in (\pi^{LL}_{LA}, \pi^{LL}_{HA})$, then the unique best response at level 1 for players is to choose the action they expect to be culturally salient; and if $\rho > \pi^{LL}_{HA}$, then the unique best response at level 1 is to choose L. Again, at level 1, all players are playing best responses to others' (level-1) strategies, so this describes the introspective equilibrium.

Finally, suppose that $T > m^0_{HH}$. Then, for any $\rho > 0$, the unique best response at level 1 is to choose L: at level 0, the proportion of players with an impulse to choose H is at most m^0_{HH} for any state (θ_A, θ_B); so, no attack can be successful. If $\rho < 0$, then players have a strictly dominant strategy to choose H. Clearly, at level 1, all players are playing best responses to others' (level-1) strategies, so this describes the introspective equilibrium. □

We are now ready to prove Proposition 3.3. We start with part (a). Observe that, when $\beta = 0$, $m^0_{LH} = m^0_{LL} = 1 - q =: m_L$ and $m^0_{HH} = m^0_{HL} = q =: m_H$; and when $\beta = \frac{1}{2}$, $m^0_{HL} = m^0_{LH} = \frac{1}{2}$. The characterization in Lemma A.7 for this case is illustrated in Figure 5: For $T \leq 1 - q$ and for $T > q$, introspective equilibrium and thus the probability of a successful attack is independent of the level of diversity. But for $T \in (1 - q, q]$, the probability of a successful attack depends on diversity. For any $T \in (1 - q, q]$, players in homogeneous societies ($\beta = 0$) choose the action they expect to be culturally salient whenever $\rho \in (1 - q, q)$, attack for $\rho < 1 - q$; and do not attack for $\rho > q$. For diverse

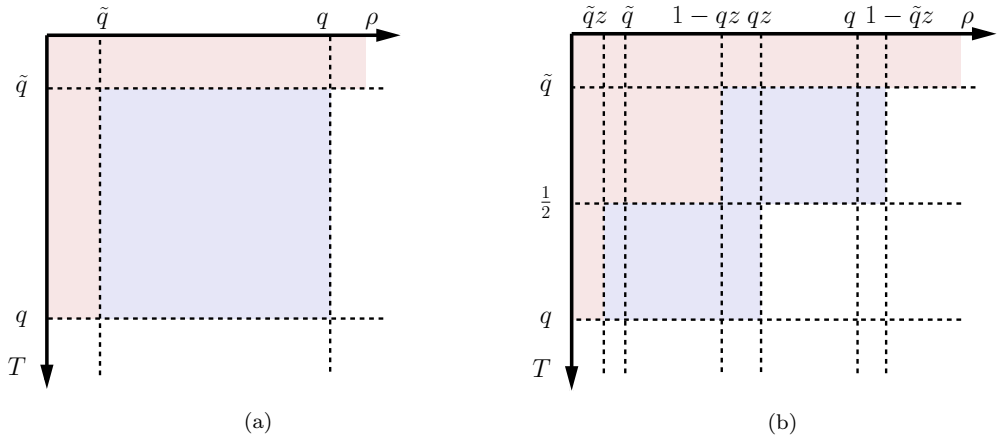

Figure 5: Introspective equilibrium for regime change models as a function of T and ρ for (a) homogeneous societies ($\beta = 0$) and (b) diverse societies ($\beta = \frac{1}{2}$). Areas shaded red and those shaded blue represent parameter combinations where players attack and choose the action they expect to be culturally salient, respectively; unshaded areas represent parameter combinations where no player attacks.

societies ($\beta = \frac{1}{2}$), for $T \in (1-q, \frac{1}{2}]$, players choose the action they expect to be culturally salient whenever $\rho \in (1-qz, 1-\tilde{q}z)$, attack for $\rho < 1-qz$; and do not attack for $\rho > 1-\tilde{q}z$. Similarly, for $T \in (\frac{1}{2}, q]$, players choose the action they expect to be culturally salient for $\rho \in ((1-q)z, qz)$, attack for $\rho < (1-q)z$; and do not attack for $\rho > q$.

Proposition 3.3(a) now follows by noting, first, that $(1-q)z < 1-q < 1-qz$ and $qz < q < 1-\tilde{q}$. Second, for parameters such that players choose the action they expect to be culturally salient in either type of society (homogeneous or diverse), the probability of the attack being successful is higher in diverse societies when the regime is weak but lower when the regime is strong. To see this, note that in homogeneous societies, if players choose the action they expect to be culturally salient ($T \in (1-q, q]$), then the proportion of players choosing to attack is q with probability $\frac{1}{2}$ and $1-q$ with probability $\frac{1}{2}$; so, an attack is successful with probability $\frac{1}{2}$. In diverse societies, if players choose the action they expect to be culturally salient ($T \in (1-q, q]$), then if $T \in (1-q, \frac{1}{2}]$, an attack is successful if attacking is culturally salient for at least one group, which happens with probability $1 - z/2 > \frac{1}{2}$; and if $T \in (\frac{1}{2}, q]$, an attack is successful if and only if attacking is culturally salient for both groups, which happens with probability $z/2 < \frac{1}{2}$.

In sum, if $T < \frac{1}{2}$, then the probability of a successful attack in a diverse society is always at least as high as in a homogeneous society (and strictly higher for intermediate values of T and ρ), both because players attack for a larger range of parameters (ρ, T), and because the likelihood of an attack succeeding is higher (for intermediate values of ρ). For $T > \frac{1}{2}$, the reverse is true: the probability of a successful attack in a homogeneous society is always at least as high as in a diverse society (and strictly higher for intermediate values of T and ρ).

