A Theory of Strategic Uncertainty and Cultural Diversity *

Willemien Kets† Alvaro Sandroni‡

September 29, 2017

Abstract

This paper presents a novel mechanism through which culture can affect behavior. Cultural diversity matters because it influences the degree of strategic uncertainty that players face. We model this by building on research in psychology on perspective taking. The model delivers comparative statics that are broadly consistent with experimental evidence and that are difficult to obtain with existing methods. In addition, it can account for a variety of disparate evidence, including why inefficient social customs persist in some societies but not in others and why exclusively targeting incentives may not help with resolving collective action problems.

*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, 2015). We are grateful to the Associate Editor and three anonymous referees for excellent suggestions. We thank Larbi Alaoui, Vincent Crawford, Yan Chen, Vessela Daskalova, Wouter Dessein, Matthew Jackson, Wouter Kager, Rachel Kranton, Bart Lipman, George Mailath, Niko Matouschek, 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, Yiqing Xing, and numerous seminar audiences and conference participants for helpful comments and stimulating discussions.

†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 today's interconnected world, 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 if groups differ in "hard," payoff-relevant factors such as information, preferences, or skills,[1] it is silent on the effect of culture if it is not directly payoff-relevant. That cultural difference matter beyond payoffs is well-documented. To give an example, mergers often fail to live up to expectations, not because of a dearth of economic benefits but due to incompatibilities in culture. Such cultural incompatibilities can hamper economic performance even if all players face the same incentives (Shrivastava, 1986).

The challenge for economic theory is thus to develop a methodology that can model how cultural differences affect economic outcomes even if they do not directly influence incentives. Equally important is to demonstrate that the methodology can deliver testable implications and yield new economic insights.

This paper takes a first step in this direction. We introduce a new solution concept based on insights from psychology under which cultural diversity matters even if it does not directly affect payoffs. This allows us to address new questions, such as how the relative economic performance of culturally homogeneous and culturally diverse societies depends on the economic environment. We derive novel comparative statics that are broadly consistent with experimental evidence and that are difficult to obtain using existing methods. This yields new economic insights in a range of economic applications.

In our model, players belong to different cultural groups. A society in which all players belong to the same cultural group is culturally homogeneous; otherwise, it is culturally diverse. Groups are identical in all payoff-relevant factors (i.e., payoffs and information). But, following the prevailing view in sociology and anthropology (e.g., DiMaggio, 1997), we assume that cultural groups differ in the mental models (or "schemas") that they use. That is, players who belong to different cultural groups have different conscious and unconscious preconceptions and beliefs that organize their thinking. As a result, players from different cultural groups may react differently to the same situation even if they face identical incentives.

While intuitive, this is difficult to capture using standard game-theoretic methods, as standard game theory lacks a formal language to model how culture influences strategic beliefs. To model the effects of culture, we thus build on research in psychology on theory of mind.

[1] This has been studied extensively in the context of organizations; see, e.g., Lazear (1999), Hong and Page (2001), Che and Kartik (2009), and Van den Steen (2010). Beyond the context of organizations, see Alesina and La Ferrara (2005) and references therein.

Theory of mind is a central concept in psychology. It refers to the ability to take another person's perspective. Perspective-taking is a process whereby an individual, starting from his own instinctive reaction to a situation, reasons about the other person's response. In our model, a player's instinctive response takes the form of an impulse, that is, a payoff-irrelevant signal that suggests an action. As there are often commonalities in how people react to a given situation, a player can use his own impulse to form a belief about other players' impulses. He can then formulate a best response to this belief. As he realizes that other players may have gone though a similar process, he may revise his belief and adjust his response. This process continues to higher orders. The limit of this process is an *introspective equilibrium*.

We study coordination games with many players. In these games, players' best response depends on what they expect other players to do. Hence, players use perspective-taking to decide on their action. To capture that players from the same cultural group share similar mental models, we assume that impulses are more strongly correlated within groups than across groups. This implies that a player's impulse is more informative of the impulses of players who belong to the same group as himself. Thus, players face less *strategic uncertainty* (i.e., a priori uncertainty about impulses) when there is a single culture.

Our main result characterizes the relative economic performance of different societies as a function of the economic environments (i.e., payoffs), where economic performance is measured by the (expected) aggregate payoff in introspective equilibrium. When the actions are highly asymmetric in terms of payoffs, then introspective equilibrium selects a unique Nash equilibrium and which Nash equilibrium is selected is independent of the degree of cultural diversity. Intuitively, if the options are highly asymmetric, then there is an obvious way to play the game, and cultural diversity does not affect economic performance.

However, in less extreme cases, cultural diversity can either be a cost or a benefit depending on the economic environment. To see this, consider a player who has an impulse to take a certain action. If the payoff structure of the game gives little guidance (i.e., the actions are nearly symmetric in terms of payoffs), then a player with an impulse to play a given action has no reason to deviate from this impulse. But, if all players follow their impulse in introspective equilibrium, then, given that impulses are more strongly correlated within groups, behavioral consistency is improved if all players belong to the same group. That is, *if the actions are nearly symmetric in terms of payoffs, then cultural homogeneity reduces the risk of miscoordination.* This echoes the insight of Kreps (1990) that culture is a source of focal principles that can facilitate coordination. In this case, culturally homogeneous societies outperform culturally diverse ones.

The situation is different when there is some asymmetry in terms of payoffs (but not so much that there is an obvious way to play the game). In this case, the like-mindedness of

culturally homogeneous societies can be a curse. In a culturally homogeneous society, strategic uncertainty is limited. Consequently, behavior is strongly guided by expectations about other players' behavior. This means that a culturally homogeneous society may get locked into playing a focal equilibrium even if all players would prefer (collectively) to switch to a different equilibrium. By contrast, culturally diverse societies lack congruent expectations, so choices are more strongly guided by payoff considerations. We show that *if there is some asymmetry in payoffs and coordinating on the "good" Nash equilibrium is not too risky, then cultural diversity reduces the risk of inefficient lock-in.* In this case, culturally diverse societies have higher payoffs in equilibrium than culturally homogeneous ones.[2]

In sum, cultural diversity can be a double-edged sword. It can be a cost or a benefit depending on the economic environment. This is consistent with the view in sociology that *culture both enables and constrains* (Sewell, 1992). On the one hand, a common culture enables effective coordination; but, on the other hand, shared beliefs may constrain players in their choices if these shared beliefs anchor players' expectations which then become self-fulfilling. These subtle effects are above and beyond any direct impact that cultural diversity may have on the payoffs.

After deriving the main result and describing its testable implications, we use the model to shed light on a range of economic applications. As has been well-documented, social customs can persist even if they are inefficient, and this is associated with significant welfare costs (Akerlof, 1976, 1980). Using the model, we characterize the economic and sociocultural conditions under which inefficient social customs can be sustained. An implication of the analysis is that a small change in economic conditions can lead to a divergence in economic outcomes for societies that are identical in all payoff-relevant aspects.

We next turn to collective action problems. If all players benefit from each others' investment, then a society faces a free-riding problem. If, in addition, there are social controls in place that make that a player who free-rides suffers a social sanction that increases with the proportion of players who invest, then there is a coordination component: cooperation may be viable but only if sufficiently many other players invest. We show that if investment is socially optimal but risky, then free-riding cannot be eliminated entirely in equilibrium. Depending on the degree of cultural diversity, all players free-ride or there is partial (but not complete) investment and a positive fraction of players incurs a social sanction. In this environment, policies designed to alleviate incentive problems may not be effective if they ignore coordination problems: Increasing the cost of social sanctions (for example, by increased monitoring)

[2]When there is some asymmetry in payoffs but coordinating on the "good" Nash equilibrium is risky, there can be a tension between reducing the risk of miscoordination and avoiding inefficient lock-in; see Sections 3.3 and 4.2.

increases not only the incentive to invest, it also increases the cost of miscoordination (because free-riders incur a greater cost). This reduces the total payoff in the introspective equilibrium with partial investment while it does not affect the total payoff in the introspective equilibrium in which all players free-ride. So, increasing the cost of social sanctions may be ineffective. Instead, policies that make investing focal can lead to large welfare gains.

Thus far, we have focused on the optimal population composition. However, in most cases, societies cannot freely choose their population composition even if they can affect it at the margin. An important question is thus how the effect of small changes in the population composition compares with the effect of larger changes. Consider, for example, a culturally homogeneous organization. As we have seen, cultural homogeneity is suboptimal if concordant expectations leads players to coordinate on an inefficient but focal Nash equilibrium. Nevertheless, small changes in the composition of the workforce may lead to lower equilibrium payoffs: As we show, cultural diversity improves economic performance only once the minority reaches a critical mass. This nonmonotonic effect of diversity on economic outcomes cannot be obtained with existing approaches: If the benefits of cultural diversity are purely skill- or information-based, then one would expect that even a small minority would have a positive impact on performance.

Finally, we show how our insights on cultural diversity can be translated to analyze how the optimal organizational culture varies with the economic environment, and to show that leaders are sometimes more effective if they are more ambiguous in their communication.

The remainder of this paper is organized as follows. The next section presents the basic model. Section 3 presents the main theoretical results. Section 4 uses the model to study various applications. Section 5 compares the predictions with experimental evidence and predictions from other models. Section 7 discusses the related literature, and Section 8 concludes. Proofs are relegated to the appendices.

2 Model

2.1 Introspection

The set of players is $N = [0, 1]$. Each player $j \in N$ has two actions, denoted s^0, s^1. As is standard, we write $S_j := \{s^0, s^1\}$ for the action set of player j and $S_{-j} := \prod_{i \neq j} S_i$ for the set of action profiles for players $i \neq j$. The payoff to player $j \in N$ of choosing action $s_j \in S_j$ when the other players play according to $s_{-j} \in S_{-j}$ is $u_j(s_j, s_{-j})$. A player's payoff depends only on his own action and the proportion of players taking each action, i.e., $u_j(s_j, s_{-j}) = u(s_j, m)$, where m is the proportion of players choosing s^1 under s_{-j}.

In many games of interest, players cannot deduce from the payoff structure alone what the other players will do. Thus, even if the game is fixed and known, players often face considerable *strategic uncertainty*, that is, uncertainty about the other players' actions. A standard approach in game theory is to select a Nash equilibrium, thus resolving any strategic uncertainty players may face. We depart from standard game theory in that we explicitly model the process by which players resolve this strategic uncertainty. As observed by Schelling (1960), when facing strategic uncertainty, "[the] 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 ability to take another person's perspective. This involves introspection: to form a belief about another person, people first observe their own mental state.[3] This is a rapid and instinctive process referred to as first-person simulation (Goldman, 2006). It is followed by a slower, more deliberative process whereby individuals reason about others' mental states using a naive understanding of psychology. This may lead them to adjust their initial belief (Gopnik and Wellman, 1994). Thus, individuals use the observation of their own mental state to form an initial belief and then use "folk psychology" to adjust this belief. In decision problems, the relevant mental states are the states that govern behavior. A player's initial mental state then takes the form of an inclination to act without deliberation. We will refer to the pre-reflective inclination to take a certain action as an *impulse*.

We formalize the introspective process as follows. Each player receives an impulse, drawn from a common prior which we refer to as the impulse distribution. Impulses are not observable by other players, and do not have a direct effect on payoffs. A player's instinctive reaction is to follow his impulse. For example, if the player's impulse is $I_j = s^1$, then his pre-reflective inclination is to take action s^1. This defines the level-0 strategy σ_j^0 for each player $j \in N$ (i.e., $\sigma_j^0(I_j) = I_j$ for impulse $I_j \in S_j$). Through introspection, players realize that other players also have an impulse. Players can use their own impulse to form a belief about others' impulses: Using Bayes' rule, players can derive the distribution of other players' impulses conditional on their own impulse. The process of first-person simulation thus yields a belief about other players' instinctive responses (i.e., their level-0 strategies). Of course, a player's instinctive reaction need not be optimal given other players' actions. If a player's instinctive reaction is not a best response against the belief that others follow their impulse, then he adjusts his initial response. This defines his level-1 strategy σ_j^1: for $I_j \in S_j$, $\sigma_j^1(I_j) \in S_j$ is a best response for player j against the belief that other players follow their impulse, conditional on having

[3]These ideas also have a long history. 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."

received impulse I_j.[4] The reasoning need not stop here. As a "folk game theorist," the player realizes that the other players will go through a similar reasoning process and formulate their level-1 strategy (i.e., a best response to their beliefs about other players' level-0 strategies). Again, if his level-1 strategy is not optimal against the level-1 strategies of the other players, then he adjust his response. This defines his level-2 strategy σ_j^2. In general, for $k > 1$, the *level-k strategy* σ_j^k is a best response to the level-$(k-1)$ strategies of the other players. Players go through this reasoning process in their mind before taking a decision. Accordingly, player j's behavior is given by the limit $\sigma_j := \lim_{k\to\infty} \sigma_j^k$ of the introspective process. The profile $\sigma = (\sigma_j)_{j\in N}$ of limiting strategies defines an *introspective equilibrium.*

Theory of mind thus involves a fast, instinctive process of first-person simulation and a slower, deliberate reasoning process anchored by the initial intuitions. In case there is a conflict between a player's intuition and reasoning in the sense that they point to different actions, the player adjusts his response. As other players may likewise adjust their response, the player's contemplated course of action may not remain optimal under these modified strategies. If that is the case, the player needs to engage in further reasoning. Thus, unlike a "folk psychologist" who faces a single-person decision situation, a "folk game-theorist" must reason to higher orders, represented here by the different levels. As the introspective process converges, the strategies for each player become consistent across levels. Thus, in introspective equilibrium, the conflict between the different levels of reasoning is resolved. The next result shows that once these conflicts are resolved within each player, the resulting strategies are also consistent across players in that they form a correlated equilibrium. To state the result, we make a technical assumption to avoid measurability issues: we assume that there is an underlying (payoff irrelevant) state θ such that impulses are conditionally independent given the state. We also assume that a player's payoff is continuous in the proportion of players choosing each action.

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

By the epistemic characterization of correlated equilibrium of Aumann (1987), Proposition 2.1 implies that introspective equilibrium is always consistent with common knowledge of rationality. That is, in introspective equilibrium, every player acts as if he is rational, believes that the other players are rational and that they believe that others are rational,..., ad infinitum. Proposition 2.1 establishes that there is a remarkable consistency between the proposed solutions of psychologists and game theorists to the problem of strategic uncertainty. Even

[4]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.

though players are assumed to be "folk game theorists" who do not engage in a full-fledged equilibrium analysis, the strategies they end up choosing must be consistent with common knowledge of rationality, which is one of the central assumptions in game theory.

The intuition behind Proposition 2.1 is simple. In our model, every player receives a signal (i.e., an impulse), and his strategy is a function of his signal. If all types choose a best response given their belief, then we obtain a correlated equilibrium. So, to show that an introspective equilibrium is a correlated equilibrium, it suffices to show that in the limit, every type chooses a best response against other players' strategies. This follows because in any introspective equilibrium, players must choose actions that survive the iterative elimination of strictly dominated actions (under a common prior). These are precisely the actions that can be played in a correlated equilibrium (Aumann, 1987).

Proposition 2.1 provides a new foundation for correlated equilibrium based on central tenets in psychology. Hence, when it exists, introspective equilibrium inherits any testable implication that correlated equilibrium might have. Moreover, as we will see, the introspective process in fact selects a unique introspective equilibrium in the games that we consider. So, introspective equilibrium provides a foundation for correlated equilibrium as well as a refinement of it.

2.2 Culture

We next discuss how culture influences the introspective process. In our model, the introspective process is anchored by the impulses. A player's first instinct may be to take an action that stands out among the others. But, which action stands out may depend on the player's background. To give a well-known example, Schelling (1960, pp. 55–56) showed that Grand Central station is a salient meeting place for people who live near New York City, while other locations may be more salient for nonlocals. If culture influences how players respond to a given context, then the instinctive responses of members of different cultural groups may differ even if they face the same context and have the same incentives.

To model this, we assume that culture influences the distribution of impulses. Each player belongs to one of two *(cultural) groups*, labeled by $\gamma = A, B$. Players know which group they belong to. For each cultural group $\gamma = A, B$, there is a state $\theta_\gamma = s^0, s^1$ where each state is equally likely. Conditional on $\theta_\gamma = s$, each player $j \in \gamma$ has an impulse to play s with probability $q \in (\frac{1}{2}, 1)$, independently across players. The parameter q captures the degree to which the player is attuned to the state of his cultural group: if q is close to 1, then a player's impulse is strongly correlated with the state of his cultural group; if q is close to $\frac{1}{2}$, then his impulse and the state of his cultural group are almost independent. Thus, q measures *culture*

strength. Conditional on having impulse $I = s$, a player assigns probability

$$Q_{in} := q^2 + (1-q)^2$$

to a member of his group having impulse s. The probability Q_{in} lies strictly between $\frac{1}{2}$ and 1 and increases with culture strength (i.e., q). We emphasize that the assumption that impulses are equally likely a priori is not critical for our results; it merely simplifies the calculations. The key idea is that a player's impulse is an informative but noisy signal about other players' impulses.

We next discuss the relation between the cultures of different groups. Cultures obviously differ in many respects, but they also share commonalities. To model this, we assume that the states θ_A, θ_B of the groups are imperfectly correlated. Without loss of generality, we use the following parametrization of the joint distribution over states:

	$\theta_B = s^1$	$\theta_B = s^0$
$\theta_A = s^1$	$\frac{1}{4} + \eta$	$\frac{1}{4} - \eta$
$\theta_A = s^0$	$\frac{1}{4} - \eta$	$\frac{1}{4} + \eta$

The parameter η is the covariance between the states. If $\eta = 0$, then the states are independent; if $\eta = \frac{1}{4}$, then the states are perfectly correlated. We will take η to be (strictly) between 0 and $\frac{1}{4}$, so that the states are imperfectly correlated. If states are almost independent (i.e., η close to 0), then a player's impulse is almost completely uninformative of the impulses of members of the other group; if the states are almost perfectly correlated (i.e., η close to $\frac{1}{4}$), then a player's impulse is almost as informative of the impulses of the other group as of her own. Thus, $\delta := 1 - \eta$ reflects the *cultural distance* between the two groups. Conditional on having impulse $I = s$, a player assigns probability

$$Q_{out} := q^2 \cdot (\tfrac{1}{2} + 2 \cdot (1-\delta)) + 2q \cdot (1-q) \cdot (\tfrac{1}{2} - 2 \cdot (1-\delta)) + (1-q)^2 \cdot (\tfrac{1}{2} + 2 \cdot (1-\delta))$$

to a member of the other group having impulse s. The probability Q_{out} lies strictly between $\frac{1}{2}$ and Q_{in} and decreases with cultural distance (i.e., δ). Hence, players who belong to the same group are more likely to have the same impulse than players who belong to different groups; and this difference is more pronounced for groups that are not culturally close.[5] In Appendix A, we show formally that players face more strategic uncertainty if they interact with players from another group.

