DISCUSSION PAPER SERIES

DP16229

The Value of a Coordination Game

Willemien Kets, Wouter Kager and Alvaro Sandroni

ORGANIZATIONAL ECONOMICS

CEPR

ISSN 0265-8003

The Value of a Coordination Game

Willemien Kets, Wouter Kager and Alvaro Sandroni

Discussion Paper DP16229
Published 08 June 2021
Submitted 03 June 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Organizational Economics

The Value of a Coordination Game

Abstract

The value of a game is the payoff a player can expect (ex ante) from playing the game. Understanding how the value changes with economic primitives is critical for policy design and welfare. However, for games with multiple equilibria, the value is difficult to determine. We therefore develop a new theory of the value of coordination games. The theory delivers testable comparative statics on the value and delivers novel insights relevant to policy design. For example, policies that shift behavior in the desired direction can make everyone worse off, and policies that increase everyone's payoffs can reduce welfare.

JEL Classification: N/A

Keywords: N/A

Willemien Kets - willemien.kets@economics.ox.ac.uk
University of Oxford and CEPR

Wouter Kager - w.kager@vu.nl
VU University Amsterdam

Alvaro Sandroni - sandroni@kellogg.northwestern.edu
Kellogg School of Management, Northwestern University

Acknowledgements
We are grateful to the Associate Editor and two anonymous referees for excellent suggestions. We thank Larbi Alaoui, Miguel Ballester, Larry Blume, Vincent Crawford, David Gill, Meg Meyer, Antonio Penta, David Schmeidler, Jakub Steiner, Colin Stewart, and numerous seminar audiences for helpful comments and stimulating discussions. Luzia Bruckamp provided excellent research assistance.

The Value of a Coordination Game*

Willemien Kets† Wouter Kager‡ Alvaro Sandroni§

June 3, 2021

Abstract

The value of a game is the payoff a player can expect (ex ante) from playing the game. Understanding how the value changes with economic primitives is critical for policy design and welfare. However, for games with multiple equilibria, the value is difficult to determine. We therefore develop a new theory of the value of coordination games. The theory delivers testable comparative statics on the value and delivers novel insights relevant to policy design. For example, policies that shift behavior in the desired direction can make everyone worse off, and policies that increase everyone's payoffs can reduce welfare.

*We are grateful to the Associate Editor and two anonymous referees for excellent suggestions. We thank Larbi Alaoui, Miguel Ballester, Larry Blume, Vincent Crawford, David Gill, Meg Meyer, Antonio Penta, David Schmeidler, Jakub Steiner, Colin Stewart, and numerous seminar audiences for helpful comments and stimulating discussions. Luzia Bruckamp provided excellent research assistance.

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

‡Department of Mathematics, Vrije Universiteit Amsterdam. E-mail: w.kager@vu.nl.

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

> [In a coordination game, players] must somehow *use* the labeling of [actions] in order to do better than pure chance; and how to use it may depend more on imagination than on logic, more on poetry or humor than on mathematics. It is noteworthy that traditional game theory does not assign a "value" to this game: how well people can concert in this fashion is something that, though hopefully amenable to systematic analysis, cannot be discovered by reasoning a priori.
>
> Schelling (1960, pp. 97–98)

1 Introduction

The value of a game is the payoff a player can expect (ex ante) from playing the game. Understanding how the value changes with economic primitives is critical for policy evaluation and design. For example, when designing institutions, a planner needs to know the payoff that players can expect to receive when they interact under different institutional constraints. Likewise, before introducing a new policy, a policy maker would want to know how it changes players' welfare. In settings where economic primitives uniquely pin down behavior, defining the value of a game is simple. To give an example, in standard principal-agent problems, it is straightforward to determine what a contract is worth to both parties. However, for games with multiple equilibria, the value is often difficult to determine theoretically. For example, a policy change that leaves the set of Nash equilibria unchanged may still affect the value if it changes the way the game is played (e.g., by making an unplayed equilibrium more attractive in terms of payoffs). But while there is an extensive literature on how *equilibrium behavior* changes with payoffs in games with multiple equilibria (e.g., Milgrom and Roberts, 1990; Milgrom and Shannon, 1994; Athey, 2002; Echenique, 2002; Vives, 2005), the question of how the *value* changes with economic primitives is largely open.[1] This is problematic: While the literature shows that equilibrium behavior is monotone in payoffs for a large class of games with multiple equilibria (supermodular games),[2] this does not imply that the value will also change monotonically (Angeletos and Pavan, 2004). This means that *a policy that shifts behavior in the desired direction may make everyone worse off*. Thus, it is critical to understand how the value varies with economic primitives in games with multiple equilibria.

This paper takes a first step in this research program by deriving testable comparative statics on the value of coordination games (a subclass of supermodular games). A key difficulty is that

[1] An exception is Theorem 7 in Milgrom and Roberts (1990) which gives sufficient conditions under which one pure Nash equilibrium is Pareto-preferred over another in a game with strategic complementarities. It does not provide comparative statics. Crawford and Smallwood (1984) provide comparative statics on the value for zero-sum games; however, their methods do not extend to non-zero sum games like the ones studied here.