It remains to prove parts (b)–(c). The proof of Proposition 3.3(b) follows from the observation that the range of payoff parameters (T, ρ) for which players choose the action they expect to be culturally salient expands as q increases. The proof of Proposition 3.3(b) follows from noting that when the regime is weak, players attack for a larger range of risk parameters as d increases ($1-qz$ and $1-\tilde{q}z$ increase with d); while the reverse is true when the regime is strong ($(1-q)z$ and qz decrease with

d). □

A.5 Proofs of Propositions 3.4–3.5

We combine the proofs for Propositions 3.4–3.5. We start by characterizing social welfare as a function of payoff parameters and diversity.

Lemma A.8. [Social Welfare] *Assume that players have identical preferences and fix a class $(u(\cdot;\rho))_{\rho\in\mathbb{R}}$ of games such that $W(m;u)$ is quadratic and convex in m.*

(a) If all players choose H, expected social welfare $\widehat{W}(u;\beta)$ is independent of diversity β;

(b) If minority players choose H while majority players choose the action they expect to be culturally salient, expected social welfare $\widehat{W}(u;\beta)$ can be a non-monotone function of β: There is b_* such that $\widehat{W}(u;\beta)$ decreases with β for $\beta < b_*$ and increases with β otherwise;

(c) If all players choose the action they expect to be culturally salient, expected social welfare $\widehat{W}(u;\beta)$ in introspective equilibrium decreases with diversity β;

(d) If minority players choose L while majority players choose the action they expect to be culturally salient, expected social welfare $\widehat{W}(u;\beta)$ can be a non-monotone function of β: There is b_{**} such that $\widehat{W}(u;\beta)$ decreases with β for $\beta < b_{**}$ and increases with β otherwise;

(e) If all players choose L, expected social welfare $\widehat{W}(u;\beta)$ is independent of β.

Proof. By assumption, social welfare when a (random) proportion m of players chooses H equals $W(m;\rho) = \hat{a}m^2+\hat{b}m+\hat{c}$ for constants $\hat{a},\hat{b},\hat{c}$ (dependent on payoffs (i.e., $u(\cdot;\rho)$) but not on sociocultural factors (i.e., β, q, d)), with $W(m;\rho)$ attaining a minimum at $\underline{m} := -\frac{\hat{b}}{2\hat{a}}$. Since $W(m;u)$ is convex, $\hat{a} > 0$ and $\hat{b} < 0$. By standard arguments, if m_σ is the (random) proportion of players who choose H under a strategy profile σ (for given β), expected welfare is

$$\widehat{W}(\rho;\beta) = \hat{a}\cdot\big(\mathbb{E}_\beta[m_\sigma]^2 + \mathrm{Var}_\beta[m_\sigma]\big) + \hat{b}\mathbb{E}_\beta[m_\sigma] + \hat{c},$$

where the expectation and variance are taken over the impulse distribution (which depends on β). It thus suffices to calculate the expectation $\mathbb{E}_\beta[m_\sigma]$ and the variance $\mathrm{Var}_\beta[m_\sigma]$ of the proportion of players choosing H in introspective equilibrium.

First, if all players choose the same action, then expected equilibrium welfare is independent of β: it equals $\widehat{W}(\rho;\beta) = \hat{a}+\hat{b}+\hat{c}$ if all players choose H ($m_\sigma = 1$) and $\widehat{W}(\rho;\beta) = \hat{c}$ if all players choose L ($m_\sigma = 0$).

Second, if all players choose the action they expect to be culturally salient, then $\mathbb{E}[m_\sigma] = \frac{1}{2}$. Moreover,

$$\mathrm{Var}[m_\sigma] = \tfrac{1}{2}(1+\eta)(q-\tfrac{1}{2})^2 + \tfrac{1}{4}(1-\eta)\big(BQ+(1-B)(1-Q)-\tfrac{1}{2}\big),$$

where $B := \beta^2+(1-\beta)^2$ and $Q := q^2+(1-q)^2 > \frac{1}{2}$. Notice that, because $\beta \le \frac{1}{2}$, B decreases with β. Hence, the change in expected welfare with β is proportional to $-\frac{d\mathrm{Var}[m_\sigma]}{dB}$, which equals $-(2Q-1) < 0$.

Third, if minority players choose H while majority players choose the action they expect to be culturally salient, then

$$\widehat{W}(\rho;\beta) = \hat{a}\cdot\left((\tfrac{1}{2}+\tfrac{1}{2}\beta)^2 + (1-\beta)^2\cdot(q-\tfrac{1}{2})^2\right) + \hat{b}\cdot(\tfrac{1}{2}+\tfrac{1}{2}\beta) + \hat{c},$$

and it follows from the first- and second-order conditions that $\widehat{W}(\rho;\beta)$ attains its (unique) minimum at

$$\mathrm{b}_* := 1 - \frac{1-\underline{m}}{\frac{1}{2}+2(q-\frac{1}{2})^2}.$$

Fourth, if minority players choose L while majority players choose the action they expect to be culturally salient, then

$$\widehat{W}(\rho;\beta) = \hat{a}\cdot(1-\beta)^2\cdot[(q-\tfrac{1}{2})^2+\tfrac{1}{4}] + \tfrac{\hat{b}}{2}(1-\beta) + \hat{c}.$$

From the first- and second-order conditions, this function attains its minimum at

$$\mathrm{b}_{**} := 1 - \frac{\underline{m}}{\frac{1}{2}+2(q-\frac{1}{2})^2}.$$ □

We are now ready to prove Proposition 3.4. By assumption, $W(1;\rho) - W(0;\rho)$ is decreasing in m. This is equivalent to the minimum $\underline{m} = \underline{m}(\rho)$ being increasing in ρ. We will use this throughout.