[5]The group structure itself may also affect impulses. For example, if two people arrive at a door at the same time, then, if one of them is a woman and the other one is a man, then at least in some societies, there is the clear expectation that the man holds the door and let the woman go first (see Blume, 2000, for a related point). We abstract away from this effect.

Even though this model is admittedly stylized, it captures some intuitive ideas. Individuals may differ in the set of circumstances under which they are inclined to take a certain action. To give an example, when communicating with a subordinate, some people may be inclined to go by that subordinate's desk and speak with him directly while others are inclined to communicate by email. While there can be considerable individual heterogeneity even within a given cultural group, the differences tend to be greater across groups. For example, some cultures put more emphasis on personal relations than others. Since both cultural differences and individual heterogeneity play a role, differences between cultures are gradual and probabilistic rather than absolute and discrete. This is in line with the prevailing view in anthropology that "cross-cultural differences are not necessarily categorical, i.e., 'These people do x; those people don't do x.' Instead, ... [m]any differences are differences in statistical frequency, i.e., 'These people do x much more frequently than those people'; other differences are context-sensitive, i.e., 'These people do x under these circumstances; those people do not do x under the same circumstances but they do x under different circumstances'" (Ochs, 1988, p. 135). In our model, this is captured by random impulses: in a given situation, some players are inclined to choose action s^0, while others are inclined to choose s^1. Cultural differences are captured by the assumption that impulses are more strongly correlated within groups than across groups (i.e., $Q_{in} > Q_{out}$).[6]

An advantage of this probabilistic approach is that there is no need to specify exactly the set of circumstances under which an individual has an inclination to take a given action. Rather, we model the *aggregate* patterns of relations between the instinctive reactions of different groups (across situations). As we shall see, this "detail-free" approach will allow us to shed light on aggregate behavior even if the model is silent on how a given individual reacts to a given situation.

2.3 Cultural diversity

To analyze how diversity affect economic outcomes, we compare the strategic behavior of players who belong to different societies. A *society* is a tuple $\mathcal{S} = (\alpha, q, \delta)$, where α is the proportion of players who belong to group A, q measures culture strength, and δ is the cultural distance between groups. The society's characteristics (i.e., α, q, δ) are common knowledge among the players. Since groups are symmetric, we can assume without loss of generality that group A forms the *majority* while group B forms the *minority* (i.e., $\alpha \geq \frac{1}{2}$). A

[6] Another way cultural difference may manifest themselves is that groups may differ in their inclination to take a certain action. For much of the paper, we abstract away from this so as to focus on the effect of strategic uncertainty per se, though see the discussion of dysfunctional cultures in Section 6.

society is *culturally homogeneous* if it is dominated by a single cultural group (i.e., α close to 1); conversely, a society is *culturally diverse* if players are roughly equally divided over groups (i.e., α is close to $\frac{1}{2}$). We will refer to α as the *population composition*. While the characteristics of a society (i.e., α, q, δ) are not directly payoff relevant, we will see that they can nevertheless influence strategic behavior, through their effect on the impulse distribution.

3 Coordination

3.1 Game

In this section, we apply the concepts of introspection and culture to study coordination problems. The payoff to a player of choosing an action increases linearly with the proportion of players $i \neq j$ choosing that action. That is, if player $j \in N$ chooses action $s = s^0, s^1$ and the other players play according to the action profile $s_{-j} := (s_i)_{i \neq j}$, then the payoff to j is

$$u_j(s, s_{-j}) = \begin{cases} m_0(s_{-j}) \cdot v_0 - c_0 & \text{if } s = s^0; \\ m_1(s_{-j}) \cdot v_1 - c_1 & \text{if } s = s^1; \end{cases} \tag{3.1}$$

where $m_n(s_{-j})$ is the proportion of players $j \neq i$ choosing action s^n under s_{-j}, and v_0, v_1, c_0, c_1 are constants satisfying $v_0, v_1 \geq 0$ (with at least one strict inequality). So, the payoff function consists of two parts: the intrinsic utility or cost of taking action s^n (given by $-c_n$) and a term $m_n \cdot v_n$ that increases with the proportion m_n of players that take that action. We assume that either action can be a best response provided sufficiently many other players choose it, that is,

$$\begin{aligned} v_0 - c_0 &> -c_1; \\ v_1 - c_1 &> -c_0. \end{aligned} \tag{3.2}$$

So, the game has two strict Nash equilibria,[7] and players have an incentive to coordinate their action. Without loss of generality, coordinating on s^1 (weakly) Pareto dominates coordinating on s^0, i.e.,

$$v_1 - c_1 \geq v_0 - c_0.$$

We refer to s^1 as the *efficient* action and to the Nash equilibrium in which all players choose s^1 as the efficient Nash equilibrium. Likewise, we refer to s^0 as the *inefficient* action and to the Nash equilibrium in which all players choose s^0 as the inefficient Nash equilibrium.

[7] In addition, the game has unstable Nash equilibria in which a proportion $m^* := \frac{v_0 + c_1 - c_0}{v_0 + v_1}$ of players choose s^1. Besides being unstable, this class of Nash equilibria has the unappealing property that the proportion of players choosing s^1 *decreases* with the payoff v_1 (and increases with c_1).

3.2 Introspective equilibrium

We start by characterizing the introspective equilibrium for this class of games. A key statistic is the *risk parameter*, defined by

$$\rho := \frac{v_0 + c_1 - c_0}{v_0 + v_1}.$$

The risk parameter depends only on the payoff parameters. It is a measure of the risk associated with choosing the efficient action. If a player expects that the proportion of players that choose the efficient action is at least ρ, then it is a best response for him to choose the efficient action. If the risk parameter is close to 0 (i.e., v_1 high and c_1 low), then playing the efficient action is optimal for a player unless he expects most other players to choose the inefficient action; in this sense, the efficient action is not very risky. On the other hand, if the risk parameter is close to 1 (i.e., v_1 low and c_1 high), then choosing the inefficient action is a best response for a player unless he expects most players to choose the efficient action. In this sense, choosing the efficient action is risky. Thus, if the risk parameter is close to 0 or 1, then there is significant asymmetry between the actions in terms of payoffs. In the intermediate case (i.e., ρ close to $\frac{1}{2}$), the payoff structure of the game gives little guidance: both actions are equally attractive in the sense that a player who assigns equal probability to players choosing each action is indifferent. The following result characterizes the introspective equilibrium as a function of the risk parameter ρ:

Theorem 3.1. [Introspective Equilibrium: Existence and Characterization] *Consider a society $\mathcal{S} = (\alpha, q, \delta)$ and suppose the game (3.1) has risk parameter ρ. Then, an introspective equilibrium exists: there exist $\rho_1 = \rho_1(\alpha, q, \delta)$, $\rho_2 = \rho_2(\alpha, q, \delta)$, $\rho_3 = \rho_3(\alpha, q, \delta)$, $\rho_4 = \rho_4(\alpha, q, \delta)$, with $0 < \rho_1 < \rho_2 < \frac{1}{2} < \rho_3 < \rho_4 < 1$, such that*

(a) If $\rho \geq \rho_4$, then there is an introspective equilibrium in which all players choose the inefficient action.

(b) If $\rho \in [\rho_3, \rho_4]$, then there is an introspective equilibrium in which players from the minority group choose the inefficient action, while players from the majority group follow their impulse.

(c) If $\rho \in [\rho_2, \rho_3]$, then there is an introspective equilibrium in which all players follow their impulse.

(d) If $\rho \in [\rho_1, \rho_2]$, then there is an introspective equilibrium in which players from the minority group choose s^1, while players from the majority group follow their impulse.

(e) If $\rho \leq \rho_1$, then there is an introspective equilibrium in which all players choose s^1.

By establishing existence of introspective equilibrium, Theorem 3.1 demonstrates that in coordination games, any conflict between instinctive reactions and reasoned responses can be resolved: after sufficiently many steps of reasoning, the process converges to an introspective equilibrium. Therefore, a game-theoretic analysis based on theory of mind is suitable for analyzing strategic behavior in coordination games.

In addition, Theorem 3.1 provides a full characterization of introspective equilibrium for the games that we consider. The first main implication is that for any society and impulse distribution, *the proportion of players who choose the efficient action increases as the risk parameter falls.* This is intuitive: a fall in risk parameter corresponds to an increase in the payoff to the efficient action. The second main implication is that for any society and impulse distribution, *if the risk parameter is sufficiently low, introspective equilibrium selects the efficient Nash equilibrium.* At the other extreme, *if the risk parameter is sufficiently high, introspective equilibrium selects the inefficient Nash equilibrium.* This is intuitive: for extreme values of the risk parameter, payoff considerations trump any other considerations, and all players choose the same action, namely the one that is most attractive in terms of payoffs (as measured by the risk parameter). The third main implication is that *for intermediate values of the risk parameter, behavior is influenced by contextual factors (i.e., impulses) and is not consistent with Nash equilibrium.* To see the intuition, suppose that the risk parameter is $\rho = \frac{1}{2}$. In this case, the payoff structure provides little guidance, and there is scope for impulses to influence behavior. As impulses are positively correlated, a player with an impulse to play action $s = s^0, s^1$ thinks that more than half of the players have an impulse to choose s. Hence, the unique best response at level 1 for the player is to follow his impulse (i.e., to choose action s). So, impulses break the symmetry between the actions. By a simple inductive argument, there is a unique introspective equilibrium, and in this equilibrium, players follow their impulse. Hence, behavior varies with the context: in some situations, most players will choose s, while in others, most players will choose $s' \neq s$, even if the underlying game (i.e., payoff structure) is the same. As impulses are imperfectly correlated, behavior is not consistent with Nash equilibrium: the coordination rate is higher than if players randomize independently, but, unlike in the strict Nash equilibria, coordination is not perfect.[8]

To summarize, for any society, the probability that a player chooses the efficient action increases as the risk parameter falls. While introspective equilibrium selects one of the pure Nash equilibria for extremal values of the risk parameter, for intermediate values of the risk

[8]Behavior is also not consistent with the unstable Nash equilibrium (footnote 7). An easy way to see this is that the introspective process cannot converge to an unstable fixed point.

parameter, it predicts non-Nash behavior. While the exact thresholds between regimes (i.e., $\rho_1, \rho_2, \rho_3, \rho_4$) depend on the specifics of the sociocultural environment, this general pattern obtains for any society. Theorem 3.1 thus delivers testable predictions on how strategic behavior varies with the economic environment. Section 5 compares the predictions to experimental evidence.

Theorem 3.1 has the following corollary:

Corollary 3.2. [Introspective Equilibrium: Generic Uniqueness] *Consider a society* $\mathcal{S} = (\alpha, q, \delta)$ *and a game* (3.1)*. Then, the introspective equilibrium is unique for generic values of the payoff parameters.*[9]

So, for any society, introspective equilibrium is unique for generic values of the payoff parameters. This result follows from the fact that for a given society, the game has multiple introspective equilibria if and only if the payoffs are such that the risk parameter is equal to one of the thresholds in Theorem 3.1 (i.e., $\rho = \rho_1, \rho_2, \rho_3, \rho_4$), and this is nongeneric.

Together with Proposition 2.1, these results imply that introspective equilibrium is a strict refinement of introspective equilibrium.[10] As a refinement of correlated equilibrium, introspective equilibrium depends both on economic factors (i.e., ρ) and on sociocultural factors (i.e., α, q, δ). This is natural: choices are often determined both by economic and by sociocultural factors. For example, as we discuss in Section 4.1, in a society with a strong culture, payoff considerations receive less weight than in a society with a weak culture. While intuitive, this dependence of introspective equilibrium on non-payoff factors makes it difficult to predict behavior for a particular game if the impulse distribution is unobservable.[11] However, even if we do not observe the impulse distribution, the (generic) uniqueness of introspective equilibrium makes it possible to derive testable implications as it yields comparative statics that are independent of the impulse distribution. Moreover, since the impulse distributions can be related to cultural variables, introducing impulses into the model allows us to provide a formal and systematic understanding of how changes in variables that are not directly payoff relevant, such as culture, may become payoff-relevant by influencing how the game is played. The introspective process thus has a dual purpose: to act as an equilibrium refinement, and to

[9] That is, the set of payoff parameters (v_0, v_1, c_0, c_1) for which there is a unique introspective equilibrium has Lebesgue measure 1 in the set of all payoff parameters.

[10] As we demonstrate in Appendix B, this is true even if we consider the set of introspective equilibria across all societies.

[11] In a recent paper, Agranov, Caplin, and Tergiman (2015) develop an experimental protocol that could be used to track players' impulses and their reasoning process. The mechanism keeps track of subjects' provisional choices (i.e., their best decision at any given point in time), thus measuring subjects' instinctive responses and revised choices.

provide a formal model of how culture may shape economic outcomes.[12] We next apply this to study the effect of cultural diversity on economic performance.

3.3 Cultural diversity and economic performance

We compare the economic performance of different types of societies in different economic environments. We measure economic performance by the expected total payoff in introspective equilibrium. Consider a game (3.1) with payoff parameters (v_0, v_1, c_0, c_1) and society $\mathcal{S} = (\alpha, q, \delta)$ and let σ be an introspective equilibrium for the society in this game. Then, the *(expected) total payoff* is

$$\Pi(\sigma; \alpha, q, \delta) := \mathbb{E}_{\alpha,q,\delta}\Big[m(\sigma) \cdot \big(m(\sigma) \cdot v_1 - c_1\big) + \big(1 - m(\sigma)\big) \cdot \big((1 - m(\sigma)) \cdot v_0 - c_0\big)\Big],$$

where $m(\sigma)$ is the proportion $m(\sigma)$ of players who choose s^1 under σ and where the expectation $\mathbb{E}_{\alpha,q,\delta}$ is taken over the impulse distribution (governed by the society's characteristics α, q, δ). The *optimal population composition* α^* is the population composition that maximizes the total payoff in introspective equilibrium.

The following result shows that the optimal population composition varies in a systematic way with the relative attractiveness of the efficient action as measured by its risk parameter ρ:

Theorem 3.3. [Coordination: Optimal Population Composition] *Fix c_0, c_1. Then there exist $\rho^1 = \rho^1(q,\delta), \underline{\rho} = \underline{\rho}(q,\delta), \overline{\rho} = \overline{\rho}(q,\delta)$, and $\rho^0 = \rho^0(q,\delta)$, with $0 < \rho^1 < \underline{\rho} < \frac{1}{2} < \overline{\rho} \leq \rho^0 < 1$, such that:*

(a) If $\rho < \rho^1$, then the total payoff is independent of the population composition: for any population composition, there is a unique introspective equilibrium, and in this introspective equilibrium, all players choose the efficient action s^1.

(b) If $\rho > \rho^0$, then the total payoff is independent of the population composition: for any population composition, there is a unique introspective equilibrium, and in this introspective equilibrium, all players choose the inefficient action s^0.

(c) If $\rho \in (\rho^1, \rho^0)$, then the total payoff depends on the population composition:

[12] In these respects, our approach is similar to that in the literature on global games (Carlsson and van Damme, 1993). This literature selects a unique Nash equilibrium in coordination games using unobservable perturbations of players' higher-order beliefs about payoffs, and has used this uniqueness to derive testable implications from comparative statics. Moreover, by relating unobservable belief perturbations to different types of information, the literature has been able to show how the effect of information on economic outcomes depends not only on the (payoff-relevant) content but also on *how* the information is transmitted (Morris and Shin, 2003).

(c1) If $\rho \in (\underline{\rho}, \overline{\rho})$, then cultural homogeneity is optimal: the optimal population composition is $\alpha^* = 1$.

(c2) If $\rho \in (\rho^1, \underline{\rho}) \cup (\overline{\rho}, \rho^0)$, then cultural diversity is optimal: the optimal population composition is $\alpha^* < 1$.

Theorem 3.3 consists of several parts. The first part (Theorem 3.3(a) and (b)) shows that for a range of payoff parameters, the population composition does not affect economic performance. As we have seen in Theorem 3.1, for extreme values of the risk parameter, the introspective process selects a unique Nash equilibrium: if the risk parameter is close to 0, then the efficient Nash equilibrium is selected, and if the risk parameter is close to 1, then the inefficient Nash equilibrium is selected. The first part of Theorem 3.3 adds to this by showing that for extreme values of the risk parameter, whether or not a unique Nash equilibrium is selected is independent of the population composition.[13] So, *for extreme values of the risk parameter, culturally homogeneous and culturally diverse societies earn the same payoffs.*

In less extreme cases, behavior may depend on sociocultural factors. Theorem 3.3(c) shows that *for intermediate values of the risk parameter, the equilibrium payoff depends on the population composition, and the optimal population composition varies non-monotonically with payoffs.* For lower and higher values of the risk parameter, culturally diversity is optimal, but when the risk parameter lies between these two ranges, culturally homogeneous societies have higher payoffs than culturally diverse ones.