[2] That is, the *set* of equilibria changes monotonically with payoffs (e.g., Milgrom and Roberts, 1990). The same is true for stable equilibria (Echenique, 2002).

when games have multiple equilibria, the payoff structure does not uniquely pin down behavior. Nevertheless, we often have clear intuitions about what the value should be. For example, consider the following game, denoted G_w:

	s^1	s^2
s^1	w, w	$0, 0$
s^2	$0, 0$	$1, 1$

with $w \geq 1$. The following predictions seem intuitive: (p1) If w is sufficiently large, then the value of the game G_w equals w; (p2) If w equals 1, then the value of G_w lies strictly between $\frac{1}{2}$ and 1. Intuitively, if w is sufficiently large, action s^1 stands out in terms of payoffs. The Nash equilibrium in which both players choose s^1 thus becomes the unique focal point, and the value equals w. This is so intuitive that whenever $w > 1$ the literature generally restricts attention to the (s^1, s^1) equilibrium on the grounds that it is the unique Nash equilibrium that might plausibly be played (Harsanyi and Selten, 1988).[3] By contrast, if $w = 1$, the payoff structure of the game does not give any guidance on how to play. However, oftentimes, one of the action labels somehow "stands out," e.g., through associations that connect those labels to some aspect of players' common experience or culture. In such cases, we would expect players to be more successful at coordinating than they would have been had they chosen their actions independently. At the same time, because players almost never have been exposed to exactly the same cultural and social patterns, coordination is likely to be imperfect. Thus, we would expect players to do better than in mixed Nash equilibrium (which gives a value of $\frac{1}{2}$) but less well than in pure Nash equilibrium (which gives a value of 1). Consistent with this prediction, Mehta et al. (1994, p. 668) show experimentally that if one of the actions has a salient label, then the value of G_w with $w = 1$ is 0.76.

While predictions (p1) and (p2) are intuitive, standard solution concepts all fail to make at least one of these predictions uniquely (i.e., as the only possible outcome) or fail to make both of them (see Section 5.5 for details). Motivated by this, we develop a novel theory of the value for symmetric (2×2) coordination games.[4] A key insight is that the comparative statics of the value are driven by the tradeoff between two basic coordination problems that players face: (1) How to coordinate on some Nash equilibrium (i.e., avoid *miscoordination*); and (2) How to coordinate on the "right" Nash equilibrium (i.e., avoid *coordination failure*). To capture this tradeoff we build on our earlier work (Kets and Sandroni, 2019, 2021). This earlier

[3] Indeed, we are not aware of experimental papers that study games like G_w for $w > 1$. This is consistent with our argument: If there is an obvious way to play, then there is no reason to study the game experimentally. See Schmidt et al. (2003, p. 285) for comments along these lines.

[4] Focusing on simple stylized games allows us to obtain the cleanest insights. However, our methods can be extended to more general coordination games.

work introduced a behavioral solution concept, *introspective equilibrium*, that allows for both economic and non-economic factors to influence behavior. This makes it possible to understand how the relative importance of miscoordination and coordination failure varies with economic factors (i.e., payoff parameters). When payoff considerations dominate any non-economic factors (as in G_w for w large), there is no miscoordination, but there can be coordination failure. By contrast, when the payoff structure provides little guidance (as in G_w for $w = 1$), there can be miscoordination; however, because non-economic factors can help players coordinate, the value exceeds the expected payoff in the mixed Nash equilibrium. Hence, this novel theory is able to deliver the intuitive predictions (p1) and (p2) (Section 3.2.1).

We can go well beyond this to formalize other intuitions that cannot be captured by standard theory. Consider, for instance, the following variant of G_w, denoted by $\widetilde{G}_x$:

	s^1	s^2
s^1	w, w	$-c, 0$
s^2	$0, -c$	$1+x, 1+x$

where $x, c \geq 0$ and $x + 1 < w$. We show that if w is sufficiently high, then the value of $\widetilde{G}_x$ is *lower* when $1 + x$ is close to $w > 1$ than if it is close to 1. So, the value of the game can fall when all payoffs increase (weakly) (Section 3.2.1). Intuitively, there is both a direct and an indirect effect. The *direct (payoff) effect* means that as x increases, players who coordinate on s^2 receive a higher payoff. The *indirect (strategic) effect* means that a change in x changes the way the game is played: If w is sufficiently high and x is close to 0, then players choose s^1 and the value of the game is $w > 1$. When the unplayed equilibrium (s^2, s^2) becomes more attractive in terms of payoffs (i.e., x close to $w - 1$), there is miscoordination. This leads to a reduction in value even though all payoffs are weakly higher. This insight cannot be formalized using traditional game-theoretic models, as an increase in x does not change the set of Nash equilibria.[5] However, it is critical for understanding why policy changes that create direct benefits may ultimately reduce welfare. For example, if labor supply decisions are strategic complements, an increase in unemployment benefits (a positive direct effect) may make workers worse off if it induces miscoordination (a negative strategic effect) (Lindbeck et al., 1999).

As illustrated by these simple examples, a theory of the value of coordination games can provide nontrivial welfare implications that cannot be derived with standard game-theoretic models. We therefore develop a general theory of the value of coordination games. Our first main result (Theorem 1) focuses on the comparative statics of the value as a function of a key parameter, the parameter $\rho = (u_{22} - u_{12})/(u_{11} - u_{21} + u_{22} - u_{12})$ that summarizes the best-response

[5] Standard refinements are also unable to formalize this intuition: Payoff dominance selects (s^1, s^1); and if c is not too large, the same is true for risk dominance. See Section 5.5 for further discussion.

correspondence (where u_{nm} is the payoff to a player who chooses s^n when the other player chooses s^m). Theorem 1 is an important starting point in that the parameter ρ fully characterizes the comparative statics for equilibrium behavior.[6] However, as Theorem 1 demonstrates, the value typically depends on all payoff parameters (i.e., $u_{11}, u_{22}, u_{21}, u_{12}$), not just the parameter ρ. This means that making more specific predictions requires considering specific applications. We therefore consider various applications, selected to showcase the breadth of economic insights that can be obtained through studying the comparative statics on the value.