We first consider the case $\rho \leq \frac{1}{2}$. By Lemma A.5, there are three possibilities (depending on β and ρ) for the introspective equilibrium: (i) players choose the action they expect to be culturally salient (if $\beta > \frac{Q_{in}-(1-\rho)}{Q_{in}}$); or (ii) players choose H (if $\beta \in (\frac{1-Q_{out}-\rho}{Q_{in}-Q_{out}}, \frac{Q_{in}-(1-\rho)}{Q_{in}})$); or (iii) minority players choose H while majority players choose the action they expect to be culturally salient (for $\beta < \frac{1-Q_{out}-\rho}{Q_{in}-Q_{out}}$). For ease of exposition, denote expected welfare in introspective equilibrium under these three strategy profiles for given diversity β by $\widehat{W}_{CS}(\beta), \widehat{W}_H(\beta)$, and $\widehat{W}_{minH}(\beta)$, respectively. (Notice, we consider social welfare under each of these strategy profiles for any combination of β and ρ, regardless of whether the strategy profile is an introspective equilibrium for that particular parameter configuration.)

Fix the payoff parameters $(\hat{a},\hat{b},\hat{c})$. First, we compare $\widehat{W}_{CS}(0)$ to $\widehat{W}_{minH}(\beta)$. Using the expressions in the proof of Lemma A.8, for given $\beta \in [0,\frac{1}{2}]$, $\widehat{W}_{minH}(\beta) > \widehat{W}_{CS}(0)$ if and only if

$$-\frac{\hat{b}}{2\hat{a}} < 1 - (2-\beta)\cdot\left[(q-\tfrac{1}{2})^2+\tfrac{1}{4}\right]. \tag{A.10}$$

Notice that this inequality is easier to satisfy for larger values of β (given $\underline{m} = -\frac{\hat{b}}{2\hat{a}}$) and for smaller values of ρ (as $\underline{m}$ is increasing in ρ).

We next compare $\widehat{W}_{CS}(0)$ to $\widehat{W}_H(\beta)$. Because $\widehat{W}_H(\beta)$ is independent of β, we can define $\widehat{W}_H := \widehat{W}_H(\beta)$ (where β is arbitrary). Using the expressions in the proof of Lemma A.8, $\widehat{W}_H > \widehat{W}_{CS}(0)$ if and only if

$$-\frac{\hat{b}}{2\hat{a}} < 1 - \left[(q-\tfrac{1}{2})^2+\tfrac{1}{4}\right]. \tag{A.11}$$

Finally, we compare $\widehat{W}_H$ to $\widehat{W}_{minH}(\beta)$. By similar arguments as before, $\widehat{W}_H > \widehat{W}_{minH}(\beta)$ if and only if

$$-\frac{\hat{b}}{2\hat{a}} < 1 - (1-\beta)\cdot\left[(q-\tfrac{1}{2})^2+\tfrac{1}{4}\right]. \tag{A.12}$$

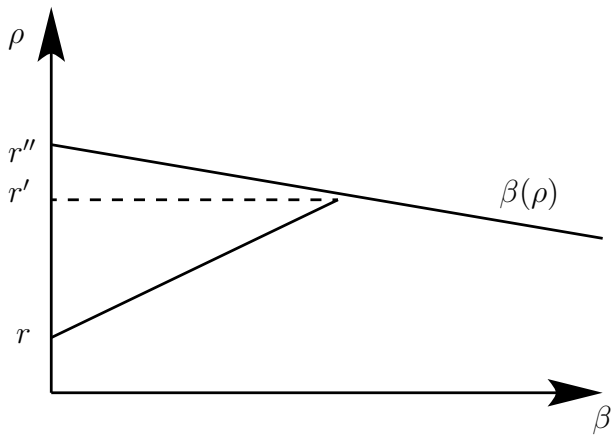

Figure 6: The different regimes for $\rho(m^2)$ and the line $\beta(\rho)$.

We can use Eqs. (A.10)–(A.12) to characterize the conditions on payoffs under which equilibrium welfare is maximized when $\beta = 0$ (i.e., cultural homogeneity is socially optimal). Notice that the left-hand side of the inequalities in Eqs. (A.10)–(A.12) is equal to the minimum $\underline{m}$ of $W(m;u)$ (and thus depends only on payoff parameters) while the right-hand side of each inequality depends only on non-payoff parameters. If we denote the right-hand side of Eqs. (A.10)–(A.12) by $m^1(\beta), m^2, m^3(\beta)$, respectively, then for every $\beta \in [0, \frac{1}{2}]$, $m^1(\beta) < m^2 \leq m^3(\beta)$ (with strict inequality if $\beta = 0$). Because the risk parameter ρ is increasing in $\underline{m}$, we can define an increasing function $\rho(\underline{m})$ that maps each $\underline{m}$ into the corresponding risk parameter ρ.