To see the intuition why culturally homogeneous societies have a higher total payoff than culturally diverse ones for intermediate values of the risk parameter (Theorem 3.3(c1)), suppose that the risk parameter is $\rho = \frac{1}{2}$. As we have seen, there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse. In this case, the payoff structure of the game provides little guidance, and players' actions are guided by contextual clues. But since culture shapes players' reaction to the context (i.e., $Q_{in} > Q_{out}$), players are more likely to coordinate when there is a single culture. So, shared cultural perceptions improve behavioral consistency, and a culturally homogeneous society has a higher total payoff than culturally diverse societies. This argument of course extends to any game with risk parameter sufficiently close to $\frac{1}{2}$. So, *when the payoffs provide little guidance, culturally homogeneous societies minimize the risk of miscoordination.*

While it is intuitive that shared cultural perceptions facilitate coordination, it is less clear that cultural diversity can ever be optimal when players are trying to coordinate their actions, at least when groups are identical in payoff-relevant aspects such as skills or access to resources.

[13] In terms of Theorem 3.3, the threshold ρ^1 (resp. ρ^0) is the minimum (resp. maximum) of the threshold $\rho_1(\alpha, q, \delta)$ (resp. of $\rho_4(\alpha, q, \delta)$) over α.

Surprisingly, Theorem 3.3(c2) shows that this is not the case: even though groups are identical in all payoff-relevant aspects in our model, cultural diversity can have an economic benefit. In fact, for a range of payoff parameters, culturally diverse societies are better able to coordinate their actions, not in spite of the greater strategic uncertainty that characterizes these societies, but because of it.

To build intuition, suppose that the risk parameter is less than $\frac{1}{2}$, say, $\rho = 0.3$. In this case, the efficient action s^1 is the unique best response for a player who expects more than 30% of the players to choose s^1. However, choosing the efficient action is not devoid of risk: if the society tilts towards the inefficient action s^0, a player would be better off choosing s^0. Suppose $q = 0.9$. Then, in a culturally homogeneous society (i.e., $\alpha = 1$), there is little strategic uncertainty: a player thinks it is likely that other players have the same impulse as he does (i.e., Q_{in} is close to 1). Given that players have an incentive to coordinate their actions, the unique best response at level 1 for a player is to follow his impulse. The same is true at higher levels. In a culturally homogeneous society, players thus coordinate on the inefficient Nash equilibrium with positive probability.[14]

Now consider a society where some players belong to a different group (i.e., $\alpha < 1$). Then, players from the minority group may face significant strategic uncertainty: their impulse is not very informative of other players' impulses. In the extreme case where the minority is very small and the cultural difference is large, the belief of a player from the minority group is close to the prior for any impulse he might have. So, a player from the minority group expects that roughly half of the players have an impulse to play the efficient action, regardless of his impulse. Since this exceeds the risk parameter (as $\rho < \frac{1}{2}$), the unique best response for a player from the minority group is to choose the efficient action regardless of his impulse. This, in turn, makes it attractive for players from the majority group to choose the efficient action. In particular, if the minority is not too small, then at level 2, the unique best response for players from the majority group is to choose the efficient action regardless of their impulse. Then, in the unique introspective equilibrium, players all choose the efficient action.[15]

So, paradoxically, the lack of congruent expectations that characterizes culturally diverse societies allow players to perfectly coordinate their actions. Moreover, all players choose the efficient action, giving them the highest possible payoff. By contrast, in culturally homogeneous societies, players are locked into the inefficient Nash equilibrium with positive probability. Hence, *if the risk parameter is sufficiently small, cultural diversity reduces the risk of ineffi-*

[14] This is true also for models in which the ex ante probability that players receive an impulse to choose the inefficient action is small, though the welfare effect is obviously smaller.

[15] In terms of Theorem 3.1, for given q, δ, if $\rho \in (\rho^1(q,\delta), \underline{\rho}(q,\delta))$, there exists $\alpha < 1$ such that $\rho < \rho_1(\alpha, q, \delta)$ but $\rho > \rho_1(1, q, \delta)$.

cient lock-in. The optimal population composition trades off two factors: on the one hand, the minority group must be sufficiently small for the minority to face significant strategic uncertainty; on the other hand, it must be sufficiently large to affect the incentives for the players in the majority group.

In effect, impulses anchor the reasoning process. As is well-documented, anchoring can lead to sub-optimal decisions in single-person decision problems (Tversky and Kahneman, 1974). In strategic situations, the distortive effect of anchors is reinforced: a sub-optimal anchor can be locked in if players believe that other players' reasoning process is anchored in a sub-optimal impulse. This source of self-fulfilling expectations can lead players to coordinate on an inefficient action. In culturally diverse societies, the lack of congruent expectations mitigates the power of anchors: if there is significant strategic uncertainty, then payoff considerations trump expectations, and players coordinate on the efficient action.

If the risk parameter is greater than $\frac{1}{2}$, cultural diversity can also be beneficial. However, the intuition is different. If the risk parameter is close to 1, choosing the efficient action is a best response for a player only if he thinks the vast majority of players will choose it. But if a player's impulse is not very informative of the impulses of others, as in culturally diverse societies, the player cannot hold such an extreme belief, and his unique best response is to choose the inefficient action. So, in this case, strategic uncertainty leads players to coordinate on the inefficient action. By contrast, in a culturally homogeneous society, players with an impulse to play an action think that most players have the same impulse and will thus follow their impulse (provided ρ is not too high, i.e., $\rho \in (\overline{\rho}, \rho^0)$). Since impulses are imperfectly correlated, this leads to some miscoordination. There is thus a tradeoff between the risk of miscoordination and the cost of coordinating on the inefficient action. If the loss of not choosing the efficient action is small, then coordinating on the inefficient action yields higher payoffs than tolerating miscoordination, and cultural diversity is optimal. We will come back to this point in Section 4.2.

In sum, cultural diversity is beneficial when the risk parameter is not too close to $\frac{1}{2}$ (so that one action has a clear edge in terms of payoffs) while at the same time not being too close to 0 or 1 (so that this edge is not so strong so as to deny any significant role for contextual clues); and it is costly when the risk parameter is close to $\frac{1}{2}$. The optimal population composition is thus a nonmonotonic function of the risk parameter. This is driven by the fact that players face more strategic uncertainty in a culturally diverse society than in a culturally homogeneous society. Strategic uncertainty leads to lower payoffs when the payoff structure of the game gives little guidance, but can reduce miscoordination in games with significant payoff asymmetries. So, *the same factor – strategic uncertainty – can be a cost or a benefit*, depending on the economic environment. These subtle effects of cultural diversity are driven entirely by strategic

considerations and are therefore above and beyond any direct effect that diversity might have on payoffs.[16]

Theorem 3.3 provides testable implications on the relative performance of different types of societies. While the precise values of the thresholds (i.e., $\rho^1, \rho^0, \underline{\rho}, \overline{\rho}$) depend on (potentially unobservable) sociocultural factors, the same comparative statics obtain for any impulse distribution. Moreover, the results are robust: for example, we can allow for heterogeneity in beliefs (i.e., q) and relax the assumption that each impulse is equally likely a priori. And, while introspective equilibrium is the limiting outcome of a reasoning process of potentially infinitely many steps, our results also go through if players can perform only finitely many levels of reasoning. In particular, all results presented here require at most two levels of reasoning.[17] Thus, our "detail-free" approach to modeling culture provides testable and robust predictions on aggregate patterns of behavior and the resulting economic outcomes.

Beyond the general results provided in Theorem 3.3, the model can also shed light on the subtle interplay between economic and sociocultural factors in various applications. This is the topic we turn to next.

4 Applications

4.1 Social customs

A social custom is an act whose utility depends on the actions of others (Akerlof, 1976, 1980; Dasgupta, 1988). Whether an individual chooses to adhere to a social custom depends on the intrinsic utility associated with the act as well as the actions of others. This means that players have an incentive to coordinate their actions, so that there are generally multiple equilibria (Akerlof, 1980, Prop. I, III). This equilibrium multiplicity can lead to coordination failure: a society can be locked into the Nash equilibrium where everyone obeys the custom even though

[16] There is an interesting parallel between the benefits of diversity and the benefits of reasoning through more steps. If there is little asymmetry in payoffs, then initial beliefs drive behavior. In this case, reasoning does not accomplish much: a player always acts on impulse, no matter how many steps of reasoning he performs. In this case, cultural homogeneity is optimal: if all players act on impulse then highly correlated impulses lead to improved behavioral consistency. By contrast, if there is significant asymmetry in payoffs, then there is more scope for the reasoning process to influence play. Cultural diversity is now valuable because it increases strategic uncertainty and this steers the reasoning process to a more desirable outcome. We thank an anonymous referee for pointing this out.

[17] In games with many actions, this may no longer hold; however, the comparative statics are robust to relaxing the assumption that players can perform infinitely many steps of reasoning. (This holds, for example, for a game in Kets and Sandroni (2015) with a continuum of actions.).

everyone would benefit if everyone would abandon it.[18] Moreover, it implies that societies that are identical in all payoff-relevant aspects can end up experiencing very different outcomes. While this has long been recognized in economics, standard game-theoretic models are silent on what drives equilibrium selection in a given society: Why do some societies abandon the custom and flourish while others continue to obey the custom when these societies face the same economic environment? Here we use the model developed in the previous sections to address these questions.

We consider a simplified (static) version of the model in Akerlof (1980). Players choose whether to obey a code of behavior or custom. Breaking the custom (i.e., disobeying the code) results in a social sanction (e.g., loss of reputation, embarrassment, ostracization). The larger the proportion of the population that subscribes to the code, the greater the social sanction. For example, if more people follow the custom, deviations from the custom tend to be punished more severely (e.g., ostracization by a larger group). So, if the social custom is to play s^0, then the payoff to s^0 is $w_0 > 0$ and the payoff to s^1 is $w_1 - c\mu_0$, where μ_0 is the proportion of the population that obeys the code (i.e., plays s^0) and $w_1, c > 0$ are constants, with c measuring the severity of the social sanction. The custom is disadvantageous (i.e., $w_0 < w_1$) but nevertheless self-enforcing (i.e., $w_0 > w_1 - c$). So, there are two strict Nash equilibria: one in which all players obey the code (i.e., $\mu_0 = 1$) and one in which no player obeys the code (i.e., $\mu_0 = 0$). The Nash equilibrium in which nobody obeys the code is efficient (i.e., $w_1 > w_0$), but potentially risky (i.e., $c > 0$). This is a special case of the model (3.1) if we take $(v_0, v_1, c_0, c_1) = (0, c, -w_0, -(w_1 - c))$. Hence, the risk parameter is

$$\rho = 1 - \frac{(w_1 - w_0)}{c}.$$

The *cost of obeying the custom* is the difference in intrinsic benefits of disobeying and following the code, i.e., $w_1 - w_0$.

We characterize the conditions under which a custom can persist. Say that a custom *persists* (for a given society and risk parameter) if in every introspective equilibrium, players obey the custom with positive probability; if the custom does not persist, then it is *abandoned*. By our previous results, inefficient customs can persist only if the costs of obeying them is not too

[18] Examples of culturally accepted but inefficient practices abound, ranging from everyday matters like holiday gift-giving (Waldfogel, 1993) and the sending of Christmas cards (Schelling, 1978, pp. 31–33) to economically important phenomena such as corruption (Bardhan, 1997), oppressive customs (Akerlof, 1976), codes that preclude transactions at market-clearing prices (Akerlof, 1980), cultures of mutual distrust (Dasgupta, 1988), informal mutual assistance systems that adversely affect incentives to enter more productive occupations (Hoff and Sen, 2006), increased violence associated with a culture of honor (Nisbett and Cohen, 1996), and excessive spending by poor households in developing countries on lavish weddings and funerals (Banerjee and Duflo, 2011).

high (i.e., ρ is not too small). More precisely, Theorem 3.1 shows that a custom persists if and only if $\rho > \rho_1(\alpha, q, \delta)$. We refer to $\rho_1(\alpha, q, \delta)$ as the *persistence threshold*.

Proposition 4.1. [Persistence of Inefficient Customs: Payoffs] *Societies gain from abandoning an inefficient custom: for any society* (α, q, δ),

$$\Pi(\sigma^{\downarrow\rho_1(\alpha,q,\delta)}) - \Pi(\sigma^{\uparrow\rho_1(\alpha,q,\delta)}) \geq \frac{w_1 - w_0}{2q^2 + 2 \cdot (1-q)^2}, \tag{4.1}$$

where $\Pi(\sigma^{\uparrow\rho_1(\alpha,q,\delta)})$ *(resp.* $\Pi(\sigma^{\uparrow\rho_1(\alpha,q,\delta)})$*) is the total expected payoff as the risk parameter* ρ *approaches the persistence threshold* $\rho_1(\alpha, q, \delta)$ *from above (resp. below).*[19]

Proposition 4.1 shows that coordination failure can substantial: When the custom is abandoned, there is a discontinuous increase in the total payoff.[20] The inefficiency associated with obeying the custom is directly proportional to the cost $w_1 - w_0$ of obeying the custom . Moreover, it is lower when strategic uncertainty is limited (i.e., q close to 1), reflecting the greater behavioral consistency associated with strong cultures.

Whether or not a society abandons a custom depends not only on economic factors (i.e., ρ) but also on sociocultural factors. The following result makes this dependence precise. The result follows from the proof of Theorem 3.1, so we refer to it as a corollary.

Corollary 4.2. [Persistence of Inefficient Customs: Culture] *Inefficient customs persist for a larger range of payoff parameters if groups have a strong culture or are culturally close. That is, for any population composition* α*, the persistence threshold* ρ_1 *decreases with* q *and increases with* δ*.*

Corollary 4.2 shows that economic factors and cultural factors jointly determine whether a custom persists or is abandoned: a custom can persist if it is not too costly (i.e., ρ sufficiently high) and groups have a weak culture and are not culturally close (i.e., q close to $\frac{1}{2}$ and δ small). The intuition is similar to before. When groups have a strong culture or are culturally close, players face little strategic uncertainty. Consequently, they have an incentive to choose the inefficient action if they think other players will choose it. So, cultural expectations become self-fulfilling. By contrast, a lack of congruent expectations makes it possible to abandon the custom. An implication of this result is that for a given economic environment (i.e., ρ), a society with a strong culture may obey the custom even if a society with a weak culture that

[19] The total expected payoff near the persistence threshold (i.e., $\Pi(\sigma^{\downarrow\rho_1(\alpha,q,\delta)})$ and $\Pi(\sigma^{\uparrow\rho_1(\alpha,q,\delta)})$) is well-defined since the introspective equilibrium is unique near the persistence threshold. See the proof.

[20] The discontinuity is not an artifact of our assumption that players are ex ante identical in terms of beliefs or payoffs; the discontinuity is preserved in richer models that allow for heterogeneity in payoffs or belief parameters.

is otherwise similar has abandoned it. The idea that societies that are identical in all (payoff relevant) respects may end up with different outcomes because they play according to different equilibria is of course not new. However, existing accounts have not been able to characterize the economic and sociocultural conditions that lead a society to adopt one equilibrium versus another, as they abstract away from sociocultural factors; moreover, they generally predict that both persistence and abandonment are equilibria and thus do not yield sharp predictions that can be tested.

We next turn to the question how cultural diversity affects the persistence of inefficient customs. Theorem 3.1 has the following interesting implication: for a range of payoff parameters, two societies that are identical in all respects except the population composition may react very differently to a small change in economic fundamentals:

Corollary 4.3. [Cultural Diversity and Divergence] *Fix q, δ. There exist a nonempty open interval T for the risk parameter and a function $\underline{a} : T \to (\frac{1}{2}, 1)$ such that for $\rho \in T$, the custom persists in a society (α, q, δ) if and only if $\alpha > \underline{a}(\rho)$. Moreover, for $\rho \in T$, if the custom persists in two societies $\mathcal{S} = (\alpha, q, \delta)$, $\mathcal{S}' = (\alpha', q, \delta)$ and $\alpha > \alpha'$, then there is $\rho' < \rho$ such that if the risk parameter decreases to ρ', the custom is abandoned in $\mathcal{S}'$ yet persists in $\mathcal{S}$.*

The first claim states that for a range of payoff parameters, an inefficient custom persists if and only if the society is sufficiently homogeneous. The second claim concerns changes in the payoff parameters. When the cost $w_1 - w_0$ of obeying the custom increases, the risk parameter falls. So, Corollary 4.3 shows that under certain conditions, *as the cost of obeying the custom increases, a culturally diverse society may cease to follow the code while a culturally homogeneous society may continue to obey it.* Hence, the economic performance of societies may diverge when economic fundamentals change even if the societies are identical in all payoff-relevant aspects. Given that a small change in economic fundamentals may lead to a large change payoffs (cf. Eq. (4.1)), the effects can be substantial. This prediction holds for any impulse distribution and can thus be tested without knowing the cultural parameters.

To illustrate the result, consider Figure 1. Figure 1 plots the total payoff as a function of the cost of obeying the custom, for a culturally homogeneous (dashed green line) and a culturally diverse society (solid blue line). The former does well when the cost of obeying the custom is low; but as this cost increases, its payoff decreases gradually. In contrast, a culturally diverse society sees a discontinuous jump in payoffs as it switches from a regime where it follows the code to one where the code is abandoned. Thus, close-knit societies perform at least as well as pluralistic societies when the cost of following outdated practices is low but do worse as the cost increases. This is intuitive, but difficult to formalize with standard game-theoretic models as these models do not take into account sociocultural factors. In particular, standard

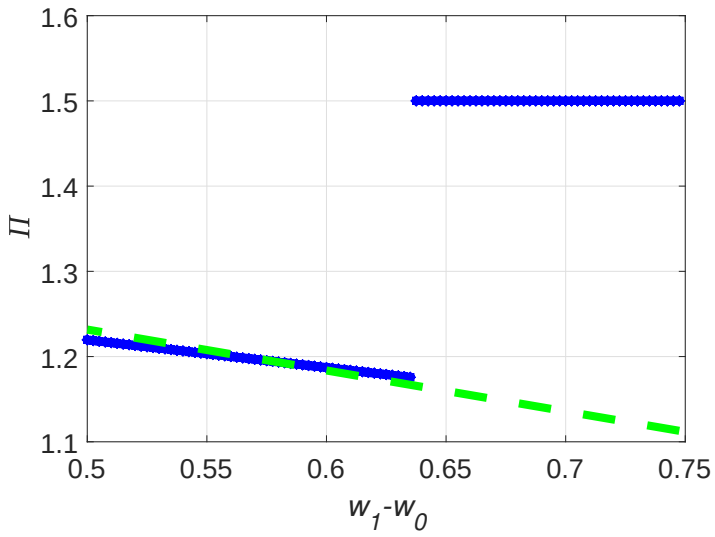

Figure 1: The total payoff as a function of $w_1 - w_0$ for societies with $\alpha = 0.65$ (solid blue line) and with $\alpha = 0.95$ (dashed green line) for $w_1 = 1.5, c = 1, q = 0.99, \delta = 0.995$, and w_0 ranging from 0.5 to 0.75.

models of social customs have multiple equilibria and are silent on the conditions that drive equilibrium selection. Modeling players' reasoning process and their reactions to contextual clues thus has the potential to shed light on the drivers of equilibrium transitions and economic divergence.