Our first economic application addresses the key question whether policies to stimulate investment are necessarily welfare improving. Consider a setting where players can choose whether to invest or not, with both full investment and no investment being Nash equilibria. Full investment is Pareto optimal, and players who invest receive a subsidy $s > 0$ regardless of whether the other player invests. Even though there is no apparent downside to this policy (because it (weakly) increases the payoffs to any action profile), our results show that this policy may reduce players' welfare by creating miscoordination (Theorem 2). This is true even though the policy destabilizes a Pareto-inferior equilibrium (i.e., reduces coordination failure). These insights apply more widely. For example, in societies that rely on informal contracts, strengthening judicial enforcement may be counterproductive if this destabilizes cooperation, even if it leads to an expansion of the formal sector, i.e., it eliminates coordination failure (Dixit, 2004, Ch. 2). Thus, policies that shift behavior in the desired direction may make everyone worse off. Because our results provide a full characterization of the conditions under which a policy has a net positive or negative effect on the value, the theory also shows how to implement the policy in a way that ensures it has the desired welfare effects. Hence, the theory not only points to the limitations inherent in focusing on comparative statics on equilibrium behavior in policy design, but also offers guidance on how to ensure that a desired change in behavior also leads to a welfare improvement.

The second economic application we consider is to collusion. It has been argued that the risk of miscoordination renders tacit collusion impracticable so that policy makers need not worry about tacit collusion and can focus on deterring explicit collusion (e.g., Motta, 2004, p. 190). To examine this claim, we adopt a simple reduced-form model of collusion, where a lack of collusion corresponds to coordination failure and where tacit collusion can lead to miscoordination. Perhaps surprisingly, it turns out that it generally pays for firms to attempt to collude: In terms of our model, miscoordination tends to be less costly than coordination failure (Theorem 3). This suggests that industry lobbies have an incentive to lobby for changes that make collusion more attractive, such as improving the ease of detection or increasing the

[6] This holds for any solution concept that assumes that players choose best responses, not just introspective equilibrium.

frequency of interaction (Ivaldi et al., 2003). Understanding when parties have an incentive to change their economic environment requires doing comparative statics on the value; comparative statics on equilibrium behavior, while providing valuable insights into when collusion can be sustained, does not provide any insight into this key question.

The main contribution of our paper is that it is one of the very few papers that provides comparative statics on the value of coordination games. Relative to the existing literature which focuses largely on comparative statics on equilibrium behavior, this delivers important new insights. As we show, comparative statics on the value are crucial for understanding how and when policies that shift behavior in the desired direction improve or reduce welfare. They are also crucial for understanding how and when policies that provide only benefits in the absence of strategic interactions may ultimately reduce welfare, and for when parties have an incentive to influence their economic environments.

To develop this theory of the value, we build on our earlier work (Kets and Sandroni, 2019, 2021). Although the present paper uses the solution concept (introspective equilibrium) introduced in our earlier work, there are several fundamental differences between this paper and our earlier work. First and foremost, Kets and Sandroni (2019, 2021) focus on different questions and do not provide any comparative statics on the value.[7] As a consequence, none of our main results (Theorem 1–3) are suggested by or follow from our earlier work; the same is true for the intuitive predictions for G_w and $\widetilde{G}_x$ (Propositions 2–3). A smaller but still substantial contribution relative to our earlier work is that we use an axiomatic approach when defining introspective equilibrium in the current paper (Section 2.2). By contrast, Kets and Sandroni (2019, 2021) use a simple parametric model that is a special (limiting) case of the current framework. The current axiomatic approach not only makes the results stronger (since the results hold across all environments that satisfy our axioms), it also elucidates the main drivers behind the results.

Importantly, our predictions are testable even if the relevant non-economic factors are unobservable. This is because we focus on comparative statics: Our results predict how the value changes with payoffs (e.g., an increase in w in G_w or x in $\widetilde{G}_x$) for a broad range of assumptions on the non-economic factors that govern behavior. Because we obtain the same qualitative comparative statics on the value *regardless* of the exact assumptions on non-economic factors, our predictions can be tested even if the relevant non-economic factors cannot be observed. By contrast, much of the behavioral game theory literature estimates the relevant behavioral parameters from data, with the notable exception of Alaoui and Penta (2016) and Alaoui et al.

[7] Kets and Sandroni (2021) study the economic impact of diversity. Diversity is taken to affect what we call introspective type spaces here, and the focus is on how diversity affects behavior, with some limited results on the welfare implications of diversity (Propositions 3.4–3.5). Diversity increases the scope for miscoordination but may reduce coordination failure. Kets and Sandroni (2019) only consider miscoordination. Neither of these earlier papers provides comparative statics on the value.

(2020) who provide testable comparative statics for level-k models and test their predictions experimentally.[8] Instead of assuming that the relevant parameters can be estimated from the data, we focus on predictions that are *independent* of the precise behavioral parameters.

The outline of this paper is as follows. Section 2 introduces the model and presents an auxiliary result (Proposition 1). Section 3 presents the main results. Section 4 presents a dynamic application. Sections 5 discusses the model and connects our results to the related literature. Section 6 concludes. Omitted proofs can be found in the appendix.

2 Model

2.1 Coordination

We consider symmetric (2×2) coordination games. There are two actions, s^1 and s^2. Payoffs are given by

	s^1	s^2
s^1	u_{11}, u_{11}	u_{12}, u_{21}
s^2	u_{21}, u_{12}	u_{22}, u_{22}

where $u_{11} > u_{21}$, $u_{22} > u_{12}$, and $u_{11} \geq u_{22}$. Thus, both (s^1, s^1) and (s^2, s^2) are strict Nash equilibria and the equilibrium in which both players choose s^1 is (weakly) Pareto dominant. All payoff parameters are common knowledge.

As is well-known, the best-response correspondence can be summarized by the parameter

$$\rho := \frac{u_{22} - u_{12}}{u_{11} - u_{21} + u_{22} - u_{12}}.$$

Action s^1 is a best response for a player if and only if it assigns probability at least ρ to the other player choosing s^1. That is, (s^1, s^1) is ρ-dominant in the sense of Morris et al. (1995) (also see Selten, 1995). We therefore refer to ρ as the *dominance parameter*. It is obviously closely related to risk dominance: (s^1, s^1) is risk dominant if $\rho < \frac{1}{2}$, and (s^2, s^2) is risk dominant if $\rho > \frac{1}{2}$.