We consider different regimes for $\rho(m^2)$. The boundary values r, r', r'' for the different regimes are indicated in Figure 6, which reproduces the equilibrium characterization in Figure 4 for $\rho \leq \frac{1}{2}$; for example, $r = 1 - Q_{in}$. If cultural homogeneity is socially optimal (i.e., $\beta = 0$ is the unique value that maximizes social welfare in introspective equilibrium), then we write $\bar{\beta} = 0$; otherwise, we write $\bar{\beta} \neq 0$. Then, by the above arguments, for $\rho(m^2) < r$, $\bar{\beta} \neq 0$ for $\rho < r$ and $\bar{\beta} = 0$ for $\rho > r$. For $\rho(m^2) \in (r, r')$, then $\bar{\beta} \neq 0$ for $\rho < \rho(m^2)$ and $\bar{\beta} = 0$ for $\rho > \rho(m^2)$. Next suppose $\rho(m^2) \in (r', r'')$. Given ρ, write $\beta(\rho)$ for the value of β such that $\rho = \tilde{\beta}\tilde{Q}_{out} + \beta\tilde{Q}_{in}$; see Figure 6. For $\rho > \rho(m^2)$, clearly $\bar{\beta} = 0$. Notice that, if there is $\rho \in (r', \rho(m^2))$ such that $\widehat{W}_{minH}(\beta(\rho)) > W_{CS}(\beta = 0)$ (so $\bar{\beta} \neq 0$ for ρ), then $\widehat{W}_{minH}(\beta(\rho')) > W_{CS}(\beta = 0)$ for $\rho' \in (r', \rho)$. So, there is $\tilde{\rho}$ such that for $\rho \in (r', \tilde{\rho})$, $\bar{\beta} \neq 0$ and for $\rho \in (\tilde{\rho}, \rho^2(m^2))$, $\bar{\beta} = 0$. Likewise, for $\rho(m^2) \in (r'', \frac{1}{2})$, there is $\tilde{\rho}$ such that for $\rho < \tilde{\rho}$, $\bar{\beta} \neq 0$ and for $\rho > \tilde{\rho}$, $\bar{\beta} = 0$. So, there is $\underline{\rho} < \frac{1}{2}$ such that for $\rho < \underline{\rho}$, $\bar{\beta} \neq 0$, and for $\rho \in (\underline{\rho}, \frac{1}{2}]$, $\bar{\beta} = 0$.

Next suppose $\rho \geq \frac{1}{2}$. Again, by Lemma A.5, there are three possibilities (depending on β and ρ) for the introspective equilibrium: (i) players choose the action they expect to be culturally salient (if $\beta > \frac{Q_{in}-\rho}{Q_{in}}$); or (ii) players choose L (if $\beta \in (\frac{1-Q_{out}-(1-\rho)}{Q_{in}-Q_{out}}, \frac{Q_{in}-\rho}{Q_{in}})$); or (iii) minority players choose L while majority players choose the action they expect to be culturally salient (for $\beta < \frac{1-Q_{out}-(1-\rho)}{Q_{in}-Q_{out}}$). As before, denote expected welfare in introspective equilibrium under these three strategy profiles for given β by $\widehat{W}_{CS}(\beta), \widehat{W}_L(\beta)$, and $\widehat{W}_{minL}(\beta)$, respectively, and fix the payoff parameters $(\hat{a}, \hat{b}, \hat{c})$. We can again define $\widehat{W}_L := \widehat{W}_L(\beta)$ (for arbitrary β). By a similar argument as before, $\widehat{W}_{minL}(\beta) > \widehat{W}_{CS}(\beta = 0)$ if and only if

$$-\frac{\hat{b}}{2\hat{a}} > (2-\beta) \cdot \left[(q - \tfrac{1}{2})^2 + \tfrac{1}{4}\right]. \tag{A.13}$$

Likewise, $\widehat{W}_L > \widehat{W}_{CS}(\beta = 0)$, if and only if

$$-\frac{\hat{b}}{2\hat{a}} > (q - \tfrac{1}{2})^2 + \tfrac{1}{4}. \tag{A.14}$$

	H	L
H	u_{cc}	$(1-\delta)\,u_{\text{cd}}+\delta\,u_{\text{dd}}$
L	$(1-\delta)\,u_{\text{dc}}+\delta\,u_{\text{dd}}$	u_{dd}

Figure 7: The reduced game

Finally, $\widehat{W}_L > \widehat{W}_{minL}(\beta)$ if and only if

$$-\frac{\hat{b}}{2\hat{a}} > (1-\beta)\cdot\left[(q-\tfrac{1}{2})^2+\tfrac{1}{4}\right]. \tag{A.15}$$

Again, if we write $m^3 := (2-\beta)\cdot\left[(q-\frac{1}{2})^2+\frac{1}{4}\right]$, $m^2 := (q-\frac{1}{2})^2+\frac{1}{4}$; and $m^1 := (1-\beta)\cdot\left[(q-\frac{1}{2})^2+\frac{1}{4}\right]$ for the right-hand side expressions in (A.13)–(A.15), then we have $m^1 \leq m^2 < m^3$ (with strict inequality if $\beta > 0$). We can then apply a similar argument as before to show that there is $\overline{\rho} > \frac{1}{2}$ such that $\bar{\beta} = 0$ for $\rho \in [\frac{1}{2}, \overline{\rho}$ and $\bar{\beta} \neq 0$ for $\rho > \overline{\rho}$. □

A.6 Proof of Proposition 4.1

When the prior probability p can take on any value in $(0,1)$, we can write the joint distribution of θ_A and θ_B as

	$\theta_B = \text{c}$	$\theta_B = \text{d}$
$\theta_A = \text{c}$	$p^2+\tilde{\eta}$	$p(1-p)-\tilde{\eta}$
$\theta_A = \text{d}$	$p(1-p)-\tilde{\eta}$	$(1-p)^2+\tilde{\eta}$

where $\tilde{\eta} \in (0, p\,(1-p))$.

We start by showing that the infinitely repeated game can be analyzed using the (one-shot) linear game in Figure 7, which we refer to as the *reduced game* (where the impulse distribution for the reduced game is the same as the impulse distribution for the repeated game at $t = 0$, with H taking the place of c and L taking the place of d).