4.2 Collective action problems

Many economic problems have both a coordination and a cooperation component. Consider, for example, the case of urban decay. The value of a dwelling depends for an important part on the quality of the neighborhood. In turn, the quality of a neighborhood depends on residents investing effort to improve their neighborhood, e.g., by cleaning up litter and planting flower beds. There is a free-riding problem: all residents benefit from investment in the neighborhood, but investing may be individually costly. In addition, collective action problems can have a coordination component: If players who free-ride incur a social sanction that increases with the proportion of players who invest, then cooperation may be viable but only if sufficiently many other players invest. In the case of investing in neighborhood quality, social sanctions take the form of informal social controls.[21] Here we use the model developed in Section 3 to

[21] In the words of the eminent writer and urban activist Jane Jacobs (1961), "[Order] is kept primarily by an intricate, almost unconscious, network of voluntary controls and standards among people and enforced by the people themselves" (p. 40).

shed light on the tradeoff between coordination and incentive problems.

A player can choose whether to invest (i.e., play s^1) or to free-ride (i.e., play s^0). The payoff to investing increases with the proportion of players who invest: If a proportion m of players invest, then the payoff to a player who invests is

$$mv_1 - c_1,$$

where $c_1 > 0$ is the effort cost. Players who free-ride derive a benefit from the investment of other players; but, in addition, they do incur a social sanction. Thus, if a proportion m of players invest, then the payoff to a player who does not invest is

$$mv_1 - mC,$$

where $C > 0$. Since residents benefit from others' investments regardless of whether they invest themselves (i.e., all players get mv_1, regardless of their action), there is an incentive problem. In addition, there is a coordination problem: thanks to the social sanction, players have an incentive to invest if they expect a large proportion of players to invest. Indeed, if we take $v_0 = -(v_1 - C)$ and $c_0 = -(v_1 - C)$, then this fits the model (3.1) in Section 3, with risk parameter

$$\rho = \frac{c_1}{C}.$$

If $c_1 \in (0, C)$, then, by (3.2), the game has two strict Nash equilibria: one in which all players invest effort, and one in which no player invests effort.

We ask whether culturally homogeneous or culturally diverse societies are better at solving collective action problems. We focus on the case where the effort cost c_1 is high relative to the social sanction C (i.e., $\rho > \frac{1}{2}$) but not too high to render population composition irrelevant (i.e., $\rho < \rho^0$). In this case, cultural diversity is beneficial if and only if the culture is weak:

Corollary 4.4. [Collective Action] *Assume that $\rho \in (\frac{1}{2}, \rho^0)$. Then, cultural diversity is optimal (i.e., $\alpha^* < 1$) whenever*

$$Q_{in} < 1 - \frac{v_1 - c_1}{C}; \tag{4.2}$$

otherwise, cultural homogeneity is optimal (i.e., $\alpha^* = 1$).

The proof of Corollary 4.4 follows directly from the proof of Theorem 3.3 and is thus omitted. The result says that if the effort cost is relatively high (i.e., $\rho > \frac{1}{2}$), then cultural diversity is optimal if there is a weak culture (i.e., Q_{in} small) and the benefit of coordinating on the efficient action is small compared to the social sanction (i.e., $\frac{v_1 - c_1}{C}$ small).

The intuition is as follows. If the risk parameter is high (i.e., $\rho > \frac{1}{2}$), free-riding is relatively attractive and so cannot be eliminated entirely in equilibrium (Theorem 3.1). Hence, any

reduction in free-riding comes at the expense of an increase in miscoordination. The optimal population composition thus has to trade off two types of problems: either it can prioritize coordination or it can reduce free-riding.

To analyze the optimal population composition, we compare equilibrium behavior in different societies. In a culturally diverse society, the impulse of a player from the minority group is not very informative of the impulses of other players. So, even if a player from the minority group has an impulse to invest (i.e., play s^1), he expects only a small proportion of players to have an impulse to invest. Since the risk parameter is high (i.e., $\rho > \frac{1}{2}$), investing is risky. Hence, at level 1, it is optimal for a player from the minority group to not invest (i.e., play s^0), regardless of his impulse. This, in turn, leads players from the majority group to not invest at level 2. Hence, in a culturally diverse society, no player invests in introspective equilibrium. This means foregoing the benefit of investing, but it ensures successful coordination.

By contrast, in a culturally homogeneous society, a player's impulse is informative of other players' impulses. Since players have an incentive to coordinate their actions, it is optimal for players to follow their impulse to match other players' actions. As impulses are imperfectly correlated, it yields some miscoordination; however, it also gives positive investment.

The tradeoff between the cost of miscoordination and free-riding depends on culture strength. If the culture is weak, then the coordination problems are so severe that it is better to tolerate free-riding rather than pay the cost associated with miscoordination. In this case, culturally diverse societies have a higher total payoff than culturally homogeneous ones, even if there is more investment in the latter. Conversely, if there is a strong culture, then the cost associated with miscoordination is small, and it is optimal to have positive investment even if it comes at the expense of less coordination. In this case, culturally homogeneous societies have a higher total payoff than culturally diverse societies.

For given culture strength q (and thus Q_{in}), the optimal tradeoff between free riding and miscoordination depends on the payoff parameters. Naturally, if the benefit $v_1 - c_1$ of coordinating on the efficient action increases, cultural homogeneity is optimal for a larger range of parameters (i.e., the right-hand side of (4.2) falls). Intuitively, if the benefits associated with full investment (viz., $v_1 - c_1$) increase, then increasing investment takes greater priority over limiting miscoordination. Since culturally homogeneous societies have a higher investment rate and less free-riding than culturally diverse societies, the payoffs to the former increase relative to the latter's.

Perhaps surprisingly, a decrease in the cost C of social sanctions also increases the range of parameters for which it is optimal to reduce free-riding (i.e., the right-hand side of (4.2) falls). The decrease in the cost of social sanctions has two effects. The first effect is direct: a decrease in the cost of social sanctions increases the payoff to not investing. The second effect

is strategic in nature. From the perspective of society, a social sanction not only provides an incentive to invest, it also acts as a miscoordination cost: a social cost is incurred whenever some players invest and some players do not. When the cost of social sanctions falls, the cost of miscoordination decreases, making a "mixed" state with partial (but not full) investment viable. As culturally homogeneous societies have more investment than culturally diverse societies, this strategic effect makes the former more attractive relative to the latter.

Thus, even if investment is costly (i.e., $\rho > \frac{1}{2}$), some societies can sustain positive investment. Positive investment is easier to sustain in equilibrium if the society is culturally homogeneous, as *shared cultural perceptions allow players to coordinate their actions so that no effort is wasted*. This suggests the possibility that at least some of the problems associated with battling collective action problems are not due to incentive problems. Rather, the problem may be one of coordination. If that is the case, policies designed to alleviate incentive problems may not be effective. For example, if the cost C of social sanctions is high (e.g., thanks to effective monitoring), then a culturally diverse society has a higher total payoff than a culturally homogeneous society. However, there is no investment in equilibrium and each player earns 0. Instead of increasing the cost of social sanctions, a culturally diverse society would be better of if the coordination among members of different groups were to be improved. For example, in the case of urban decay, a local government could act to coordinate community efforts. This improved coordination would lead to less wasted effort, increase the incentives to invest, and make it possible to sustain positive investment in equilibrium.

4.3 Quota

Our analysis thus far has focused on the optimal population composition, with an eye to comparing the economic performance of societies in different environments. Another important question concerns the effect of small changes to the population composition. Consider, for example, an organization that is culturally homogeneous. How would policies that aim to increase minority representation by a little bit influence economic performance (i.e., total payoffs)?

We use the modeled developed in Section 3 to study this issue. We view an organization as a society and we write (β, q, δ) for an organization (α, q, δ) for which a proportion $\beta = 1 - \alpha$ of players belongs to the minority group. The following corollary of Theorem 3.3 shows that if an organization has only a small minority, then it may be worse off than if it had none at all:

Corollary 4.5. [Effect of Small Minorities] *Culturally homogeneous organizations can have a higher total payoff than organizations with a small minority. Let $\rho \in (\rho^0, \rho^1)$ and suppose that if $\rho < 1 - Q_{out}$, then $2v_1 \cdot (1 - Q_{in}) - c_1 > 2v_0 Q_{in} - c_0$. Then there is $\varepsilon > 0$ such that an*

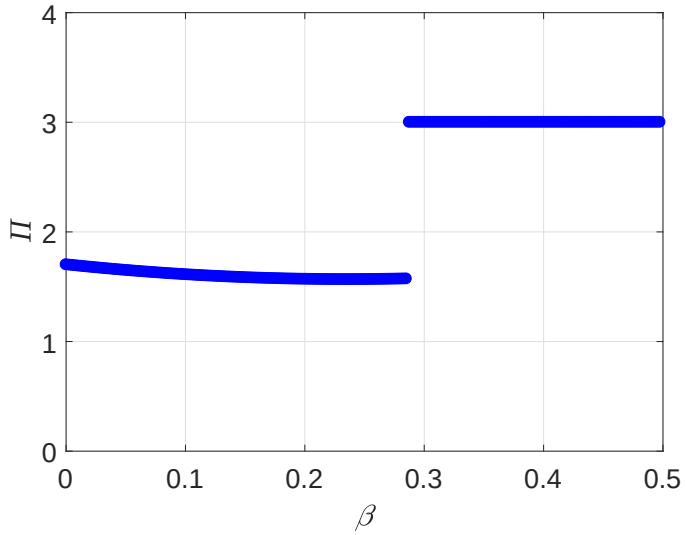

Figure 2: The total payoff as a function of the minority share β (for parameters $v_1 = 4, v_0 = 1, c_1 = 1, c_0 = 0.5, q = 0.99, \delta = 0.995$).

organization (β, q, δ) *with* $\beta = 0$ *has a strictly higher total payoff than a society* (β', q, δ) *with* $\beta' \in (0, \varepsilon)$.

Together with Theorem 3.3(c), Corollary 4.5 shows that the effects of cultural diversity can be nonmonotonic: organizations can gain from having a sizeable minority (Theorem 3.3(c)), but having a small minority may be worse than having none at all. This is illustrated in Figure 2. Intuitively, cultural diversity is beneficial only if it reduces miscoordination. As we have seen, cultural diversity can reduce miscoordination if two conditions are met: (1) strategic uncertainty leads the minority to choose a fixed action s and (2) the minority is sufficiently large so that this gives the entire society an incentive to coordinate on s. Since the behavior of a token minority has a negligible impact on the incentives of other players, the latter condition fails to hold if the minority is small. In this case, the behavior of the minority and the majority is poorly coordinated, and cultural diversity increases the risk of miscoordination rather than mitigating it. Only when the minority reaches a critical mass does it influence the incentives for other players and can diversity facilitate coordination.

The potential nonmonotonic relation between cultural diversity and its benefits has potentially important implications for the design of diversity policies. To give an example, the present results suggest that there is no economic rationale for instituting a low quota, at least in the present static setting. An organization that sets a low quota (i.e., β close to 0) earns a lower payoff than a culturally homogeneous organization (i.e., $\beta = 0$) even if it could earn a substantially higher total payoff if it had a sizeable minority. So, quota may not have eco-

nomic benefits and can even be costly unless the quota is sufficiently high and there are enough suitable candidates from the minority group. This is consistent with empirical evidence. The empirical literature on quota generally has failed to find a significant positive relation between quota and firm performance and has suggested that this may be due to the minority group not having reached a critical mass (e.g., Chapple and Humphrey, 2014, and references therein). Another implication is that if there are only few employees who belong to a minority group, the firm may benefit by concentrating these employees in a single division until the group reaches a critical mass. It is not clear how to derive similar implications in a traditional economic model. For example, if the benefits of diversity are information- or skill-based, as is often assumed, one would expect that a small minority already brings significant economic benefits. By contrast, if cultural diversity affects rate of miscoordination, then cultural diversity is beneficial only if the minority can reach a critical mass.

4.4 Organizational culture

We can apply the model to shed light on the effects of corporate culture on firm performance. For simplicity, we restrict attention to organizations that are culturally homogeneous (i.e., $\alpha = 1$) throughout this section. The following result characterizes the optimal organizational culture for different economic environments:

Proposition 4.6. [Organizational culture] *Fix δ and $q' > q$. Consider two organizations $\mathcal{S} = (\alpha, q, \delta)$ and $\mathcal{S}' = (\alpha, q', \delta)$ with $\alpha = 1$, and write Q_{in} (resp. Q'_{in}) for $Q_{in}(q) = q^2 + (1-q)^2$ (resp. $Q_{in}(q') = (q')^2 + (1-q')^2$). Then:*

(a) If $\rho < 1 - Q'_{in}$, then the total payoff is independent of culture strength: for both $\mathcal{S}$ and $\mathcal{S}'$, there is a unique introspective equilibrium, and in this introspective equilibrium, all players choose the efficient action s^1.

(b) If $\rho > Q'_{in}$, then the total payoff is independent of culture strength: for both $\mathcal{S}$ and $\mathcal{S}'$, there is a unique introspective equilibrium, and in this introspective equilibrium, all players choose the inefficient action s^0.

(c) If $\rho \in (1 - Q'_{in}, Q'_{in})$, then the total payoff depends on culture strength:

(c1) If $\rho \in (1 - Q'_{in}, 1 - Q_{in})$, then a weak culture is optimal: the total payoff in $\mathcal{S}$ is greater than the total payoff in $\mathcal{S}'$.

(c2) If $\rho \in (1 - Q_{in}, Q_{in})$, then a strong culture is optimal: the total payoff in $\mathcal{S}'$ is greater than the total payoff in $\mathcal{S}$.

(c3) *If $\rho \in (Q_{in}, Q'_{in})$, then a weak culture is optimal if and only if q' is sufficiently small, i.e., $Q'_{in} < \frac{2v_0 - c_0 + c_1}{v_1 + v_0}$.*

Proposition 4.6 mirrors Theorem 3.3, illustrating that the effect of a weak organizational culture (resp. strong organizational culture) is not dissimilar to that of cultural diversity (resp. homogeneity). While the boundaries between the different regimes depend on (potentially unobservable) cultural factors, the same basic pattern emerges across societies so that the comparative static results on payoffs can be tested without observing the organization's culture. Again, we have a non-monotonic relation: a weak organizational culture is optimal for high and low values of the risk parameter while a strong organizational culture is optimal for intermediate values of the risk parameter. Moreover, for extreme values of the risk parameter (i.e., ρ close to 0 or 1), economic performance is independent of cultural factors: if the payoff to an action is sufficiently high, then all players choose it in introspective equilibrium, independent of any cultural influence. For intermediate values of the risk parameter, cultural factors play a role. If the payoff structure of the game provides little guidance (cf. (c2)), then it is optimal to have a strong organizational culture. The intuition is the same as before. If the actions are (nearly) symmetric in terms of payoffs, then choices must be guided by situational cues. If there is a strong culture, players are inclined to respond to a given situational cue in the same way. Thus, a strong culture facilitates coordination. On the other hand, if the risk parameter is low (cf. (c1)), then a lack of congruent expectations allows an organization to avoid inefficient lock-in and a weak culture is optimal. Finally, if the risk parameter is high (cf. (c3)), then whether or not a weak culture is optimal depends on culture strength. As we have seen earlier, for high values of the risk parameter, there is a tradeoff between reducing the rate of miscoordination and avoiding inefficient lock-in, and which type of culture is optimal depends on the relative costs of each.

Proposition 4.6 sheds light on how organizations can improve their performance in case their culture is a poor fit with the economic environment. For example, an organization with a strong culture that is locked into an inefficient Nash equilibrium could benefit from starting fresh at a "greenfield" site, and this is indeed common practice among organizations if changing practices is too costly (Brynjolfsson and Milgrom, 2013). On the other hand, an organization with a weak culture that experiences frequent mishaps and misunderstandings could choose to use simple rules of action to improve coordination even if that means that behavior is not well adapted to the circumstances (Dessein and Santos, 2006).

There is a tight connection between organizational culture and leadership. One role for a leader is to be a coordinator, that is, to announce the equilibrium to be played (Hermalin,

2013).[22] Outside the realm of formal game theory, such announcements rarely take the form "coordinate on action s." Rather, they take the form of prescriptions (e.g., "make it a priority to meet the demands of the biggest customers") that, while (reasonably) specific, can be interpreted in different ways (prioritize at what cost?). So, messages can be ambiguous. Consider a leader who communicates with two subordinates. The leader observes an underlying state $\theta = s^0, s^1$ drawn from a common prior with full support. He sends a public message to both players to inform them of the state.[23] After receiving the message, players play a coordination game with payoffs given by

	s^1	s^0
s^1	V, V	$0, 0$
s^0	$0, 0$	$1, 1$

where $V \geq 1$. This is a special case of the model (3.1) with risk parameter $\rho = \frac{v}{V+v}$ (assuming players are randomly matched in pairs). Suppose that players take the message at face value: a player's impulses is his idiosyncratic interpretation of the leader's message. Assume that conditional on $\theta = s$, each player's interpretation of the message is s with probability $q > \frac{1}{2}$, independently across players. So, player's impulses are correlated. By choosing his communication style, the leader can influence q. Then, Theorem 3.1 suggests that the leader's optimal communication style depends on the payoff V. If $V = 1$, then it does not matter which Nash equilibrium players coordinate on; rather, the priority is to avoid miscoordination. In that case, it is optimal for the leader to send an unambiguous message to improve behavioral consistency (i.e., $q \to 1$). On the other hand, if V is greater than 1, then, by choosing a message that is sufficiently ambiguous (i.e., q close to $\frac{1}{2}$, a leader can ensure that players coordinate on the efficient Nash equilibrium. Thus, depending on the economic environment, it may be optimal for leaders to use ambiguous messages.