2.2 Introspection

As coordination games have multiple equilibria, players face considerable strategic uncertainty, that is, uncertainty about the other player's action. As observed by Schelling (1960), when a player is uncertain about another player's action, "[the] objective is to make contact with

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

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* (Apperly, 2012). Theory of mind is a central concept in psychology and refers to the cognitive ability to take another person's perspective. Following Kets and Sandroni (2021), we model this perspective-taking process as follows: Each player has an impulse to choose an action. Each player's first instinct is to follow his impulse. But, through introspection, players realize that the other player also has an impulse. This may lead them to adjust their response. This process continues to higher levels until no player wishes to adjust his choice.

To formally model this introspective process we need to model players' beliefs. We do so using type spaces, as is standard; to emphasize that our type spaces encode not just beliefs but also players' impulses, we refer to them as *introspective type spaces*. That is, each player $j \in \{1,2\}$ has an introspective type $t_j \in T := [0,1]$, drawn from a common prior on $T \times T$ with distribution function $F(t_1,t_2)$. Each introspective type $t_j \in T$ is associated with an *impulse* $\mathcal{I}_j(t_j) \in \{s^1,s^2\}$. The functions $\mathcal{I}_j(\cdot)$ that map types into impulses are common knowledge. Players know their own impulse (i.e., the impulse functions are measurable) but not the other player's impulse. A player's first instinct is to follow his impulse. This defines his *level-0 strategy* σ_j^0 (i.e., $\sigma_j^0(t_j) = \mathcal{I}_j(t_j)$). For any $k > 0$, the *level-k strategy* σ_j^k for each player j is a best response to the level-$(k-1)$ strategy σ_{-j}^{k-1} of the other players.[9] A player's behavior is given by the limit $\sigma_j := \lim_{k\to\infty} \sigma_j^k$ of the introspective process. If these limiting strategies exist, then $\sigma = (\sigma_j)_j$ is an *introspective equilibrium*.[10]

Introspective equilibrium thus models situations where an iterative perspective-taking process allows players to reach consistent expectations (i.e., the introspective process converges). This makes the model suitable for describing situations where initial beliefs may not be consistent (i.e., $\sigma^0 \neq \sigma$) but where the social context (e.g., shared culture, salient action labels) can help players reach consistent expectations. Introspective equilibrium is thus a good model for situations where the assumption from Nash or correlated equilibrium that players' initial expectations are correct (i.e., $\sigma^0 = \sigma$) is perhaps too strong, yet the social context guides players' initial expectations in a way that allows them to reach consistent expectations.

[9] 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.

[10] As the terminology suggests, the introspective process bears some resemblance to level-k and cognitive hierarchy models (see Nagel (1995), Stahl and Wilson (1995), Costa-Gomes et al. (2001), and Camerer et al. (2004) for early references, and see Crawford et al. (2013) for a survey). The key distinction is that introspective equilibrium uses impulses to model the effects of non-economic factors. Another difference is that introspective equilibrium does not presume that players are boundedly rational. This is to emphasize that results are not driven by bounded rationality. However, this is not critical: Our results are robust to relaxing the assumption that $k \to \infty$.

The introspective type space models players' impulses as well as their beliefs about the other player's impulses, their beliefs about the other's beliefs, and so on. Since such beliefs are typically difficult to measure, we focus on comparative statics that hold across a large class of introspective type spaces. We consider any introspective type space that satisfies the following conditions:

(SYM) Players are ex ante identical. That is, players have the same impulse function (i.e., $\mathcal{I}_1 = \mathcal{I}_2$) and the cumulative distribution function F induced by the common prior is symmetric (i.e., $F(y,z) = F(z,y)$).

(MON-I) Impulses are monotone in type: There is a threshold $\tau^0 \in (0,1)$ such that for each player j, a introspective type t has an impulse to choose s^1 (i.e., $\mathcal{I}_j(t) = s^1$) if and only if $t \geq \tau^0$.

(MON-B) Beliefs are monotone in type: The conditional probability $F(\tau \mid t)$ that the other player has an introspective type at most τ (given the player's type t) is strictly decreasing in t.

(REG) The common prior $F(\cdot)$ has a continuous density f with full support on the interior of $T \times T$ and the limits $\lim_{t \downarrow 0} F(t \mid t)$ and $\lim_{t \uparrow 1} F(t \mid t)$ exist.

Assumptions (MON-I) and (MON-B) ensure that the game is a monotone supermodular game (Vives, 2005; Van Zandt and Vives, 2007). They imply that high types think it likely that the other player has a high type and thinks that the other player has a high type, and so on. In particular, players with an impulse to choose action s thinks it is likely that the other player has an impulse to choose s. For example, if an action label is salient to a player, then he would typically expect it to be salient to the other player as well. Assumption (SYM) captures the idea that players have symmetric roles (i.e., are ex ante identical), and Assumption (REG) is a technical regularity condition.[11] In the remainder of the paper, we will assume (SYM)–(REG); an introspective type space can then be summarized by its distribution function $F(\cdot)$ and the threshold τ^0. We thus write $\mathcal{T} = (F, \tau^0)$ for an introspective type space.

For some of our results, we require an additional assumption. Let the *rank belief* of an introspective type t in an introspective type space (F, τ^0) be the probability $F(t \mid t)$ that the type assigns to the other player having a lower type than itself (i.e., $t_{-j} \leq t$) (e.g., Morris et al., 2016). Then, an introspective type space (F, τ^0) induces *non-monotone rank beliefs* if it satisfies the following condition:

(NMRB) There exists $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$ or there exists $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both).

[11] For example, all conditional distributions are well-defined; see, e.g., Grimmett and Stirzaker (2020).