Lemma A.9. [The reduced game] *The infinitely repeated prisoner's dilemma game is strategically equivalent to the reduced game. That is, in any introspective equilibrium of the repeated game, players either choose the grim-trigger strategy or defect in every period. Moreover, a player chooses the grim-trigger strategy in the introspective equilibrium of the repeated game if and only if he chooses action H in the introspective equilibrium of the reduced game.*

The proof is relegated to the online appendix. Importantly, the reduced game is a linear game with risk parameter

$$\rho = \frac{(1-\delta)\,(u_{\text{dd}}-u_{\text{cd}})}{(1-\delta)\,(u_{\text{cc}}+u_{\text{dd}}-u_{\text{cd}}-u_{\text{dc}})+\delta\,(u_{\text{cc}}-u_{\text{dd}})}.$$

By a similar argument as before (cf. Lemma A.5), for $\beta \in \{0, \frac{1}{2}\}$, there exist $\rho_H(\beta;p)$ and $\rho_L(\beta;p)$ with $\rho_H(\beta;p) < \rho_L(\beta;p)$ such that if diversity is β, then, in introspective equilibrium, a player with impulse I chooses H if $\rho < \rho_I(\beta;p)$ and chooses L if $\rho > \rho_I(\beta;p)$. (If $\rho = \rho_I(\beta;p)$, the player's choice

depends on the tie-breaking rule.) To derive the cutoffs $\rho_H(\beta;p)$ and $\rho_L(\beta;p)$, write I_j^G for the impulse of a player from group G. Then, by Bayes' rule,

$$\begin{aligned}
\rho_H(0;p) &= \frac{pq^2+(1-p)(1-q)^2}{pq+(1-p)(1-q)};\\
\rho_L(0;p) &= \frac{q(1-q)}{p(1-q)+(1-p)q};\\
\rho_H(\tfrac{1}{2};p) &= \tfrac{1}{2}\left(\frac{pq^2+(1-p)(1-q)^2}{pq+(1-p)(1-q)}+\frac{(pq+(1-p)(1-q))^2+\tilde{\eta}(2Q-1)}{pq+(1-p)(1-q)}\right);\\
\rho_L(\tfrac{1}{2};p) &= \tfrac{1}{2}\left(\frac{q(1-q)}{p(1-q)+(1-p)q}+\frac{(p^2+(1-p)^2)q(1-q)+p(1-p)Q-\tilde{\eta}(2Q-1)}{p(1-q)+(1-p)q}\right);
\end{aligned}$$

where $Q:=q^2+(1-q)^2$. It can be checked that $0<\rho_L(0;p)<\rho_L(\frac{1}{2};p)<\rho_H(\frac{1}{2};p)<\rho_H(0;p)<1$, with each cutoff $\rho_I(\beta;p)$ increasing in p.

We are now ready to characterize the cooperation rate $\Gamma(\beta)$ for homogeneous and diverse societies ($\beta\in\{0,\frac{1}{2}\}$). Clearly, if $\rho>\rho_H(\beta;p)$ (players defect in every period), then $\Gamma(\beta)=0$; and if $\rho<\rho_L(\beta;p)$ (all players choose grim trigger), then $\Gamma(\beta)=1$. In the intermediate case $\rho\in(\rho_L(\beta;p),\rho_H(\beta;p))$, players choose the action they expect to be culturally salient. In that case,

$$\begin{aligned}
\Gamma(0) &= p\,q^2+(1-p)\,(1-q)^2;\\
\Gamma(\tfrac{1}{2}) &= (p^2+\tilde{\eta})\,q^2+\tfrac{1}{2}(p\,(1-p)-\tilde{\eta})+((1-p)^2+\tilde{\eta})\,(1-q)^2.
\end{aligned}$$

If players choose the action they expect to be culturally salient in either society (i.e., $\rho\in(\rho_L(\frac{1}{2};p),\rho_H(\frac{1}{2};p))$), then the cooperation rate is higher in homogeneous societies ($0<\Gamma(\frac{1}{2})<\Gamma(0)<1$).

The result now follows by choosing $p=\hat{p}$ appropriately, using that the cutoffs $\rho_I(\beta;p)$ are functions of p. Define

$$\underline{\rho}:=\lim_{p\downarrow 0}\rho_L(\tfrac{1}{2};p); \qquad \text{and} \qquad \underline{\rho}:=\lim_{p\uparrow 1}\rho_L(\tfrac{1}{2};p).$$

Then, set $\hat{p}=0$ if $\rho\leq\underline{\rho}$; $\hat{p}=1$ if $\rho\geq\overline{\rho}$; and take $\hat{p}$ such that $\rho=\rho_L(\frac{1}{2};\hat{p})$ for $\rho\in(\underline{\rho},\overline{\rho})$. □

A.7 Proof of Proposition 4.2

Let $\rho_I(p):=\mathbb{E}[m^0\mid I]$ be the conditional expectation of the share of players (workers) with an impulse to choose the high action given the prior probability p that H is culturally salient. Then, as in the proof of Proposition 4.1,

$$\begin{aligned}
\rho_H(p,q) &= \frac{p\,q^2+(1-p)\,(1-q)^2}{p\,q+(1-p)\,(1-q)};\\
\rho_L(p,q) &= \frac{q\,(1-q)}{p\,(1-q)+(1-p)\,q}.
\end{aligned}$$

So, at level 1, an employee with impulse I chooses H at level 1 if $\rho<\rho_I(p)$ and chooses L if $\rho>\rho_I(p)$, where

$$\rho=\tfrac{1}{2}+\frac{\lambda}{1-\lambda}(\tfrac{1}{2}-\tau),$$

as before. It is easy to check that the level-2 strategy is identical to the level-1 strategy. Hence, in introspective equilibrium, employees choose H if $\rho < \rho_L(p)$; they choose L if $\rho > \rho_H(p)$, and they choose practices consistent with the cluster they expect to be culturally salient when $\rho \in (\rho_L(p), \rho_H(p))$.