5 Introspection and culture revisited

5.1 Introspection

We return to the predictions of introspective equilibrium to compare them with experimental evidence. The *first main prediction* of our model is that players may choose the inefficient action even if they would benefit from coordinating on the efficient action. This prediction has

[22] We thank an anonymous referee for suggesting this application.

[23] We assume that the leader commits to reporting the state truthfully so as to focus exclusively on the role of leadership in facilitating coordination.

received ample experimental support for a range of coordination games (see, e.g., Van Huyck, Battalio, and Beil, 1990; Cooper, DeJong, Forsythe, and Ross, 1990, 1992; Straub, 1995). The intuition behind the theoretical result is that, at low levels of reasoning, introspective players allow for the possibility that other players act on impulse even if that is not in those players' self interest. As a result, individual incentives reflect not only the benefits of successful coordination but also the risk of miscoordinating, and this may lead players to choose the inefficient action. This feature is partly familiar from existing solution concepts which involve some form of perturbations or trembles (such as global games (Carlsson and van Damme, 1993), quantal response equilibrium (McKelvey and Palfrey, 1995), and the tracing procedure (Harsanyi and Selten, 1988)) or that are derived from an evolutionary or learning process (e.g., Fudenberg and Levine, 1999). However, a novel prediction not captured by existing models is that the gap between individual incentives and socially optimal behavior is smaller when there is more strategic uncertainty in the sense that societies that experience more strategic uncertainty can avoid inefficient lock-in for a larger range of payoff parameters. Moreover, the degree of strategic uncertainty is determined by sociocultural factors, rather than by mistakes or trembles. This opens up the possibility to endogenize the strategic uncertainty, as discussed in Section 6.

The *second main prediction* of our model is that coordination is successful only if there is sufficient asymmetry between the actions in terms of payoffs (i.e., ρ close to 0 or 1). If there is limited asymmetry between the actions in terms of payoffs, both actions are chosen with positive probability and behavior is not consistent with Nash equilibrium. These predictions are strongly supported by the data. We focus much of the discussion on symmetric (2×2) coordination games, which are a special case of our model.[24] For these games, Straub (1995) and Schmidt, Shupp, Walker, and Ostrom (2003), among many others, show that for intermediate values of the risk parameter, behavior is not consistent with Nash equilibrium: players coordinate at a higher rate than in mixed Nash equilibrium, but at a lower rate than in pure Nash equilibrium.[25] Existing equilibrium selection methods cannot account for these findings. Payoff dominance selects the same Nash equilibrium independent of the risk parameter, as do team reasoning theories (Sugden, 1993). Risk dominance makes the stark prediction that play-

[24] If we write u_{nm} for the payoff to a player who chooses s^n while the other player chooses s^m, then if players are drawn randomly from a large population and matched in pairs to play the game, their expected payoff is given by (3.1) with payoff parameters $(v_0, v_1, c_0, c_1) = (u_{00} - u_{01}, u_{11} - u_{10}, -u_{01}, -u_{10})$ and risk parameter $\rho = \frac{u_{00}-u_{10}}{u_{11}+u_{00}-u_{01}-u_{10}}$.

[25] We are not aware of any experimental studies that study games with extreme values for the risk parameter. This could be a selection effect: if the interest is in testing competing hypotheses, there is no reason to select games for which there is an obvious way to play so that all theories make the same prediction; see Schmidt, Shupp, Walker, and Ostrom (2003, p. 285) for a remark along these lines.

ers coordinate on the efficient action (with probability 1) whenever the risk parameter is less than $\frac{1}{2}$, while they coordinate on the inefficient action whenever the risk parameter is greater than $\frac{1}{2}$. So, risk dominance cannot explain that coordination is only partially successful when there is limited asymmetry among the actions; a fortiori, it cannot explain the observed non-Nash behavior. Since the risk-dominant Nash equilibrium is selected by global games methods (Carlsson and van Damme, 1993), evolutionary models (Young, 1993; Kandori, Mailath, and Rob, 1993), and quantal response equilibrium (McKelvey and Palfrey, 1995), these methods cannot explain the observed behavior either.[26] This also holds for other concepts. Most notably, Crawford and Haller (1990), in their study of how players use asymmetries in the game to coordinate, derive the stark prediction that players coordinate (with probability 1) whenever there is *some* asymmetry between actions, no matter how small. By predicting that coordination succeeds only if there is sufficient asymmetry between the actions, our model provides a more nuanced and arguably more realistic view than existing concepts.

The *third main prediction* of our theory is that for intermediate values of the risk parameter (i.e., ρ close to $\frac{1}{2}$), behavior is strongly influenced by contextual factors, and behavioral consistency improves when strategic uncertainty is reduced. Experimental support for the influence of contextual factors comes from a variety of sources. First, there is extensive evidence that past experience influences strategic behavior even when there are no incentives to build reputation or signal intentions (e.g., Schmidt, Shupp, Walker, and Ostrom, 2003). To the extent that history shapes impulses, this is consistent with our results.[27]

A second type of evidence for this prediction involves the saliency of alternatives. Arguably, when the payoff structure of the game provides little guidance (i.e., ρ close to $\frac{1}{2}$) and one action is more salient than the others, players have an inclination to select the salient alternative (e.g., Mehta, Starmer, and Sugden, 1994, p. 659). If that is the case, then a greater asymmetry in salience between the actions corresponds to a higher q in our model. Then, the model predicts that the coordination rate is higher for games where one alternative is significantly more salient than the others (Proposition 4.6(c2)). Experiments on pure coordination games[28] provide support for this hypothesis: In these games, subjects are often remarkably successful at coordinating even if they can choose from a large number of alternatives (Schelling, 1960; Mehta, Starmer, and Sugden, 1994; Bardsley, Mehta, Starmer, and Sugden, 2010), suggesting that impulses are correlated.

[26] The noisy introspection model of Goeree and Holt (2004) predicts non-Nash behavior in at least some coordination games. However, it is unclear how predictions vary with payoffs and thus whether the model can reproduce the observed comparative statics.

[27] Of course, to model how the history of play affects impulses, a dynamic model is needed; see Section 6.

[28] In a *pure coordination game*, players earn 1 if they choose the same alternative and 0 otherwise (so that $\rho = \frac{1}{2}$ if there are two actions).

A third source of evidence that contextual clues influence behavior comes from individual variation in perspective-taking ability. An individual with superior perspective-taking abilities presumably has a highly informative signal about other players' impulses and will thus be better at coordinating. Curry and Jones Chesters (2012) show that in the pure coordination games of Mehta, Starmer, and Sugden (1994), subjects with superior perspective-taking ability (as measured by a self-report questionnaire) have a higher probability of coordinating when matched against the population, consistent with our theory.

Together, this evidence suggests that when there is limited asymmetry between the actions in terms of payoffs, strategic behavior is responsive to contextual clues and stronger clues improve coordination. This is consistent with our theory. At the same time, it is difficult to capture using existing approaches, as existing models do not account for the influence of contextual factors on play.[29]

5.2 Cultural differences

We next turn to the predictions on how cultural differences affect behavior. In our model, players find it easier to anticipate the impulses of members of their own group. Moreover, players from the same group are more likely to have the same impulse. One implication of these assumptions is that when there is limited asymmetry in actions in terms of payoffs, culturally homogeneous groups are more successful at coordinating than culturally diverse groups (Theorem 3.3(c1)). This is consistent with experimental evidence. Weber and Camerer (2003) present an experiment where groups of subjects are allowed to develop a common "culture" while playing a pure coordination game. They show that culturally homogeneous groups have a higher coordination rate than culturally diverse groups, consistent with our theory. Jackson and Xing (2014) contrast the behavior of subjects residing in India versus the U.S. in a battle-of-the-sexes game. They find that subjects are better able to predict how subjects of their own group would play. Moreover, the two groups differ in the actions that they take. To the extent that actions are a function of impulses, these findings support our assumption that players from the same group are more likely to have the same impulse and that players find it easier to anticipate the impulses of members of their own group. Consistent with our predictions, Jackson and Xing find that subjects are more successful at coordinating when they are matched with a member of their own group. To the best of our knowledge, there are no other models that can explain the experimental results of Weber and Camerer (2003)

[29] An exception is the literature on the determinants of salience. For example, Bacharach and Bernasconi (1997) develop and test a theory of which alternative (among a set of options) is salient. However, this literature is not concerned with how the coordination rate varies with the relative salience of the alternatives.

and Jackson and Xing (2014).

6 Discussion

General games While introspective equilibrium provides intuitive predictions in the games we study, we cannot guarantee existence for all games. Proving existence can be challenging because introspective equilibrium is more than a consistency requirement. It is also the limit of a specific reasoning process. Hence, showing existence requires more than a fixed point theorem.[30] So, as for other equilibrium refinements that are derived from some type of process, such as global games or learning models, proving existence requires making use of a game's structure. We expect that the greatest progress can be made for classes of games where other, similar processes have been shown to converge, such as games with strategic complementarities and potential games.

Coevolution of culture and behavior In our analysis, we have taken the distribution of impulses as given. In practice, we would expect players' impulses to be influenced by their past experiences. Indeed, shared experiences presumably mediate the effect of culture on beliefs. One important question is in which economic environments heterogeneity in impulses can persist. The coevolution of culture and behavior is an important area for further study. We leave this for future work.

Dysfunctional cultures Thus far, we have assumed that groups are ex ante identical in terms of their impulses. However, some groups may have a dysfunctional culture in the sense that the members of the group are likely to have an impulse to choose the inefficient action.[31] We can generalize the model to accommodate this. Suppose group $\gamma_D = A, B$ has a dysfunctional culture in the sense have that its members have an impulse to play the efficient action with probability $p \in (0, \frac{1}{2})$. Group $\gamma_S \neq \gamma_D$ has a superior culture: its members have an impulse to play the efficient action with probability $\frac{1}{2}$, as before. Then, the joint distribution of impulses is described by

[30] Indeed, like learning processes, the introspective process may fail to converge. To see this, suppose there are two players who receive a payoff of 1 if they choose different actions and receive 0 otherwise. Then, a player who thinks the other player is likely to choose s^1 should choose s^0 and vice versa. Hence, the introspective process will cycle indefinitely even though the best response correspondence has a fixed point. For a discussion of related issues in the context of learning models, see Fudenberg and Levine (1999).

[31] We are grateful to an anonymous referee for suggesting this application.

	$\theta_D = s^1$	$\theta_D = s^0$
$\theta_S = s^1$	$\frac{1}{2}p + \eta$	$\frac{1}{2} \cdot (1-p) - \eta$
$\theta_S = s^0$	$\frac{1}{2}p - \eta$	$\frac{1}{2} \cdot (1-p) + \eta$

where $\eta < \frac{1}{2}p$ measures the cultural closeness of the two groups, as before. Our results extend to this case. In particular, there is a range of payoff parameters such that cultural diversity is optimal. The intuition is the same as before: If the risk parameter is low but not too low, a culturally diverse society can avoid inefficient lock-in. This has the important implication that at least in some economic environments, a group with a superior culture can benefit from integrating with a group with a dysfunctional culture. We leave a full exploration of these issues for future work.

7 Related literature

This paper contributes to the literature in behavioral economics by developing a new methodology to study the effects of culture. Our model is related to the models studied in the literature on level-k reasoning; see Nagel (1995), Stahl and Wilson (1995), Costa-Gomes, Crawford, and Broseta (2001), and Costa-Gomes and Crawford (2006) for influential early papers, and see Crawford, Costa-Gomes, and Iriberri (2013) for a recent survey. It distinguishes itself from level-k models in two main ways. First, introspective equilibrium can explain why a unique Nash equilibrium is selected when there is a strong asymmetry between the actions while behavior is not consistent with Nash equilibrium when the payoff structure gives little guidance. As we discussed in Section 5, this is critical for explaining experimental evidence. Second, by introducing correlation in impulses, we can model the effects of culture in a natural way.

Among equilibrium selection methods, our model is most closely related to best-reply learning models. The predictions we obtain are markedly different, however. While most learning models select the risk-dominant Nash equilibrium, our introspective process may not select a Nash equilibrium even if one of the equilibria is both payoff- and risk-dominant. More generally, our model predicts that coordination is successful only if there is sufficient asymmetry among the actions. This is more in line with the experimental evidence than the stark predictions of equilibrium refinements, which predict that coordination succeeds with probability 1 whenever there is some asymmetry between the actions.

A central feature of our model is that it allows for non-Nash behavior in some games while selecting a unique Nash equilibrium in others. While there are other concepts that allow for this, introspective equilibrium is the only concept that we are aware of that delivers a systemic

understanding of how the type of behavior predicted (i.e., Nash or non-Nash) depends on the payoff structure of the game. We obtain the testable prediction that a unique Nash equilibrium is selected when there are significant asymmetries among actions in terms of payoffs, while for games where the payoff structure gives little guidance, behavior is sensitive to contextual clues and is not consistent with Nash equilibrium. This allows us to better match the experimental data, as discussed in Section 5.

Our paper also contributes to the literature on culture in economics. Kuran and Sandholm (2008) take the culture of a group to be defined by the preferences and equilibrium behaviors of its members. In our model, groups can differ in their culture even if they have identical preferences. Arrow (1974) and Crémer (1993) define culture as the shared knowledge base of a group. Kreps (1990) suggests that culture is a source of focal principles that can help select an equilibrium. In our model, players with a shared background have similar behavioral tendencies. This provides a formal mechanism through which culture can facilitate equilibrium selection. We exploit this mechanism in Kets and Sandroni (2015) to develop a theory of homophily, that is, the tendency for people to interact primarily with people like themselves. This shows that culture can shape social interactions.

A burgeoning literature studies the costs and benefits of diversity. The literature has shown that diversity can be costly if it is associated with preference heterogeneity and conflict (Van den Steen, 2010). On the other hand, diversity can be beneficial if diverse groups have access to more information or skills (Lazear, 1999; Hong and Page, 2001; Prat, 2002) or if differences in opinions provide incentives to acquire costly information (Che and Kartik, 2009; Van den Steen, 2010). Diversity can also improve performance when players have less information about the payoffs of other groups if this leads them to choose an action that is more responsive to the economic environment (Grout, Mitraille, and Sonderegger, 2015). In these models, the benefits of diversity stem from factors that directly affect payoffs, such as preferences, payoff-relevant information, and skills, while we abstract away from such factors. So, any effect of diversity that we identify is above and beyond the effects identified in the existing literature. In a recent experimental paper, Le Coq, Tremewan, and Wagner (2015) demonstrate that group composition can affect strategic behavior even if does not influence payoffs. However, their results are driven by a different mechanism than ours.[32]

[32] Le Coq, Tremewan, and Wagner (2015) induce group identity among the participants in their experiment, that is, a sense of belonging to a social category. The concepts of identity and culture are closely related but distinct. Indeed, in the experiment, induced identity does not have a consistent effect on beliefs. Without such a consistent effect on beliefs, our mechanism cannot operate.

8 Conclusions

Economic models often shy away from modeling endogenous expectations (i.e., beliefs about actions) because they can lead to multiplicity, as in the case of sunspots. We circumvent the problem of multiplicity and self-fulfilling expectations by "anchoring" beliefs. We do so by building on research in psychology on theory of mind. This yields a simple yet versatile solution concept, introspective equilibrium, that allows us to derive testable comparative static results and show how a subtle interplay between economic fundamentals and sociocultural factors determines economic outcomes. Overall, introspective equilibrium delivers results that are broadly consistent with experimental evidence and that are difficult to obtain with existing solution concepts.

In this paper, we illustrate the applicability of introspective equilibrium by studying the effect of culture on economic outcomes in coordination games. We identify a novel channel through which cultural diversity can affect economic performance and show that culture matters even if it is not directly payoff relevant. We apply this insight to a range of applications. Among other contributions, the model sheds light on how small changes in economic conditions can lead to large disparities between societies that are identical in all payoff-relevant aspects; demonstrates that standard policies to resolve collective action problems can be ineffective if social pressure is ignored; and shows that policies to improve the representation of minorities in an organization can have nonmonotonic effects on economic performance.

On the methodological side, introspective equilibrium offers a theory that both sheds light on the economic conditions under which equilibrium analysis is relevant and delivers sharp predictions when it is not. This feature of introspective equilibrium makes it possible to derive unambiguous comparative statics results regardless of whether behavior can be expected to be consistent with Nash equilibrium. Introspective equilibrium thus enlarges the scope of standard economic tools.

There are many directions for future research. The applications in Section 4 offer a glimpse of the economic insights that can be obtained by modeling players' reasoning process explicitly. On the experimental side, the model provides a unified explanation of seemingly disparate experimental findings and suggests new directions for experimental work. On the theory side, a pressing issue is to develop a dynamic model that can explain how behavior and expectations co-evolve with a view to developing a theory of equilibrium transitions: why do some societies get stuck in a "bad" equilibrium, while others are able to jump to a "good" equilibrium? Introspective equilibrium thus has the potential to deliver economic insights well beyond the question of diversity.

Appendix A Strategic uncertainty

In the main text, we claimed that culturally diverse societies face more strategic uncertainty than culturally homogeneous societies. Here we make that claim precise. The degree of uncertainty in a society is equal to the variance V in impulses. In particular, if all players choose the same action, then there is no strategic uncertainty (i.e., $V = 0$), while if each action is chosen with equal probability, then strategic uncertainty is maximal (i.e., $V = \frac{1}{2}$).