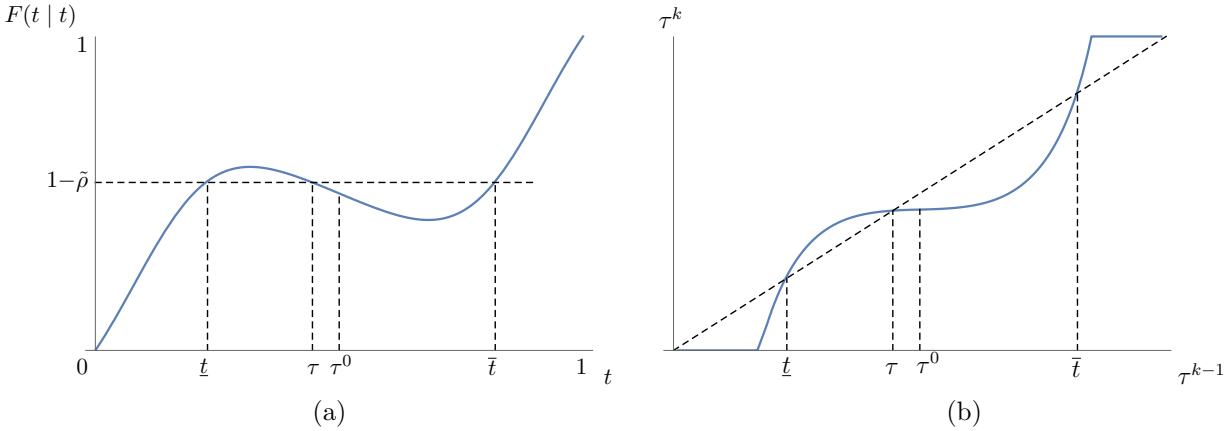

Figure 1: (a) The rank belief function $F(t \mid t)$ for an introspective type spaces that induces non-monotone rank beliefs; (b) The corresponding best response function for dominance parameter $\tilde{\rho}$.

As its name suggest, this condition ensures that the rank belief function $F(\cdot \mid \cdot)$ is non-monotone (by Lemma 3 in the appendix). Figure 1(a) illustrates the condition for an example type space. Assumption (NMRB) is more novel and therefore less well understood than the other conditions. We therefore first illustrate its implications and properties in the context of concrete applications (Sections 3.2.1 and 3.2.2) before discussing it in more abstract terms in Section 5.

2.3 Preliminary results

Before presenting the main results, we briefly summarize the basic properties of introspective equilibrium.[12]

Proposition 1. [Introspective Equilibrium]

(a) Every introspective equilibrium induces a correlated equilibrium.

(b) Every coordination game has an introspective equilibrium, and it is essentially unique.[13]

(c) The introspective equilibrium is in monotone strategies: there is a threshold $\tau \in T$ such that, in introspective equilibrium, any introspective type $t < \tau$ chooses s^2 and any introspective type $t > \tau$ chooses s^1.

(c1) If s^1 is a best response for type τ^0 at level 1 (i.e., $F(\tau^0 \mid \tau^0) \leq 1 - \rho$), then the equilibrium threshold is the largest $\tau \leq \tau^0$ such that $F(\tau \mid \tau) = 1 - \rho$ if such a τ exists, and $\tau = 0$ otherwise.

[12]Parts (a), (b), and (d) were proven in Kets and Sandroni (2021), albeit for a slightly different class of introspective type spaces.

[13]By "essentially unique," we mean that introspective equilibrium uniquely determines behavior for all but a measure-0 set of types.

(c2) If s^2 is a best response for type τ^0 at level 1 (i.e., $F(\tau^0 \mid \tau^0) \geq 1-\rho$) then the equilibrium threshold is the smallest $\tau \geq \tau^0$ such that $F(\tau \mid \tau) = 1-\rho$ if such a τ exists, and $\tau = 1$ otherwise.

(d) Introspective equilibrium is monotone in payoffs: For any introspective type space, the probability that a player chooses action s in introspective equilibrium (weakly) increases with the payoffs to that action (holding other payoffs fixed).

(e) There is miscoordination (i.e., $\tau \in (0,1)$) for an open set $(\underline{\rho}, \overline{\rho})$ of dominance parameters if and only if the introspective type space induces non-monotone rank beliefs (i.e., satisfies (NMRB)).

The proof can be found in the appendix. Proposition 1 establishes that *introspective equilibrium satisfies the minimal properties that any good theory of coordination should have.* By the characterization of correlated equilibrium of Aumann (1987), part (a) implies that introspective equilibrium is consistent with common knowledge of rationality. So, introspective equilibrium is consistent with standard rationality assumptions despite its psychological foundations. However, since correlated equilibria need not be Nash equilibria, introspective equilibrium allows for non-Nash behavior. This will be critical for characterizing the costs of miscoordination. Parts (b)–(c) show that introspective equilibrium makes sharp predictions and is fully characterized by a simple threshold. Thus, when players face identical decision problems (i.e., players have a coordination motive and payoffs are symmetric) and expect others to have similar beliefs (by (MON-B) and (MON-I)), the social context (e.g., shared culture) can help players reach consistent expectations.[14] The uniqueness result will allow us to derive testable hypotheses on how the value depends on the economic environment (i.e., payoff parameters). Part (d) shows that introspective equilibrium has intuitive properties: when the payoffs to an action increase, players are more likely to choose that action, consistent with experimental evidence (e.g., Straub, 1995; Schmidt et al., 2003). Part (e) shows that introspective equilibrium allows for miscoordination: If an introspective type space induces non-monotone rank beliefs, then for some values of the dominance parameter, players choose both actions with positive probability (i.e., $\tau \in (0,1)$).