The cost of incentivizing agents to choose H (i.e., $\tau \geq \frac{1}{2}$) equals $c := \max\{0, \tau - \frac{1}{2}\}$. By the above argument, the minimum c to incentivize all players to choose H in introspective equilibrium is

$$c_H(p,q) = \max\Big\{0, \big(\frac{1}{\lambda} - 1\big)\,\big(\tfrac{1}{2} - \rho_L(p,q)\big)\Big\}.$$

The result now follows by noting that $c_H(p,q)$ decreases with p and that it decreases with q if and only if

$$p < \frac{q^2}{q^2 + (1-q)^2} =: p^*. \qquad \square$$

A.8 Proof of Proposition 4.4

A first observation is that regimes benefit more from investing in state capacity (i.e., increasing T) if the society is diverse. The proof of Proposition 3.3 implies that a regime has an incentive to invest in state capacity when the society is diverse but not if it is homogeneous: If $T \in (1-q, q)$ and $\rho \in ((1-q)\,(1-\frac{1}{2}d), 1-(1-q)\,(1-\frac{1}{2}d))$, for diverse societies ($\beta = \frac{1}{2}$), the probability of a successful attack is strictly greater when the regime is weak ($T < \frac{1}{2}$) than when the regime is strong ($T > \frac{1}{2}$) while for homogeneous societies ($\beta = 0$), it is independent of T. To analyze the tradeoff between investing in state capacity and other options, fix a family $(P^\lambda)_{\lambda \in \mathbb{R}}$ of downward sloping functions such that the opportunity cost of investing in state capacity is increasing in λ (i.e., $P^{\lambda'}(\rho) < P^\lambda(\rho)$ for all ρ whenever $\lambda' > \lambda$) and that allow for arbitrarily high opportunity cost ($P^\lambda(\rho) \to -\infty$ as $\lambda \to \infty$). The result then follows from Lemma A.7 (Figure 5): For any policy function $P(T)$, if a regime in a homogeneous society can reduce the probability of a successful attack by increasing state capacity, then so can a regime in a diverse society, but the converse does not hold: For some policy functions, the optimal state capacity for a diverse society exceeds the optimal state capacity for a homogeneous society. $\square$

References

Acemoglu, D., T. Verdier, and J. A. Robinson (2004). Kleptocracy and divide-and-rule: A model of personal rule. *Journal of the European Economic Association 2*, 162–192.

Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association 1*, 146–157.

Akerlof, G. (1976). The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics 90*, 599–617.

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

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

Akerlof, G. A. and R. E. Kranton (2010). *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-being*. Princeton University Press.

Akerlof, R. and R. Holden (2017). Capital assembly. Working paper, Warwick.

Akerlof, R., R. Holden, and L. Rayo (2017). Network externalities and market dominance. Working paper, Warwick.

Alaoui, L. and A. Penta (2016). Endogenous depth of reasoning. *Review of Economic Studies*. Forthcoming.

Alesina, A., R. Baqir, and W. Easterly (1999). Public goods and ethnic divisions. *Quarterly Journal of Economics 114*, 1243–1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*, 155–194.

Alesina, A., P. Giuliano, and B. Reich (2018). Nation-building and education. Working paper.

Alesina, A., J. Harnoss, and H. Rapoport (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth 21*, 101–138.

Alesina, A. and E. La Ferrara (2002). Who trusts others? *Journal of Public Economics 85*, 207–234.

Alesina, A. and E. La Ferrara (2005). Ethnic diversity and economic performance. *Journal of Economic Literature 43*, 762–800.

Alesina, A. and E. Spolaore (1997). On the number and size of nations. *Quarterly Journal of Economics 112*, 1027–1056.

Alesina, A. and E. Spolaore (2003). *The Size of Nations*. MIT Press.

Alesina, A., E. Spolaore, and R. Wacziarg (2000). Economic integration and political disintegration. *American Economic Review 90*, 1276–1296.

Aliprantis, C. and K. Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide* (3rd ed.). Springer.

Angeletos, G.-M., C. Hellwig, and A. Pavan (2006). Signaling in a global game: Coordination and policy traps. *Journal of Political Economy 114*(3), 452–484.

Apperly, I. (2012). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal 99*, 116–131.

Aumann, R. J. (1987). Correlated equilibria as an expression of Bayesian rationality. *Econometrica 55*, 1–18.

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies 80*, 429–462.

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

Bisin, A. and T. Verdier (2000). "beyond the melting pot": Cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics 115*(3), 955–988.

Bisin, A. and T. Verdier (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory 97*, 298–319.

Blonski, M., P. Ockenfels, and G. Spagnolo (2011). Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics 3*, 164–192.

Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and Economic Behavior 5*, 387–424.

Blume, L. E. (1995). The statistical mechanics of best-response strategy revision. *Games and Economic Behavior 11*, 111–145.

Boix, C. and M. W. Svolik (2013). The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships. *Journal of Politics 75*(2), 300–316.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*, 803–843.

Breed, W. and T. Ktsanes (1961). Pluralistic ignorance in the process of opinion formation. *Public Opinion Quarterly 25*, 382–392.

Brynjolfsson, E. and P. Milgrom (2013). Complementarity in organizations. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*. Princeton University Press.

Bueno De Mesquita, E. (2010). Regime change and revolutionary entrepreneurs. *American Political Science Review 104*(3), 446–466.

Bursztyn, L., A. González, and D. Yanagizawa-Drott (2018). Misperceived social norms: Female labor force participation in Saudi Arabia. Working paper.

Byrne, D. P. and N. De Roos (2019). Learning to coordinate: A study in retail gasoline. *American Economic Review 109*(2), 591–619.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics 119*, 861–898.

Chapple, L. and J. E. Humphrey (2014). Does board gender diversity have a financial impact? Evidence using stock portfolio performance. *Journal of Business Ethics*, 709–723.

Che, Y.-K. and N. Kartik (2009). Opinions as incentives. *Journal of Political Economy 117*, 815–860.

Chen, R. and Y. Chen (2011). The potential of social identity for equilibrium selection. *American Economic Review 101*, 2562–2589.