The variance in impulses is a function of the probability that players have the same impulses. For a society $\mathcal{S} = (\alpha, q, \eta)$, if a player $j \in N$ belongs to the minority, then the probability that another player $i \neq j$ has the same impulse is

$$Q^{\min} := \alpha \cdot Q_{out} + (1 - \alpha) \cdot Q_{in}.$$

Likewise, if a player j belongs to the majority, then the probability that another player $i \neq j$ has the same impulse is

$$Q^{\text{maj}} := \alpha \cdot Q_{in} + (1 - \alpha) \cdot Q_{out}.$$

Since $\alpha \geq \frac{1}{2}$ and $Q_{in} > Q_{out} > \frac{1}{2}$, $Q^{\text{maj}} \geq Q^{\min} > \frac{1}{2}$ (with strict inequality if $\alpha > \frac{1}{2}$). Then, the degree of strategic uncertainty that a player in the minority and the majority face is given by

$$V^{\min} := \quad Q^{\min} \cdot (1 - Q^{\min}); \text{ and } \quad V^{\text{maj}} := \quad Q^{\text{maj}} \cdot (1 - Q^{\text{maj}}),$$

respectively. The majority faces less strategic uncertainty than the minority (i.e., $V^{\text{maj}} < V^{\min}$). We can also define aggregate strategic uncertainty $V^{\mathcal{S}}$ for the society by

$$V^{\mathcal{S}} := \alpha \cdot V^{\text{maj}} + (1 - \alpha) \cdot V^{\min}.$$

Thus, players face less strategic uncertainty in culturally homogeneous societies than in culturally diverse ones (i.e., $V^{\mathcal{S}}$ is decreasing in the majority share α).

Appendix B Comparison with correlated equilibrium

We compare the set of all introspective equilibria for a given game (across all societies) to the set of correlated equilibria for the game. Relative to correlated equilibrium, introspective equilibrium has considerable cutting power. A first observation is that *for any game* (3.1)*, the set of introspective equilibria (across all societies) is always a strict subset of the class of correlated equilibria.*[33] To make this claim precise, we can identity each society $\mathcal{S} = (\alpha, q, \delta)$

[33]We thus restrict attention to impulse distributions that are associated with some society (i.e., impulse distributions characterized by α, q, δ). Without any restrictions on the class of impulse distributions, any

with the impulse distribution it generates (Section 2.2). Write Δ for the class of impulse distributions that are associated with some society. To be able to compare the set of introspective equilibria (profiles of mappings from impulses to actions) to correlated equilibria (distributions over action profiles), we consider the distributions over action profiles induced by introspective equilibrium. That is, for $\rho \in [0,1]$ and $\mu \in \Delta$, let $\Sigma_\mu(\rho)$ be the set of distributions over action profiles induced by some introspective equilibrium for the society described by μ and risk parameter ρ. By Corollary 3.2, $\Sigma_\mu(\rho)$ has at least one element; and for generic values of ρ, it has precisely one element. For $\rho \in [0,1]$, let $\Sigma(\rho) = \bigcup_{\mu\in\Delta} \Sigma_\mu(\rho)$. With some abuse of terminology, we refer to $\Sigma(\rho)$ as the set of introspective equilibria (across all societies $\mu \in \Delta$) for risk parameter ρ. Let $\mathcal{C}(\rho)$ be the set of correlated equilibria for risk parameter ρ.[34] Then, the following claim, which is a corollary of Theorem 3.1, shows that introspective equilibrium can always rule out certain behaviors that are consisteint with correlated equilibrium:

Corollary B.1. [The Cutting Power of Introspective Equilibrium (I)] *For any $\rho \in [0,1]$, the set $\Sigma(\rho)$ of all introspective equilibria (for some society) is a strict subset of the set $\mathcal{C}(\rho)$ of all correlated equilibria.*

Proof. Let $\rho \in [0,1]$. Then, there is a correlated equilibrium in which all players choose s^1 as well as a correlated equilibrium in which all players choose s^0 (this follows because both are pure Nash equilibria). If $\rho < \frac{1}{2}$, then, by Theorem 3.1, for every $\mu \in \Delta$, there is no introspective equilibrium in $\Sigma_\mu(\rho)$ such that all players choose s^0. Likewise, if $\rho > \frac{1}{2}$, then, by Theorem 3.1, for every $\mu \in \Delta$, there is no introspective equilibrium in $\Sigma_\mu(\rho)$ such that all players choose s^1. Finally, if $\rho = \frac{1}{2}$, then, by Theorem 3.1, for every $\mu \in \Delta$, there is no introspective equilibrium in $\Sigma_\mu(\rho)$ such that all players choose s^0, and there is no introspective equilibrium in $\Sigma_\mu(\rho)$ such that all players choose s^1. □

A second observation is that in some limiting cases, the set of introspective equilibria (across all societies) collapses to a singleton, as the following corollary of Theorem 3.1 demonstrates.[35]

correlated equilibrium is an introspective equilibrium for some impulse distribution. This follows from the revelation principle: fix a correlated equilibrium and take the impulse distribution to be the distribution over action profiles generated by the correlated equilibrium. Then the game has a unique introspective equilibrium in which all players follow their impulse, and this introspective equilibrium coincides with the original correlated equilibrium. See Myerson (1994) for a version of the revelation principle for complete-information game and a discussion in the context of correlated equilibrium.

[34] Note that also for correlated equilibrium, specifying the risk parameter ρ is sufficient to pin down the incentive constraints: any two games (3.1) with the same risk parameter have the same set of correlated equilibria.

[35] The limit of a collection of sets is the set-theoretic limit.

Corollary B.2. [The Cutting Power of Introspective Equilibrium (II)] *As ρ goes to 0, 1, or $\frac{1}{2}$, the set of introspective equilibria (across all societies) converges to a singleton:*

(a) As $\rho \to 0$, the set of introspective equilibria (across all societies) converges to the unique strategy profile where all players choose the efficient action regardless of their impulse;

(b) As $\rho \to 1$, the set of introspective equilibria (across all societies) converges to the unique strategy profile where all players choose the inefficient action regardless of their impulse;

(c) As $\rho \to \frac{1}{2}$, the set of introspective equilibria (across all societies) converges to the unique strategy profile where all players follow their impulse.

Again, the proof follows directly from Theorem 3.1. So, in the limit that the risk parameter goes to 0, 1, or $\frac{1}{2}$, the set of introspective equilibria (across all societies) collapses to a singleton, and the limiting introspective equilibrium is independent of sociocultural factors. By contrast, the set of correlated equilibria does not converge to a singleton when the risk parameter goes to 0, 1, or $\frac{1}{2}$. Instead, it is a continuum (except in the trivial case $v_1 = v_0 = 0$ and $c_1 = c_0$). To see this, note that for any $\rho \in [0, 1]$, the set of correlated equilibria contains at least the strict Nash equilibria as well as the nonstrict pure Nash equilibrium in which a proportion ρ of players chooses s^1 (footnote 7); the claim now follows by noting that, except when $v_1 = v_0 = 0$ and $c_1 = c_0$, at least two of these Nash equilibria have different payoff profiles, and the set of correlated equilibrium payoff profiles includes the convex hull of Nash equilibrium payoff profiles.

Appendix C Proofs

C.1 Proof of Proposition 2.1

We will consider a slightly more general setting. In the applications we consider, players belong to different cultural groups. That is, the set of players is partitioned into a finite set Γ of groups. Each group $\gamma \in \Gamma$ contains a continuum of (identical) players, and players know which group they belong to. Denote the proportion of players who belong to group $\gamma \in \Gamma$ by $\alpha_\gamma \in [0, 1]$ (so $\sum_\gamma \alpha_\gamma = 1$). For each group $\gamma \in \Gamma$, there is an underlying state θ_γ (taking values in Θ_γ) such that conditional on $\theta := (\theta_\gamma)_{\gamma \in \Gamma} \in \Theta$, with $\Theta := \prod_{\gamma \in \Gamma} \Theta_\gamma$, the impulses of players are independent. (The statement of Proposition 2.1 thus refers to the special case where $|\Gamma| = 1$.) Then, for every $\theta \in \Theta$ and $\gamma \in \Gamma$, there is $p_{\theta,\gamma} \in [0, 1]$ such that the realized proportion of players in γ with an impulse to play s^1 is $p_{\theta,\gamma}$ (with probability 1). Since the

tie-breaking rule is identical across players,[36] at every level $k = 0, 1, \ldots$, any two players from the same group have the same level-k strategy. The expected payoff to a player $j \in N$ if he has impulse I_j and belongs to group $\gamma_j \in \Gamma$ from choosing action $s_j \in S_j$ when other players play according to their level-k strategies σ^k_{-j} is then

$$U_j(s_j, \sigma^k_{-j}; I_j, \gamma_j) := \int u\big(s_j, \sum_\gamma \alpha_\gamma m^{\sigma^k_{-j}}_{\theta,\gamma}\big) dF((p_{\theta,\gamma})_\gamma \mid I_j, \gamma_j)$$

where $F((p_{\theta,\gamma})_\gamma \mid I_j, \gamma_j)$ is the player's posterior over $(p_{\theta,\gamma})_\gamma$ given his impulse I_j and group γ_j and where $m^{\sigma^k_{-j}}_{\theta,\gamma}$ is the proportion of players $i \neq j$ in group γ that choose s^1 under the strategy profile σ^k_{-j}. That is, if all players in γ follow their impulse under σ^k_{-j}, then $m^{\sigma^k_{-j}}_{\theta,\gamma} = p_{\theta,\gamma}$; if all players in γ choose s^1 under σ^k_{-j}, then $m^{\sigma^k_{-j}}_{\theta,\gamma} = 1$, etc.

We need to show that, if for each player $j \in N$, the level-k strategies σ^k_j converge to a strategy σ_j, then the profile $\sigma = (\sigma_j)_{j \in N}$ of limiting strategies is a correlated equilibrium. That is, we need to show that for $j \in N$ and $I_j = s^0, s^1$,

$$\sigma_j(I_j) \in \arg\max_s U_j(s, \sigma_{-j}; I_j, \gamma_j)$$

where γ_j is the group that j belongs to. Since the set of impulses is finite and players' level-k strategy depends only on the group they belong to, there is $K < \infty$ such that for all $k \geq K$ and $j \in N$, $\sigma^k_j = \sigma^{k-1}_j$. So, for every $k > K$, every player plays a best response against the other players' strategies, that is, for $k > K$, $j \in N$, $I_j = s^0, s^1$,

$$\sigma^k_j(I_j) \in \arg\max_s U_j(s, \sigma^k_{-j}; I_j, \gamma_j),$$

and the result follows. □

Remark 1. The proof extends to more general settings. For example, the proof can be extended to environments where there is heterogeneity in payoffs or beliefs so that each player has a continuum of types and the convergence of the introspective process potentially takes infinitely many steps (provided that the payoff functions satisfy suitable continuity assumptions). ◁

C.2 Proof of Theorem 3.1

Recall that the risk parameter is

$$\rho = \frac{v_0 + c_1 - c_0}{v_0 + v_1}.$$

[36] This assumption is stronger than necessary: it merely helps us avoid measurability problems.

That is, if a player believes that a proportion $m \geq \rho$ (resp. $m \geq \rho$) of players chooses s^1, then choosing s^1 is a best response (resp. the unique best response).

For brevity, refer to players in the majority group (resp. minority group) as *majority players* (resp. *minority players*). At each level $k > 1$, each player forms a belief about the proportion of players choosing s^1 given the level-$(k-1)$ strategies and his impulse. Since players are symmetric in the sense that any two players with the same impulse who belong to the same group have the same beliefs, this belief depends only on the player's impulse and on the group he belongs to. Denote the expected proportion of players who choose s^1 at level 1 for a majority player (resp. minority player) with impulse $I = 0, 1$ by $m^{k-1}_{\text{maj},I}$ (resp. $m^{k-1}_{\text{min},I}$).

Level 0 All players follow their impulse.

Level 1 At level 1, each player chooses a best response against the belief that players follow their impulse. Given the belief that players follow their impulse, we have

$$\begin{aligned} m^0_{\text{maj},s^1} &= \alpha Q_{in} + (1-\alpha)\cdot Q_{out}; \quad m^0_{\text{maj},s^0} = \alpha\cdot(1-Q_{in}) + (1-\alpha)\cdot(1-Q_{out}); \\ m^0_{\text{min},s^1} &= \alpha Q_{out} + (1-\alpha)\cdot Q_{in}; \quad m^0_{\text{min},s^0} = \alpha\cdot(1-Q_{out}) + (1-\alpha)\cdot(1-Q_{in}). \end{aligned}$$

Note that $m^0_{\text{maj},s^1} \geq m^0_{\text{min},s^1} > m^0_{\text{min},s^0} \geq m^0_{\text{maj},s^0}$ (with strict inequalities if $\alpha > \frac{1}{2}$). Thus:

(a_1) If

$$\rho \geq m^0_{\text{maj},s^1}$$

then, at level 1, it is a best response for players to choose s^0 (and this is the unique best response if the inequality is strict).

(b_1) If

$$m^0_{\text{min},s^1} \leq \rho \leq m^0_{\text{maj},s^1}$$

then, at level 1, it is a best response for minority players to choose s^0 (regardless of their impulse) while it is a best response for majority players to follow their impulse (and these are the unique best responses if the inequalities are strict).

(c_1) If

$$m^0_{\text{min},s^0} \leq \rho \leq m^0_{\text{min},s^1}$$

then, at level 1, it is a best response for players to follow their impulse (and this is the unique best response if the inequalities are strict).

(d_1) If

$$m^0_{\text{maj},s^1} \leq \rho \leq m^0_{\text{min},s^0}$$

then, at level 1, it is a best response for minority players to choose s^1 (regardless of their impulse) while it is a best response for majority players to follow their impulse (and these are the unique best responses if the inequalities are strict).

(e_1) If

$$\rho \leq m^0_{\text{maj},s^1}$$

then, at level 1, it is a best response for players to choose s^1 (and this is the unique best response if the inequality is strict).

Figure 3(a) plots the level-1 parameter boundaries as a function of α (where the values for Q_{in}, Q_{out} correspond to the values for q, δ in the example in Section 3).

Level 2 As we have seen, if players believe that all players follow their impulse, then, for the parameter range in (a_1), playing s^0 is a best response. At level 2, players thus choose a best response against the belief that all players choose s^0. By (3.2), the best response is to choose s^0. So, if

$$\rho \geq \alpha Q_{in} + (1-\alpha) \cdot Q_{out},$$

then, at level 2, players choose s^0 (and this is the unique best response if the inequality is strict). Likewise, if players believe that all players follow their impulse, then, for the parameter range in (e_1), playing s^1 is a best response. At level 2, players thus choose a best response against the belief that all players choose s^1. By (3.2), the best response is to choose s^1. So, if

$$\rho \leq \alpha \cdot (1 - Q_{in}) + (1-\alpha) \cdot (1 - Q_{out})$$

then, at level 2, players choose s^1 (and this is the unique best response if the inequality is strict). If players believe that all players follow their impulse, then, for the parameter range in (c_1), it is a best response for players to follow their impulse. So, if

$$\alpha \cdot (1 - Q_{out}) + (1-\alpha) \cdot (1 - Q_{in}) \leq \rho \leq \alpha Q_{out} + (1-\alpha) \cdot Q_{in}$$

then, at level 2, players follow their impulse (and this is the unique best response if the inequalities are strict).

Hence, it remains to consider the parameter ranges in (b_1) and (d_1). First consider the parameter range in (b_1). As we have seen, if players believe that all players follow their impulse, then it is a best response for minority players to choose s^0, while it is a best response for majority players to follow their impulse. So, at level 2, it is a best response for minority players to choose s^0 (since the incentive to choose s^0 increases in the proportion of players choosing s^0). Consider a majority player. At level 2, he chooses a best response against the

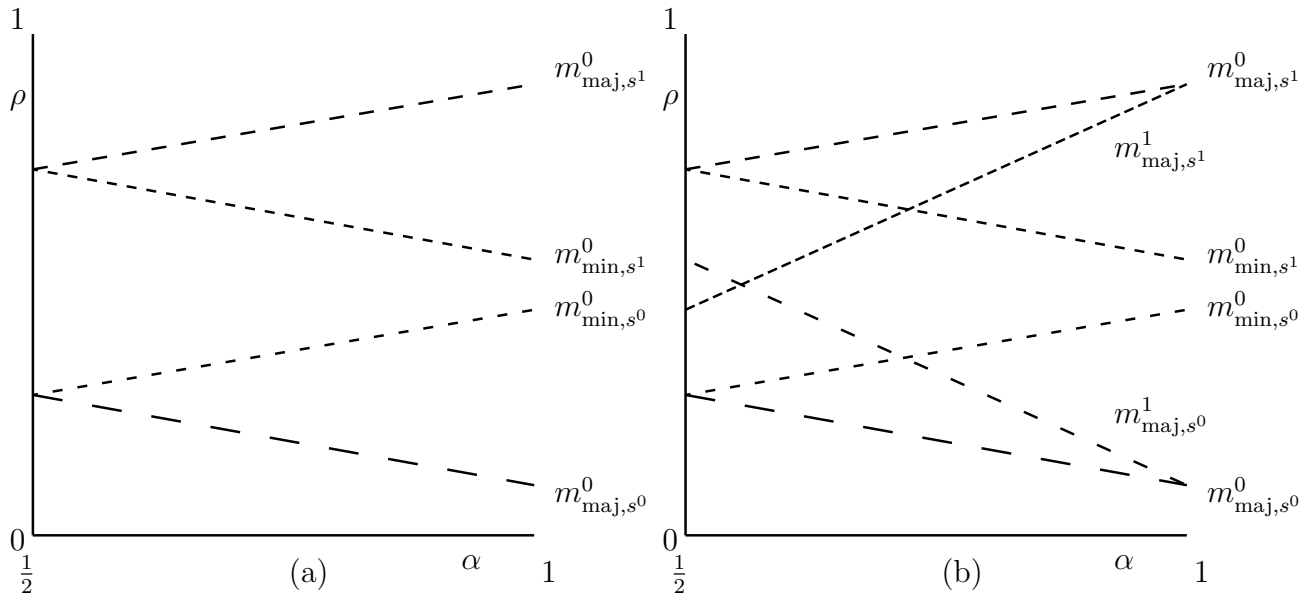

Figure 3: (a) The level-1 parameter bounds as a function of the population composition (for given Q_{in}, Q_{out}); (b) The level-2 parameter bounds.

belief that the minority players choose s^0 and that the majority players follow their impulse. If the player has an impulse to choose s^0, then it is a best response to play s^0.[37] If a player has an impulse to choose s^1, then it is best response to go against his impulse and choose s^0 instead if

$$\rho \geq m^1_{\text{maj},s^1},$$

where $m^1_{\text{maj},s^1} = \alpha Q_{in}$ (and this is the unique best response if the inequality is strict).