Proposition 1 also serves *distinguishes introspective equilibrium from other solution concepts*: no other solution concept that we are aware of satisfies all the properties listed in Proposition 1. For example, Nash and correlated equilibrium do not yield a unique prediction for coordination games (i.e., fail (b)), while concepts that select a unique equilibrium (such as global games) do not allow for miscoordination (and thus fail an analogue of (e)). And while mixed Nash allows

[14] However, when players face different decision problems (as, e.g., in zero sum games or battle of the sexes), it is less clear that introspection leads to consistent expectations, at least without a richer theory of mind.

for miscoordination, its comparative statics are the opposite of (d): mixed Nash equilibrium predicts that players are *less* likely to choose an action when its payoffs increase.

To see the intuition behind Proposition 1(b)–(c), suppose that action s^1 is a strict best response for τ^0 at level 1, that is,

$$F(\tau^0 \mid \tau^0)\, u_{12} + (1 - F(\tau^0 \mid \tau^0))\, u_{11} > F(\tau^0 \mid \tau^0)\, u_{22} + (1 - F(\tau^0 \mid \tau^0))\, u_{21},$$

or, equivalently, $F(\tau^0 \mid \tau^0) < 1 - \rho$. By (MON-B), there is a unique $\tau^1 < \tau^0$ such that, at level 1, each type chooses s^1 if $t > \tau^1$ and chooses s^2 if $t < \tau^1$ (i.e., τ^1 solves $F(\tau^0 \mid \tau^1) = 1 - \rho$ or $\tau^1 = 0$). By a simple inductive argument, for each $k > 0$, there is a (unique) level-k threshold τ^k such that types $t < \tau^k$ choose s^2 at level k, and types $t > \tau^k$ choose s^1 (i.e., τ^k solves $F(\tau^{k-1} \mid \tau^k) = 1 - \rho$ or $\tau^k = 0$). Moreover, the thresholds are decreasing in k (i.e., $\tau^k \leq \tau^{k-1}$). Intuitively, a player chooses s^1 at level k if he thinks it likely that the other player has an impulse to choose s^1 (and thus chooses s^1 at level 0) or that the other player thinks it is likely that the other has an impulse to choose s^1 (and thus chooses s^1 at level 1), and so on, up to level $k-1$. By the monotone convergence theorem, the sequence $\{\tau^k\}_k$ converges to a unique threshold $\tau < \tau^0$, where τ satisfies $F(\tau \mid \tau) = 1 - \rho$ or $\tau = 0$; an analogous argument applies when $F(\tau^0 \mid \tau^0) \geq 1 - \rho$. Thus, an introspective equilibrium exists, and it uniquely determines behavior for every $t \neq \tau$.

To see the intuition behind Proposition 1(e), consider the best-response function in Figure 1(b). The best-response function maps the level-$(k-1)$ threshold τ^{k-1} into the level-k threshold τ^k for the dominance parameter $\tilde{\rho}$ in Figure 1(a). By (MON-B), the best-response function is increasing, and by the above argument, we have $\tau^k < \tau^{k-1}$ whenever $\tau^{k-1} \in (\tau, \bar{t})$. So, the best-response function lies below the diagonal for $t \in (\tau, \tau^0)$. By an analogous argument, the best-response function lies above the diagonal whenever $\tau^{k-1} \in (\underline{t}, \tau)$. So, the best-response function intersects the diagonal from above at τ, and τ is a stable (attracting) fixed point. Condition (NMRB) ensures that such a stable interior fixed point exists (i.e., there is $\tau < \tau^0$ with $F(\tau \mid \tau) = 1 - \tilde{\rho}$). The intuition behind parts (a) and (d) is less central to our main results and is therefore omitted.

As this discussion suggests, introspective equilibrium depends on both economic and non-economic factors (i.e., τ depends on ρ as well as $\mathcal{T}$). To emphasize this dependence on the introspective type space, we will henceforth refer to $\boldsymbol{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ as the *game form* and to a pair $\mathcal{G} := (\boldsymbol{u}, \mathcal{T})$ consisting of a game form and an introspective type space as a *game* (except that we sometimes use the term "coordination game" rather than the more tedious "coordination game form"). Despite this dependence on non-economic factors, our approach allows for testable comparative statics on the value, as we show next.[15]

[15] This is not the case for Nash or correlated equilibrium, as this requires some form of selection.

3 Main results

This section presents the main results. We characterize how the value changes with payoff parameters. The value of a game is the ex ante expected payoff for a player in introspective equilibrium.[16] Formally, given a game form $\boldsymbol{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ and introspective type space $\mathcal{T} = (F, \tau^0)$, the value of the game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$ is

$$V(\boldsymbol{u}; \mathcal{T}) := \int_{T \times T} u(\sigma_1(t_1), \sigma_2(t_2)) \, \mathrm{d}F(t_1, t_2),$$

where $u(\sigma_1(t_1), \sigma_2(t_2)) \in \{u_{11}, u_{12}, u_{21}, u_{22}\}$ is the (ex-post) payoff that the player receives in introspective equilibrium (σ_1, σ_2) when he has introspective type t_1 and the other player has type t_2. By Proposition 1, the value is well-defined.[17] Because the value depends on all payoff parameters (i.e., $u_{11}, u_{12}, u_{21}, u_{22}$), we consider different types of payoff changes. Section 3.1 focuses on the comparative statics of the value as a function of the dominance parameter ρ which characterizes equilibrium behavior. Section 3.2 complements this by providing comparative statics on the value for different applications.

3.1 Benchmark theoretical result

This section provides testable comparative statics on the value as a function of the dominance parameter. This is a useful starting point, since the dominance parameter fully characterizes the best response correspondence, and thus the comparative statics of introspective equilibrium (as well as other solution concepts determined by the best response correspondence, such as Nash and correlated equilibrium). To state the result, we need some more definitions. Given an introspective type space $\mathcal{T} = (F, \tau^0)$, define

$$1 - \underline{\rho} = \sup\{F(t \mid t) \colon t \in [0, \tau^0]\};$$
$$1 - \overline{\rho} = \inf\{F(t \mid t) \colon t \in [\tau^0, 1]\};$$

see Figure 2 for an illustration. The appendix shows that $0 < \underline{\rho} \leq \overline{\rho} < 1$, and that $\underline{\rho} < \overline{\rho}$ if and only if $\mathcal{T}$ induces non-monotone rank beliefs. We have the following characterization:

[16] Our ex ante definition seems well suited for assessing the welfare implications of policies (as in our applications). In other settings, it may be more appropriate to consider an interim notion, for example, when a player decides whether or not to participate in a game (possibly at a cost) and has some information on the non-economic factors that may guide his and other players' behavior.