Costa-Gomes, M. A., V. P. Crawford, and B. Broseta (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica 69*, 1193–1235.

Crawford, V. P. (1995). Adaptive dynamics in coordination games. *Econometrica 63*, 103–143.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature 51*, 5–62.

Crémer, J. (1993). Corporate culture and shared knowledge. *Industrial and Corporate Change 2*, 351–386.

Crémer, J., L. Garicano, and A. Prat (2007). Language and the theory of the firm. *Quarterly Journal of Economics 122*, 373–407.

Dal Bó, P. and G. Fréchette (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review 101*, 411–429.

Dal Bó, P. and G. R. Fréchette (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature 56*(1), 60–114.

D'Andrade, R. (1995). *The Development of Cognitive Anthropology*. Cambridge University Press.

Demeritt, A. and K. Hoff (2018). The making of behavioral development economics. Working paper, World Bank.

Dessein, W., A. Galeotti, and T. Santos (2015). Rational inattention and organizational focus. *American Economic Review 106*, 1522–1536.

Dessein, W. and T. Santos (2006). Adaptive organizations. *Journal of Political Economy 114*, 956–995.

DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology 23*, 263–287.

Easterly, W. and R. Levine (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics 112*, 1203–1250.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies 80*(4), 1422–1458.

Egorov, G., S. Guriev, and K. Sonin (2009). Why resource-poor dictators allow freer media: A theory and evidence from panel data. *American Political Science Review 103*(4), 645–668.

Esteban, J.-M. and D. Ray (1994). On the measurement of polarization. *Econometrica*, 819–851.

Eyster, E. and M. Rabin (2005). Cursed equilibrium. *Econometrica 73*, 1623–1672.

Fearon, J. D. (2011). Self-enforcing democracy. *Quarterly Journal of Economics 126*(4), 1661–1708.

Gibbons, R. (2010). Inside organizations: Pricing, politics, and path dependence. *Annual Review of Economics*.

Gibbons, R. and R. Henderson (2013). What do managers do? Exploring persistent performance differences among seemingly similar enterprises. In R. Gibbons and J. Roberts (Eds.), *The Handbook of Organizational Economics*. Princeton University Press.

Glasze, G. and A. Alkhayyal (2002). Gated housing estates in the Arab world: Case studies in Lebanon and Riyadh, Saudi Arabia. *Environment and Planning B: Planning and Design 29*(3), 321–336.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Goldstein, I. and A. Pauzner (2005). Demand–deposit contracts and the probability of bank runs. *Journal of Finance 60*(3), 1293–1327.

Gopnik, A. and H. Wellman (1994). The "theory theory". In L. Hirschfield and S. Gelman (Eds.), *Mapping the mind: Domain specificity in culture and cognition*, pp. 257–293. Cambridge University Press.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy 102*, 912–950.

Greif, A. (2006). *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge University Press.

Greif, A. and J. Mokyr (2017). Cognitive rules, institutions, and economic growth: Douglass North and beyond. *Journal of Institutional Economics 13*(1), 25–52.

Grout, P. A., S. Mitraille, and S. Sonderegger (2015). The costs and benefits of coordinating with a different group. *Journal of Economic Theory 160*, 517–535.

Guiso, L., P. Sapienza, and L. Zingales (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives 20*, 23–48.

Guiso, L., P. Sapienza, and L. Zingales (2008). Social capital as good culture. *Journal of the European Economic Association 6*, 295–320.

Harsanyi, J. C. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press.

Heinke, M. and W. R. L. Louis (2009). Cultural background and individualistic-collectivistic values in relation to similarity, perspective taking, and empathy. *Journal of Applied Social Psychology 39*, 2570–2590.

Holmstrom, B. and P. Milgrom (1994). The firm as an incentive system. *American Economic Review*, 972–991.

Hong, L. and S. E. Page (2001). Problem solving by heterogeneous agents. *Journal of Economic Theory 97*, 123–163.

Ichniowski, C., K. Shaw, and R. W. Crandall (1995). Old dogs and new tricks: Determinants of the adoption of productivity-enhancing work practices. *Brookings Papers on Economic Activity: Microeconomics*, 1–65.

Joecks, J., K. Pull, and K. Vetter (2013). Gender diversity in the boardroom and firm performance: What exactly constitutes a "critical mass?". *Journal of Business Ethics 118*(1), 61–72.

Kandori, M., G. J. Mailath, and R. Rob (1993). Learning, mutation, and long run equilibria in games. *Econometrica 61*, 29–56.

Kaplan, S. and R. Henderson (2005). Inertia and incentives: Bridging organizational economics and organizational theory. *Organization Science 16*(5), 509–521.

Kets, W. (2019). The economic and cultural origins of cooperation. Working paper.

Kets, W., W. Kager, and A. Sandroni (2019). The value of a coordination game. Working paper.

Kets, W. and A. Sandroni (2019). A belief-based theory of homophily. *Games and Economic Behavior 115*, 410–435.

King, G., J. Pan, and M. E. Roberts (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review 107*(2), 326–343.

Knack, S. and P. Keefer (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics 112*, 1251–1288.

Knittel, C. R. and V. Stango (2003). Price ceilings as focal points for tacit collusion: Evidence from credit cards. *American Economic Review 93*(5), 1703–1729.

Kotter, J. P. and J. L. Heskett (1992). *Corporate Culture and Performance*. Free Press.

Kreps, D. M. (1990). Corporate culture and economic theory. In J. Alt and K. Shepsle (Eds.), *Perspectives on Positive Political Economy*, pp. 90–143. Cambridge University Press.

Kreps, D. M. (1996). Markets and hierarchies and (mathematical) economic theory. *Industrial and Corporate Change 5*(2), 561–595.

Kuran, T. (1987a). Chameleon voters and public choice. *Public Choice 53*, 53–78.