Next consider the parameter range in (d_1). As we have seen, if players believe that all players follow their impulse, then it is a best response for minority players to choose s^1, while it is a best response for majority players to follow their impulse. So, at level 2, it is a best response for minority players to choose s^1 (since the incentive to choose s^1 increases in the proportion of players choosing s^1). Consider a majority player. At level 2, he chooses a best response against the belief that the minority players choose s^1 and that the majority players follow their impulse. If the player has an impulse to choose s^1, then, by a similar argument as before, it is a best response to play s^1. If a player has an impulse to choose s^0, then it is best response to go against his impulse and choose s^1 instead if

$$\rho \leq m^1_{\text{maj},s^0},$$

[37] Since, for the parameter range in (b_1) s^0 is a best response against the belief that players follow their impulse, it is a fortiori a best response against the belief that the minority players choose s^0 and that the majority players follow their impulse.

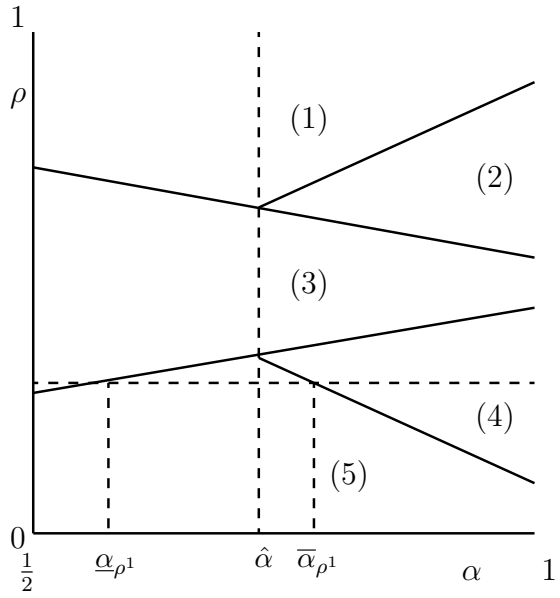

Figure 4: The introspective equilibrium as a function of the population composition (for given Q_{in}, Q_{out}), with (1) denoting the parameter region where all players choose s^0; (2) denoting the region where minority players choose s^0 and majority players follow their impulse; (3) denoting the region where all players follow their impulse; (4) denoting the region where minority players choose s^1 and majority players follow their impulse; (5) denoting the region where all players choose s^1.

where $m^1_{\text{maj},s^0} = \alpha \cdot (1 - Q_{in}) + 1 - \alpha = 1 - \alpha Q_{in}$ (and this is the unique best response if the inequality is strict). Together, these bounds describe the level-2 strategies. Figure 3(b) plots the level-2 boundaries as a function of α (where again the values for Q_{in}, Q_{out} are chosen to match the values for q, δ in the example in Section 3).

It can be checked that no player has an incentive to change his action at level 3 (or any higher level). So, the strategies played in the introspective equilibrium coincide with the level-2 strategies. So, using the expressions for $m^1_{\text{maj},s^1}, m^1_{\text{maj},s^0}, m^0_{\text{min},s^1}$, and m^0_{min,s^0}, the result follows if we define

$$\begin{aligned} \rho_4 = &\ \max\Big\{\alpha Q_{in}, \alpha Q_{out} + (1-\alpha)\cdot Q_{in}\Big\}; \\ \rho_3 = &\ \alpha Q_{out} + (1-\alpha)\cdot Q_{in}; \\ \rho_2 = &\ \alpha\cdot(1-Q_{out}) + (1-\alpha)\cdot(1-Q_{in}); \\ \rho_1 = &\ \min\Big\{1-\alpha Q_{in}, \alpha\cdot(1-Q_{out}) + (1-\alpha)\cdot(1-Q_{in})\Big\}. \end{aligned} \tag{C.1}$$

□

Theorem 3.1 shows that there are five regimes. These regimes are illustrated in Figure 4 (where (1) corresponds to Theorem 3.1(a), (2) corresponds to Theorem 3.1(b), (3) corresponds to Theorem 3.1(c), (4) corresponds to Theorem 3.1(d), and (5) corresponds to Theorem 3.1(e)).

If the risk parameter is very low, then all players choose s^1. As the risk parameter increases, players in the majority group begin to follow their impulses. As the risk parameter increases further, also the players in the minority group begin to follow their impulses. When the risk parameter increases further, the minority switches to playing s^0, regardless of their impulse. As the risk parameter increases even more, all players choose s^0.

C.3 Proof of Theorem 3.3

The proof uses Theorem 3.1. Theorem 3.1 shows that for given values of ρ, Q_{in}, and Q_{out}, there can be different introspective equilibria depending on the population composition (i.e., α). Figure 4 illustrates the different regimes. The five regimes are separated by boundaries; see Figure 4. It will be useful to write

$$\hat{\alpha} := \frac{Q_{in}}{2Q_{in} - Q_{out}}$$

for the population composition where the boundaries intersect; see Figure 4.

We calculate the total payoff for the society in introspective equilibrium for each of the regimes. Fix v_1, v_0, c_1, c_0 and define $H := \alpha^2 + (1-\alpha)^2$ and $Q := q^2 + (1-q)^2$ (so $Q = Q_{in}$). Since $q \in (\frac{1}{2}, 1)$, $Q \in (\frac{1}{2}, 1)$ and Q is strictly increasing in q; since $\alpha \in [\frac{1}{2}, 1]$, $H \in [\frac{1}{2}, 1]$ and H is increasing in α, and strictly so if $\alpha > \frac{1}{2}$.

We consider the different regimes in turn. First, if

$$\rho > \max\Big\{\alpha Q_{in}, \alpha Q_{out} + (1-\alpha)\cdot Q_{in}\Big\},$$

then in the unique introspective equilibrium, all players choose s^0 (Theorem 3.1(a), (1) in Figure 4). In this case, the total payoff is

$$\Pi_{s^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) := v_0 - c_0,$$

independent of α. Second, if

$$\alpha Q_{out} + (1-\alpha)\cdot Q_{in} < \rho < \alpha Q_{in},$$

then in the unique introspective equilibrium, minority players choose s^0 while majority players follow their impulse (Theorem 3.1(b), (2) in Figure 4). In this case, the total payoff is

$$\Pi_{\min\, s^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) := \tfrac{1}{2}\cdot\Big[\big(\alpha^2 Q + 2\cdot(1-\alpha)\big)\cdot v_0 + \alpha^2 Q v_1 - (2-\alpha)\cdot c_0 - \alpha c_1\Big].$$

Third, if

$$\alpha\cdot(1-Q_{out}) + (1-\alpha)\cdot(1-Q_{in}) < \rho < \alpha Q_{out} + (1-\alpha)\cdot Q_{in}$$

then in the unique introspective equilibrium, all players follow their impulse (Theorem 3.1(c), (3) in Figure 4). In this case, the total payoff is

$$\Pi_{\text{FI}}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) := (\tfrac{1}{4} + \eta) \cdot \big[Q \cdot (v_0 + v_1) - c_1 - c_0\big] + (\tfrac{1}{4} - \eta) \cdot \big[(QH + (1-Q) \cdot (1-H)) \cdot (v_1 + v_0) - c_1 - c_0\big].$$

Fourth, if

$$1 - \alpha Q_{in} < \rho < \alpha \cdot (1 - Q_{out}) + (1 - \alpha) \cdot (1 - Q_{in}),$$

then in the uique introspective equilibrium, players in the minority group choose s^1, while players in the majority group follow their impulse (Theorem 3.1(d), (4) in Figure 4). In this case, the total payoff is

$$\Pi_{\min\, s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) := \tfrac{1}{2} \cdot \Big[\big(\alpha^2 Q + 2 \cdot (1-\alpha)\big) \cdot v_1 + \alpha^2 Q v_0 - (2-\alpha) \cdot c_1 - \alpha c_0\Big].$$

Fifth, if

$$\rho < \min\Big\{1 - \alpha Q_{in}, \alpha \cdot (1 - Q_{out}) + (1-\alpha) \cdot (1 - Q_{in})\Big\},$$

then in the unique introspective equilibrium, all players choose s^1 (Theorem 3.1(e), (5) in Figure 4). In this case, the total payoff is

$$\Pi_{s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) := v_1 - c_1.$$

It is easy to verify that Π_{FI} is strictly increasing in α. Moreover,

$$\Pi_{s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) > \Pi_{\text{FI}}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}), \Pi_{\min\, s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out})$$

for all parameter values. Intuitively, if all players choose s^1, then there is no miscoordination; moreover, the society coordinates on the action that yields the highest coordination payoff (i.e., $v_1 - c_1 \geq v_0 - c_0$).

We can use these results to identify the population composition that maximizes the total payoff in introspective equilibrium (for given ρ). If $\rho < 1 - Q_{in}$, then all players choose s^1 in introspective equilibrium, regardless of the population composition. In this case, any population composition $\alpha \in [0, 1]$ is optimal (i.e., maximizes the expected total payoff in introspective equilibrium).

Next consider the case $\rho \in [1 - Q_{in}, 1 - \hat{\alpha} Q_{in}]$. In this case, the introspective equilibrium takes three forms, depending on the population composition (i.e., α). Define

$$\underline{\alpha}_{\rho^1} := \begin{cases} \frac{1}{2} & \text{if } \rho < \frac{1}{2} \cdot (1 - Q_{out}) + \frac{1}{2} \cdot (1 - Q_{in}); \\ \frac{\rho - (1 - Q_{in})}{Q_{in} - Q_{out}} & \text{otherwise;} \end{cases}$$

and

$$\overline{\alpha}_{\rho^1} := \frac{1-\rho}{Q_{in}}.$$

That is, $\underline{\alpha}_{\rho^1}$ is the unique $\alpha \in [\frac{1}{2}, 1]$ such that $\rho = \alpha \cdot (1 - Q_{out}) + (1 - \alpha) \cdot (1 - Q_{in})$, and $\overline{\alpha}_{\rho^1}$ is the unique $\alpha \in [\frac{1}{2}, 1]$ such that $\rho = 1 - \alpha Q_{in}$; see Figure 4. Note that $\underline{\alpha}_{\rho^1} < \overline{\alpha}_{\rho^1} < 1$. Then, if $\alpha < \underline{\alpha}_{\rho^1}$, then all players follow their impulse in the unique introspective equilibrium. If $\alpha \in (\underline{\alpha}_{\rho^1}, \overline{\alpha}_{\rho^1})$, then players choose s^1 in the unique introspective equilibrium. Finally, if $\alpha > \overline{\alpha}_{\rho^1}$, there is a unique introspective equilibrium, and in this introspective equilibrium, minority players choose s^1 while majority players follow their impulse. By the above, α is the population composition that maximizes total payoffs (for given ρ, q, η) if all players choose s^1 in introspective equilibrium. This is the case whenever $\alpha \in (\underline{\alpha}_{\rho^1}, \overline{\alpha}_{\rho^1})$. So, if s^1 is attractive relative to s^0 but not too low (i.e., $\rho \in (1 - Q_{in}, 1 - \hat{\alpha} Q_{in})$), then cultural diversity is optimal (i.e., the optimal population composition α^* satisfies $\alpha^* \in (\underline{\alpha}_{\rho^1}, \overline{\alpha}_{\rho^1})$, with $\overline{\alpha}_{\rho^1} < 1$).

Before considering the case that $\rho \in (1 - \hat{\alpha} Q_{in}, 1 - Q_{out})$, it will be helpful to consider the case $\rho \in (1 - Q_{out}, Q_{out})$. By the above, if $\rho \in (1 - Q_{out}, Q_{out})$, there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse. Since Π_{FI} is increasing in α, the optimal population composition is $\alpha^* = 1$.

For $\rho \in (1 - \hat{\alpha} Q_{in}, 1 - Q_{out})$, the introspective equilibrium takes two forms, depending on the population composition. If $\alpha < \underline{\alpha}_{\rho^1}$, then there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse; otherwise, minority players choose s^1 in introspective equilibrium while majority players follow their impulse. As Π_{FI} is increasing in α, the optimal population composition α^* satisfies $\alpha^* \geq \underline{\alpha}_{\rho^1}$. The function $\Pi_{\min\, s^1}$ is nonmonotonic in α; its second derivative is a positive constant. So, the function attains its minimum at some α. Denote the population composition at which the minimum is attained by $\tilde{\alpha}$.

We need to consider two cases for $\tilde{\alpha}$. First consider the case that $\tilde{\alpha} \geq 1$. Then, for any $\alpha < 1$, $\Pi_{\min s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out})$ is decreasing in α; moreover, $\Pi_{\min\, s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) > \Pi_{\text{FI}}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out})$ for all $\alpha < 1$. Hence, $\alpha^* = \alpha_{\rho^1} < 1$. Define $\underline{\rho}^{\tilde{\alpha} \geq 1} := 1 - Q_{out}$.

Next suppose $\tilde{\alpha} < 1$. Define $\alpha_{\text{BE}}^{s^1}$ to be the unique $\alpha \neq 1$ such that $\Pi_{\min\, s^1}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) = \Pi_{\text{FI}}(v_0, v_1, c_0, c_1; 1, Q_{in}, Q_{out})$. That is, $\alpha_{\text{BE}}^{s^1}$ is the "break-even" for the introspective equilibrium: at $\alpha = \alpha_{\text{BE}}^{s^1}$, the total payoff in the introspective equilibrium where the minority group plays s^1 (while the majority players follow their impulse) is equal to the highest expected payoff in the introspective equilibrium in which all players follow their impulse (and if $\alpha < \alpha_{BE}^{s^1}$, the total payoff is higher in the introspective equilibrium in which the minority group plays s^1 than in the introspective equilibrium in which all players follow their impulse). Then, if $\alpha_{\text{BE}}^{s^1} > \underline{\alpha}_{\rho^1}$, then cultural diversity is optimal: $\alpha^* = \alpha_{\rho^1} < 1$. If $\alpha_{\text{BE}}^{s^1} < \underline{\alpha}_{\rho^1}$, then cultural homogeneity is

optimal (i.e., $\alpha^* = 1$). Define

$$\Delta^{s^1} := \alpha_{\text{BE}}^{s^1} - \underline{\alpha}_{\rho^1}.$$

Since c_0, c_1 are fixed, Δ^{s^1} decreases with ρ. Since $\alpha^* = 1$ if $\rho \in (1 - Q_{out}, Q_{out})$ and $\alpha^* < 1$ if $\rho \in (1-Q_{in}, 1-\hat{\alpha}Q_{in})$, there is $\underline{\rho}^{\tilde{\alpha}<1} \in [1-\hat{\alpha}Q_{in}, 1-Q_{out}]$ such that $\alpha^* < 1$ if $\rho \in (1-\hat{\alpha}Q_{in}, \underline{\rho}^{\tilde{\alpha}<1})$ and $\alpha^* = 1$ if $\rho \in (\underline{\rho}^{\tilde{\alpha}<1}, Q_{out})$. (Note that $\underline{\rho}^{\tilde{\alpha}<1} < \frac{1}{2}$.) Then, define

$$\underline{\rho} := \begin{cases} \underline{\rho}^{\tilde{\alpha}\geq 1} & \text{if } \tilde{\alpha} \geq 1; \\ \underline{\rho}^{\tilde{\alpha}<1} & \text{otherwise.} \end{cases}$$

For $\rho \in (Q_{out}, \hat{\alpha}Q_{in})$, the introspective equilibrium takes two forms, depending on the population composition. Define

$$\underline{\alpha}_{\rho^0} := \begin{cases} \frac{1}{2} & \text{if } \rho \geq \frac{1}{2} \cdot (Q_{in} + Q_{out}); \\ \frac{Q_{in}-\rho}{Q_{in}-Q_{out}} & \text{otherwise;} \end{cases}$$

and

$$\overline{\alpha}_{\rho^0} := \frac{\rho}{Q_{in}}.$$

That is, $\underline{\alpha}_{\rho^0}$ is the unique $\alpha \in [\frac{1}{2}, 1]$ such that $\rho = \alpha Q_{out} + (1-\alpha) \cdot Q_{in}$, and $\overline{\alpha}_{\rho^0}$ is the unique $\alpha \in [\frac{1}{2}, 1]$ such that $\rho = \alpha Q_{in}$. Then, if $\alpha < \overline{\alpha}_{\rho^0}$, there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse; otherwise, minority players choose s^0 in introspective equilibrium while majority players follow their impulse. As Π_{FI} is increasing in α, the optimal population composition α^* satisfies $\alpha^* \geq \underline{\alpha}_{\rho^0}$. Again, the function $\Pi_{\min s^0}$ is nonmonotonic in α; its second derivative is a positive constant. So, the functions attains its minimum at some α; moreover, it can be checked that the minimum is attained at $\alpha \leq 1$. Define $\alpha_{\text{BE}}^{s^0}$ to be the unique $\alpha \neq 1$ such that $\Pi_{\min s^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) = \Pi_{\text{FI}}(v_0, v_1, c_0, c_1; 1, Q_{in}, Q_{out})$. That is, $\alpha_{\text{BE}}^{s^0}$ is the "break-even" for the introspective equilibrium: at $\alpha = \alpha_{\text{BE}}^{s^0}$, the total payoff in the introspective equilibrium where the minority group plays s^0 (while the majority players follow their impulse) is equal to the highest expected payoff in the introspective equilibrium in which all players follow their impulse. Then, it can be checked that $\alpha_{\text{BE}}^{s^0} < \overline{\alpha}_{\rho^0}$ for every $\rho \in (Q_{out}, \hat{\alpha}Q_{in})$, and it follows that cultural homogeneity is optimal (i.e., $\alpha^* = 1$).