[17] As coordination games have a unique introspective equilibrium (Proposition 1(b)), the expected equilibrium payoff for each player is well-defined. Moreover, as the introspective equilibrium is symmetric (Proposition 1(c)), the value is the same for both players.

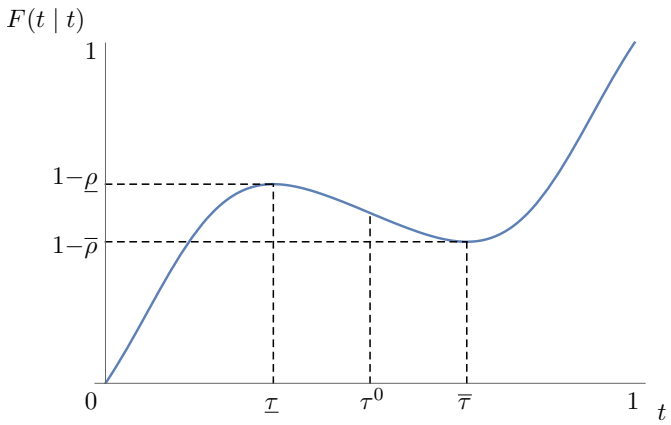

Figure 2: The rank belief function $F(t \mid t)$ from Figure 1 with bounds $\underline{\rho}, \overline{\rho}$.

Theorem 1. [The Value of a Coordination Game] *For extreme values of the dominance parameter ρ, the value is determined by Nash equilibrium, but not otherwise. That is, for any introspective type space $\mathcal{T} = (F, \tau^0)$ and game form $\boldsymbol{u}$ with dominance parameter ρ,*

(a) If $\rho < \underline{\rho}$, both players choose s^1 (i.e., $\tau = 0$) and the value is equal to the payoff in the corresponding Nash equilibrium (i.e., $V(\boldsymbol{u}; \mathcal{T}) = u_{11}$).

(b) If $\rho > \overline{\rho}$, both players choose s^2 (i.e., $\tau = 1$) and the value is equal to the payoff in the corresponding Nash equilibrium (i.e., $V(\boldsymbol{u}; \mathcal{T}) = u_{22}$).

(c) If $\rho \in (\underline{\rho}, \overline{\rho})$, players choose each action with positive probability (i.e., $\tau \in (0,1)$), and, generically, the value is not equal to the expected payoff of any of the Nash equilibria. In fact, the value is given by

$$V(\boldsymbol{u}; \mathcal{T}) = u_{11} + (u_{21} + u_{12} - 2u_{11})F(\tau) + \frac{u_{11} - u_{21}}{1 - \rho} F(\tau, \tau). \tag{1}$$

Theorem 1 has two important but distinct uses. The first use of Theorem 1 is to fully characterize the comparative statics of the value for a given introspective type space. Given any introspective type space $\mathcal{T}$, a researcher can determine the relevant properties of the type space (e.g., $\underline{\rho}, \overline{\rho}, \tau$) and directly calculate the value of any game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$. Appendix A builds on this and decomposes the comparative statics effects on the value into a direct payoff effect and an indirect strategic effect. These results are useful for environments where the type space is known or relatively easy to measure. The second use of Theorem 1 is perhaps even more important: Theorem 1 provides testable comparative statics on the value as a function of ρ that hold across type spaces. Even if an analyst does not know the introspective type space, he can still be assured that there are thresholds $\underline{\rho}, \overline{\rho}$ for the dominance parameters that delineate the different regimes for the comparative statics. So, if a data set contains games with sufficient variation in the dominance factor ρ, then Theorem 1 predicts that *when one of the actions is relatively attractive in terms of payoffs (i.e., $\rho < \underline{\rho}$ or $\rho > \overline{\rho}$), the value is equal to the payoff in*

the corresponding pure Nash equilibrium; when the payoff structure provides little guidance (i.e., $\rho \in (\underline{\rho}, \overline{\rho})$), the value is not consistent with the (expected) payoff in Nash equilibrium (pure or mixed). Because this prediction holds regardless of the introspective type space, it can be tested even if non-economic factors (i.e., $\mathcal{T}$) cannot be observed.

The intuition behind Theorem 1 is that behavior is determined by the interplay of economic and non-economic factors. For extreme values of the dominance parameter, behavior is driven by payoff considerations and the value is given by the Nash payoff (u_{11} for $\rho < \underline{\rho}$, u_{22} for $\rho > \overline{\rho}$). In contrast, for intermediate values of the dominance parameter, a player's decision may depend on non-economic factors: In that case, a player chooses action s^1 if he thinks it is sufficiently likely that the other player has an impulse to choose s^1 (i.e., assigns high probability to $t_{-j} < \tau^0$) or thinks that the other player thinks it is likely that the other has an impulse to choose s^1 (i.e., $t_{-j} < \tau^1$), and so on; otherwise, the player chooses s^2. Thus, non-economic factors prevail and some introspective types choose s^1 while others choose s^2 (i.e., $\tau \in (0,1)$). In this case, the value is not equal to one of the Nash payoffs. So, *a key determinant of the value is the relative strength of economic and non-economic factors.*