Kuran, T. (1987b). Preference falsification, policy continuity and collective conservatism. *Economic Journal 97*, 642–665.

Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public choice 61*, 41–74.

Kuran, T. (1991). The east european revolution of 1989: Is it surprising that we were surprised? *American Economic Review 81*, 121–125.

Kuran, T. (1997). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Kuran, T. and W. Sandholm (2008). Cultural integration and its discontents. *Review of Economic Studies 75*, 201–228.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer, and R. Vishny (1999). The quality of government. *Journal of Law, Economics, and Organization 15*, 222–279.

Laitin, D. D. and S. Jeon (2015). Exploring opportunities in cultural diversity. In R. A. Scott (Ed.), *Emerging Trends in the Social and Behavioral Sciences*, pp. 1–17.

Lazear, E. P. (1999a). Culture and language. *Journal of Political Economy 107*(S6), S95–S126.

Lazear, E. P. (1999b). Globalisation and the market for team-mates. *Economic Journal 109*, C15–C40.

Le Coq, C., J. Tremewan, and A. K. Wagner (2015). On the effects of group identity in strategic environments. *European Economic Review 76*, 239–252.

Little, A. T. (2012). Elections, fraud, and election monitoring in the shadow of revolution. *Quarterly Journal of Political Science 7*(3), 249–283.

Liu, Y., Z. Wei, and F. Xie (2014). Do women directors improve firm performance in China? *Journal of Corporate Finance 28*, 169–184.

Locke, J. (1690/1975). *An essay concerning human understanding*. Oxford University Press.

Lohmann, S. (1994). The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics 47*(1), 42–101.

Lorentzen, P. L. (2013). Regularizing rioting: Permitting public protest in an authoritarian regime. *Quarterly Journal of Political Science 8*, 127–158.

McKelvey, R. D. and T. R. Palfrey (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior 10*, 6–38.

Mill, J. S. (1872/1974). *A system of logic, ratiocinative and inductive*, Volume 7 of *Collected works of John Stuart Mill*. University of Toronto Press.

Milliken, F. J. and L. L. Martins (1996). Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Academy of Management Review 21*, 402–433.

Morris, S. (2000). Contagion. *Review of Economic Studies 67*, 57–78.

Morris, S. and H. Shin (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review 88*, 587–597.

Morris, S. and H. S. Shin (2003). Global games: Theory and applications. In M. Dewatripont, L. P. Hansen, and S. J. Turnovsky (Eds.), *Advances in economics and econometrics: Eighth World Congress*, Volume 1. Cambridge University Press.

Morris, S. and H. S. Shin (2004). Coordination risk and the price of debt. *European Economic Review 48*(1), 133–153.

Morris, S. and M. Yildiz (2019). Crises: Equilibrium shifts and large shocks. *American Economic Review 109*(8), 2823–2854.

Myerson, R. B. (2004). Justice, institutions, and multiple equilibria. *Chicago Journal of International Law 5*, 91–108.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review 85*, 1313–1326.

Nelson, D. W. and R. Baumgarte (2004). Cross-cultural misunderstandings reduce empathic responding. *Journal of Applied Social Psychology 34*, 391–401.

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication 24*(2), 43–51.

North, D. C. (2005). *Understanding the Process of Economic Change*. Princeton University Press.

Padró i Miquel, G. (2007). The control of politicians in divided societies: The politics of fear. *Review of Economic studies 74*(4), 1259–1274.

Page, S. E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Persson, T. and G. Tabellini (2009). Democratic capital: The nexus of political and economic change. *American Economic Journal: Macroeconomics 1*(2), 88–126.

Prat, A. (2002). Should a team be homogenous? *European Economic Review 46*, 1187–1207.

Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century. *Scandinavian Political Studies 30*(2), 137–174.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Schultz, C. G. and C. Neighbors (2007). Perceived norms and alcohol consumption: Differences between college students from rural and urban high schools. *Journal of American College Health 56*(3), 261–265.

Singer, T. and E. Fehr (2005). The neuroeconomics of mind reading and empathy. *American Economic Review 95*, 340–345.

Stahl, D. O. and P. W. Wilson (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*, 218–254.

Sunstein, C. R. (2018). Unleashed. *Social Research: An International Quarterly 85*, 73–92.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review 51*, 273–286.

Tabellini, G. (2008). The scope of cooperation: Values and incentives. *Quarterly Journal of Economics 8*, 905–950.

Tajfel, H. and J. Turner (1986). The social identity theory of intergroup behavior. In S. Worchel and W. G. Austin (Eds.), *Psychology of Intergroup Relations*, pp. 7–24.

Van Boven, L. (2000). Pluralistic ignorance and political correctness: The case of affirmative action. *Political Psychology 21*, 267–276.

Van den Steen, E. (2010). Culture clash: The costs and benefits of homogeneity. *Management Science 56*, 1718–1738.

Van Zandt, T. and X. Vives (2007). Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory 134*, 339–360.

Vives, X. (1990). Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics 19*, 305–321.

Vives, X. (2005). Complementarities and games: New developments. *Journal of Economic Literature 43*, 437–479.

Williams, H. M., S. K. Parker, and N. Turner (2007). Perceived dissimilarity and perspective taking within work teams. *Group and Organizational Management 32*, 569–597.

World Bank (2015). *World Development Report 2015: Mind, Society, and Behavior*. Washington, DC: World Bank.

Young, H. P. (1993). The evolution of conventions. *Econometrica 61*, 57–84.

Zou, X., K. Tam, M. W. Morris, S. Lee, I. Lau, and C. Chiu (2009). Culture as common sense: Perceived consensus versus personal beliefs as mechanisms of cultural influence. *Journal of Personality and Social Psychology 97*(4), 579–597.