For $\rho \in (\hat{\alpha}Q_{in}, Q_{in})$, the introspective equilibrium takes three forms, depending on the population composition. If $\alpha < \underline{\alpha}_{\rho^0}$, then there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse. If $\alpha \in (\underline{\alpha}_{\rho^0}, \overline{\alpha}_{\rho^0})$, then there is a unique introspective equilibrium, and in this introspective equilibrium, players choose s^0 regardless of their impulse. Finally, if $\alpha > \overline{\alpha}_{\rho^0}$, there is a unique introspective equilibrium, and in this introspective equilibrium, minority players choose s^0 while majority players follow their impulse.

It can be checked that for every $\rho \in (\hat{\alpha}Q_{in}, Q_{in})$, $\alpha^{s^0}_{\text{BE}} < \underline{\alpha}_{\rho^0}$. Hence, the optimal population composition is $\alpha^* = 1$ if and only if $\Pi_{s^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) < \Pi_{\text{FI}}(v_0, v_1, c_0, c_1; 1, Q_{in}, Q_{out})$ (for some α; recall that Π_{s^0} is independent of α). This holds if and only if

$$\rho < Q_{in} - \frac{v_0}{v_0 + v_1}.$$

Since c_0, c_1 are fixed, there is $\hat{\rho} \in (\hat{\alpha}Q_{in}, Q_{in}]$ such that for $\rho \in (\hat{\alpha}Q_{in}, Q_{in})$, $\alpha^* = 1$ if and only if $\rho < \hat{\rho}$ (with $\hat{\rho} = Q_{in}$ if and only if $v_0 = 0$); and $\alpha^* \in [\underline{\alpha}_{\rho^0}, \overline{\alpha}^{s^0}]$ otherwise. Given that $\alpha^* = 1$ for $\rho \in (Q_{out}, \hat{\alpha}Q_{in}]$, if we define $\overline{\rho} := \hat{\rho}$, then $\alpha^* = 1$ for any $\rho \in [\underline{\rho}, \overline{\rho}]$.

Finally, if $\rho > Q_{in}$, all players choose s^0 in introspective equilibrium, and the total payoff is independent of the population composition.

To summarize, define (for given ρ, q, η), the *payoff-maximizing introspective equilibrium* to be the introspective equilibrium when the population composition is α^*. That is, the payoff-maximizing introspective equilibrium is the introspective equilibrium at the population composition that maximizes the total payoff in equilibrium (given ρ, q, η). Then,

- For $\rho < \rho^1 := 1 - Q_{in}$, the total payoff is independent of the population composition: for any population composition, there is a unique introspective equilibrium; in this introspective equilibrium, all players choose s^1, regardless of their impulse;

- For $\rho \in (\rho^1, 1 - \hat{\alpha}Q_{in})$, in the payoff-maximizing introspective equilibrium, all players choose s^1, and the optimal population composition satisfies $\alpha^* \in (\underline{\alpha}_{\rho^1}, \overline{\alpha}_{\rho^1})$;

- For $\rho \in (1 - \hat{\alpha}Q_{in}, \underline{\rho})$, in the payoff-maximizing introspective equilibrium, minority players choose s^1 while majority players follow their impulse, and the optimal population composition is $\alpha^* = \underline{\alpha}_{\rho^1} < 1$;

- For $\rho \in (\underline{\rho}, \overline{\rho})$, in the payoff-maximizing introspective equilibrium, all players follow their impulse, and the optimal population composition is $\alpha^* = 1$.

- For $\rho \in (\overline{\rho}, Q_{in})$, in the payoff-maximizing introspective equilibrium, all players choose s^0, and the optimal population composition is $\alpha^* < 1$.

- For $\rho > \rho^0 := Q_{in}$, the total payoff is independent of the population composition: for any population composition, there is a unique introspective equilibrium; in this introspective equilibrium, all players choose s^0, regardless of their impulse.

This completes the proof. □

C.4 Proof of Proposition 4.1

We show that there is a discontinuity in total payoff when players abandon the custom. It can be checked that the persistence threshold satisfies $\rho_1(\alpha, q, \delta) < 1 - Q_{out}$ for $\alpha \in [\frac{1}{2}, 1]$, $q \in (\frac{1}{2}, 1)$, and $\delta \in (\frac{3}{4}, 1)$. Fix q, δ. We need to consider two cases: $\alpha < \hat{\alpha}$ and $\alpha > \hat{\alpha}$ (Figure 4). First, for $\alpha < \hat{\alpha}$, if ρ is sufficiently close to $\rho_1(\alpha, q, \delta)$ (say, $\rho \in (\rho_1(\alpha, q, \delta), 1 - Q_{out})$), then there is a unique introspective equilibrium, and in this introspective equilibrium, all players follow their impulse. By the proof of Theorem 3.3, the total payoff in introspective equilibrium is

$$\Pi_{\text{FI}}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) = \tfrac{1}{2} \cdot (w_1 + w_0 - c) + c \cdot \big((\tfrac{1}{4} + \eta) \cdot Q + (\tfrac{1}{4} - \eta) \cdot (Q \cdot H + (1 - Q) \cdot (1 - H)\big).$$

By the proof of Theorem 3.3, $\Pi_{\text{FI}}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out})$ is increasing in α. So, as ρ approaches $\rho_1(\alpha, q, \delta)$ from above, the increase in total payoff is greater than

$$w_1 - \tfrac{1}{2} \cdot \Big((w_0 + w_1) - \tfrac{1}{2} \cdot c \cdot (1 - Q_{in})\Big) = \tfrac{1}{2} \cdot (w_1 - w_0) + \tfrac{1}{2} \cdot c \cdot (1 - Q_{in}).$$

Second, for $\alpha > \hat{\alpha}$, if ρ is sufficiently close to $\rho_1(\alpha, q, \delta)$ (say $\rho \in (\rho_1(\alpha, q, \delta), 1 - Q_{out})$), then there is a unique introspective equilibrium, and in this introspective equilibrium, players from the minority group choose s^1 while players from the majority group follow their impulse. By the proof of Theorem 3.3, the total payoff in introspective equilibrium is

$$\Pi_{\text{mins}^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) = \tfrac{1}{2} \cdot \Big((\alpha^2 \cdot Q_{in} + 2 \cdot (1 - \alpha)) \cdot c + (2 - \alpha) \cdot (w_1 - c) + \alpha \cdot w_0\Big).$$

If $\alpha > \hat{\alpha}$, $\rho_1(\alpha, q, \delta) = 1 - \alpha Q_{in}$. For $\rho = \rho_1(\alpha, q, \delta)$,

$$\Pi_{\text{mins}^0}(v_0, v_1, c_0, c_1; \alpha, Q_{in}, Q_{out}) = \tfrac{1}{2Q_{in}} \cdot \big(w_0 - (1 - 2Q_{in}) \cdot w_1\big).$$

Hence, for $\alpha > \hat{\alpha}$, as ρ approaches $\rho_1(\alpha, q, \delta)$ from above, the expected total payoff increases from $\frac{1}{2Q_{in}} \cdot (w_0 - (1 - 2Q_{in}) \cdot w_1)$ to w_1. Thus, the increase in total payoff is

$$\tfrac{1}{2Q_{in}} \cdot (w_1 - w_0).$$

Hence, for any $\alpha \in [\frac{1}{2}, 1]$, as ρ approaches $\rho_1(\alpha, q, \delta)$ from above, the increase in expected total payoff is at least

$$\min\left\{\tfrac{1}{2} \cdot (w_1 - w_0) + \tfrac{1}{2} \cdot c \cdot (1 - Q_{in}), \tfrac{1}{2Q_{in}} \cdot (w_1 - w_0)\right\}.$$

Since $\rho > 1 - Q_{in}$, we have

$$\tfrac{1}{2} \cdot (w_1 - w_0) + \tfrac{1}{2} \cdot c \cdot (1 - Q_{in}) > \tfrac{1}{2Q_{in}} \cdot (w_1 - w_0),$$

and the result follows. □

C.5 Proof of Corollary 4.2

By the proof of Theorem 3.1,

$$\rho_1(\alpha, q, \delta) = \min\{1 - \alpha Q_{in}(q), \alpha \cdot (1 - Q_{out}(q, \delta)) + (1 - \alpha) \cdot (1 - Q_{in}(q))\},$$

where we have made the dependence of Q_{in} and Q_{out} on q and δ explicit. The result now follows from noting that Q_{in} and Q_{out} increase in q, while Q_{out} decreases in δ. That is, the probability that players have the same impulse increases with culture strength and with cultural closeness; and the persistence threshold shifts up when culture strength or cultural closeness increase. □

C.6 Proof of Corollary 4.3

By the proof of Theorem 3.1, for $\rho \in (1 - Q_{in}, \frac{1}{2} \cdot (1 - Q_{out}) + \frac{1}{2} \cdot (1 - Q_{in}))$, all players choose the efficient action in the unique introspective equilibrium if and only if the population composition α is greater than $\overline{\alpha}_{\rho^1}$ (Figure 4), where $\overline{\alpha}_{\rho^1}$ solves $1 - \overline{\alpha}_{\rho^1} Q_{in} = \rho$. For α less than $\overline{\alpha}_{\rho^1}$, there is a unique introspective equilibrium, and in this unique introspective equilibrium, players from the minority group choose s^1 while players from the majority group follow their impulse. Thus, if we take T to be the open interval $(1 - Q_{in}, \frac{1}{2} \cdot (1 - Q_{out}) + \frac{1}{2} \cdot (1 - Q_{in}))$ and $\underline{a}(\rho) := \overline{\alpha}_{\rho^1}$. □

C.7 Proof of Corollary 4.5

By the proof of Theorem 3.3,

$$\begin{array}{lll} \frac{\mathrm{d}\Pi_{s^0}}{\mathrm{d}\alpha}|_{\alpha=1} & = & 0; \\ \frac{\mathrm{d}\Pi_{\min\, s^0}}{\mathrm{d}\alpha}|_{\alpha=1} & > & 0; \\ \frac{\mathrm{d}\Pi_{\mathrm{FI}}}{\mathrm{d}\alpha}|_{\alpha=1} & > & 0; \\ \frac{\mathrm{d}\Pi_{s^1}}{\mathrm{d}\alpha}|_{\alpha=1} & = & 0. \end{array}$$

Moreover, $\frac{\mathrm{d}\Pi_{\min\, s^0}}{\mathrm{d}\alpha}|_{\alpha=1} > 0$ if and only if $\Pi_{\min\, s^1}$ attains its minimum at $\tilde{\alpha} < 1$; using the expression for the derivative of $\Pi_{\min\, s^1}$ with respect to α, we have $\tilde{\alpha} < 1$ if and only if

$$2v_1 \cdot (1 - Q) - c_1 < 2Qv_0 - c_0.$$

The proof then follows from a standard continuity argument. □

C.8 Proof of Proposition 4.6

Let $q, q' \in (\frac{1}{2}, 1)$ be such that $q' > q$, and let $\mathcal{S} = (1, q, \delta)$ and $\mathcal{S}' = (1, q', \delta)$ be (culturally homogeneous) organizations with culture strength q and q', respectively. Write Q_{in} (resp. Q'_{in})

for $Q_{in}(q) = q^2 + (1-q)^2$ (resp. $Q_{in}(q') = (q')^2 + (1-q')^2$). Note that $Q'_{in} > Q_{in}$ so that each of the regimes in Proposition 4.6 is nonempty. Also note that in culturally homogeneous societies, strategic behavior or payoffs do not depend on cultural distance δ (or Q_{out}), so we drop these variables from our notation here, trusting that no confusion will result.

We first prove (a) and (b). By Theorem 3.1, for $\rho > Q'_{in}$, there is a unique introspective equilibrium for each organization, and for each organization, all players choose s^0 in introspective equilibrium. Likewise, $\rho < 1 - Q'_{in}$, there is a unique introspective equilibrium for each organization, and for each organization, all players choose s^1 in introspective equilibrium. This yields (a) and (b).

We next turn to (c). We start with (c1). By Theorem 3.1, if $\rho \in (1 - Q'_{in}, 1 - Q_{in})$, then there is a unique introspective equilibrium for each organization. The introspective equilibria differ across organizations, though: In $\mathcal{S}'$, all players follow their impulse in introspective equilibrium, while in $\mathcal{S}$, all players choose s^1. The result then follows by noting that the total payoff is maximized when all players coordinate on the efficient action, that is, for payoff parameters v_0, v_1, c_0, c_1 such that $\rho \in (1 - Q'_{in}, 1 - Q_{in})$, $\Pi_{s^1} = (v_0, v_1, c_0, c1; 1, Q_{in}) > \Pi_{\text{FI}}(v_0, v_1, c_0, c1; 1, , Q'_{in})$.

We next turn to (c2). By Theorem 3.1, if $\rho \in (1 - Q_{in}, Q_{in})$, then there is a unique introspective equilibrium for each organization, and for each organization, all players follow their impulse in equilibrium. The result then follows by noting that the total payoff for this class of introspective equilibria is increasing in culture strength. That is, for payoff parameters v_0, v_1, c_0, c_1 such that $\rho \in (1 - Q_{in}, Q_{in})$, $\Pi_{\text{FI}}(v_0, v_1, c_0, c1; 1, Q'_{in}) > \Pi_{\text{FI}}(v_0, v_1, c_0, c1; 1, Q_{in})$.

Finally, consider (c3). By Theorem 3.1, if $\rho \in (Q_{in}, Q'_{in})$, then there is a unique introspective equilibrium for each organization. The introspective equilibria differ across organizations, though: In $\mathcal{S}'$, all players follow their impulse in introspective equilibrium, while in $\mathcal{S}$, all players choose s^1. Then, $\Pi_{s^0}(v_0, v_1, c_0, c1; 1, Q_{in}) > \Pi_{\text{FI}}(v_0, v_1, c_0, c1; 1, Q'_{in})$ if and only if $Q'_{in} < \frac{2v_0 - c_0 + c_1}{v_0 + v_1}$. □

References

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.

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

Arrow, K. J. (1974). *The Limits of Organizations*. W.W. Norton and Co.

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

Bacharach, M. and M. Bernasconi (1997). The variable frame theory of focal points: An experimental study. *Games and Economic Behavior 19*, 1–4.

Banerjee, A. V. and E. Duflo (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

Bardhan, P. (1997). Corruption and development: A review of issues. *Journal of Economic Literature 35*, 1320–1346.

Bardsley, N., J. Mehta, C. Starmer, and R. Sugden (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. *Economic Journal 120*, 40–79.

Blume, A. (2000). Coordination and learning with a partial language. *Journal of Economic Theory 95*, 1–36.

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

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

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.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1990). Selection criteria in coordination games: Some experimental results. *American Economic Review 80*, 218–233.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992). Communication in coordination games. *Quarterly Journal of Economics 107*, 739–771.

Costa-Gomes, M. A. and V. P. Crawford (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review 96*, 1737–1768.

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., 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.

Crawford, V. P. and H. Haller (1990). Learning how to cooperate: Optimal play in repeated coordination games. *Econometrica 58*, 571–595.

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

Curry, O. and M. Jones Chesters (2012). Putting ourselves in the other fellow's shoes: The role of 'theory of mind' in solving coordination problems. *Journal of Cognition and Culture 12*, 147–159.

Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations*, pp. 49–72. Blackwell.

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.

Fudenberg, D. and D. K. Levine (1999). *The Theory of Learning in Games*. MIT Press.

Goeree, J. K. and C. A. Holt (2004). A model of noisy introspection. *Games and Economic Behavior 46*, 365–382.

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

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.

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.

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

Hermalin, B. E. (2013). Leadership and corporate culture. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*, pp. 432–478. Princeton University Press.

Hoff, K. and A. Sen (2006). The kin system as a poverty trap? In S. Bowles, S. Durlauf, and K. Hoff (Eds.), *Poverty Traps*. Princeton University Press.

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

Jackson, M. O. and Y. Xing (2014). Culture-dependent strategies in coordination games. *Proceedings of the National Academy of Sciences 111*, 10889–10896.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. New York: Vintage Books.

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

Kets, W. and A. Sandroni (2015). A belief-based theory of homophily. Working paper, Northwestern University.

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.

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

Lazear, E. P. (1999). 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.

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

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

Mehta, J., C. Starmer, and R. Sugden (1994). The nature of salience: An experimental investigation of pure coordination games. *American Economic Review 84*, 658–673.

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.

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.

Myerson, R. B. (1994). Communication, correlated equilibria and incentive compatibility. Volume 2 of *Handbook of Game Theory with Economic Applications*, Chapter 24, pp. 827–847. Elsevier.

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

Nisbett, R. E. and D. Cohen (1996). *Culture Of Honor: The Psychology Of Violence In The South*. Westview Press.

Ochs, E. (1988). *Culture and language development*. Cambridge University Press.

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

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

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

Schmidt, D., R. Shupp, J. M. Walker, and E. Ostrom (2003). Playing safe in coordination games: The roles of risk dominance, payoff dominance, and history of play. *Games and Economic Behavior 42*, 281–299.

Sewell, W. (1992). A theory of structure: Duality, agency, and transformation. *American Journal of Sociology 98*, 1–29.

Shrivastava, P. (1986). Postmerger integration. *Journal of Business Strategy 7*, 65–76.

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.

Straub, P. G. (1995). Risk dominance and coordination failures in static games. *Quarterly Review of Economics and Finance 35*, 339–363.

Sugden, R. (1993). Thinking as a team: Toward an explanation of nonselfish behavior. *Social Philosophy and Policy 10*, 69–89.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.

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

Van Huyck, J. B., R. C. Battalio, and R. O. Beil (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review 80*, 234–248.

Waldfogel, J. (1993). The deadweight loss of Christmas. *American Economic Review 83*, 1328–1336.

Weber, R. A. and C. F. Camerer (2003). Cultural conflict and merger failure: An experimental approach. *Management Science 49*, 400–415.

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