We can interpret Theorem 1 in terms of the coordination problems that players face. Recall that there is *miscoordination* if players fail to coordinate on one of the pure Nash equilibria, and say that there is *coordination failure* if they coordinate on the Pareto-dominated Nash equilibrium. When non-economic factors dominate ($\rho \in (\underline{\rho}, \overline{\rho})$), behavior in introspective equilibrium is not consistent with Nash equilibrium and there is miscoordination. On the other hand, when economic factors dominate ($\rho < \underline{\rho}$ or $\rho > \overline{\rho}$), introspective equilibrium selects a pure Nash equilibrium and there is no miscoordination. However, there can be coordination failure. This is the case if action s^1 is risky ($\rho > \overline{\rho}$) but players coordinating on s^1 is Pareto optimal ($u_{11} > u_{22}$). As a result, the value can be non-monotonic in the dominance parameter (Figure 3) even though behavior is monotone in payoffs (Proposition 1). When a type space induces non-monotone rank beliefs (i.e., satisfies (NMRB)), then there is miscoordination for an open set of payoff parameters (i.e., $\overline{\rho} > \underline{\rho}$). Given the empirical relevance of miscoordination, we will henceforth assume (NMRB).

Theorem 1 also highlights another key point: *While equilibrium behavior is fully characterized by the "summary statistic" ρ (Proposition 1), the value depends on payoff parameters beyond ρ.* To see this, note that if all parameters are increased by the same amount, then equilibrium behavior stays the same (as ρ does not change); yet the value increases. Doing comparative statics on the value therefore requires a richer and more complex analysis than comparative statics on behavior. For example, to address the key question of whether miscoordination is more costly than coordination failure, a researcher needs to know the value under miscoordination. But, as Eq. (1) shows, the value under miscoordination depends on all four payoff parameters

(i.e., $u_{11}, u_{22}, u_{21}, u_{12}$). So, two games $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$, $\mathcal{G}' = (\boldsymbol{u}', \mathcal{T})$ may have a different value even if they have the same dominance parameter (i.e., $\rho(\boldsymbol{u}) = \rho(\boldsymbol{u}')$). A further complication is that the value under miscoordination (in (1)) depends on the introspective type space, both directly (through the distribution $F(\cdot)$) and indirectly (through the equilibrium threshold τ). Section 3.2 shows that it is nevertheless possible to derive testable comparative statics on the value in important economic environments.

3.2 Applications

3.2.1 Examples G_w and $\widetilde{G}_x$

This section derives novel comparative statics for the game forms G_w and $\widetilde{G}_x$ discussed in the introduction. For some of these results, we make the additional assumption that *no action is strongly salient* in the sense that if the payoff structure provides no guidance (i.e., $\rho = \frac{1}{2}$), then there is miscoordination (i.e., $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$).[18] The following result shows that the model can capture predictions (p1) and (p2) in the introduction:

Proposition 2. **[The Value of G_w]**

(p1) For $w > 1$ sufficiently large, the value of G_w equals w. That is, for every introspective type space $\mathcal{T}$, there is $\underline{w}$ such that $V((w,0,0,1);\mathcal{T}) = w$ whenever $w > \underline{w}$.

(p2) If $w = 1$ and no action is strongly salient, then the value of the game G_w is strictly between $\frac{1}{2}$ and 1. That is, for every introspective type space $\mathcal{T}$ with $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$, $V((1,0,0,1);\mathcal{T}) \in (\frac{1}{2}, 1)$.

Part (p1) yields the intuitive prediction that for w sufficiently large, the value of G_w equals w. Part (p2) shows the novel insight that for $w = 1$, the value lies strictly between that for the mixed Nash equilibrium (viz., $\frac{1}{2}$) and pure Nash equilibrium (viz., 1). This prediction is consistent with experimental evidence that subjects' payoff in the game G_w with $w = 1$ generally exceeds that in the mixed Nash equilibrium yet is less than that in pure Nash equilibrium (Mehta et al., 1994).

The contrast between (p1) and (p2) reflect our earlier finding that a key determinant of the value is the relative strength of economic and economic factors: If w is sufficiently large, economic factors dominate, and the value is given by the Nash payoff w. On the other hand, if $w = 1$, non-economic factors dominate, and the value is not consistent with Nash equilibrium. These findings are consistent with the emerging experimental literature on payoff-irrelevant signals (Duffy and Fisher, 2005). Consistent with our predictions, the evidence suggests that the outcome subjects coordinate on depends on the relative strength of economic and non-economic factors in a way that cannot be explained by Nash equilibrium or its standard refinements. In particular, the

[18] For an interpretation of this condition, see the discussion of the social salience type space below.

effects of non-economic factors are most pronounced when the payoff structure provides little guidance (as in G_w with $w = 1$) (Arifovic and Jiang, 2019), and in such cases, subjects fail to coordinate on one of the Nash equilibria (Fehr et al., 2019). On the other hand, if one of the actions stands out in terms of payoffs (as in G_w when w is sufficiently large), non-economic factors such as nudges have little effect on behavior (Bicchieri and Mercier, 2014; Bicchieri and Dimant, 2019).

We next consider $\widetilde{G}_x$.

Proposition 3. [The Value of $\widetilde{G}_x$] *If no action is strongly salient, then for $w > 1$ sufficiently large, the value of $\widetilde{G}_x$ is lower when x is close to $w - 1 > 0$ than when $x = 0$. That is, for every introspective type space $\mathcal{T}$ with $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$, there is w^* such that for $w > w^*$, $\limsup_{x \uparrow w-1} V((w, -c, 0, 1 + x); \mathcal{T}) < V((w, -c, 0, 1); \mathcal{T})$.*

Proposition 3 implies that players can be worse off if the payoffs to an unplayed equilibrium increase: In $\widetilde{G}_x$, if the payoff $u_{11} = w$ is sufficiently high and x is close to 0, then all players coordinate on the Pareto-dominant Nash equilibrium (s^1, s^1). As the payoff $u_{22} = 1 + x$ in the (initially unplayed) Nash equilibrium (s^2, s^2) approaches w, miscoordination ensues. This reflects a tension between a direct payoff effect and an indirect strategic effect. The direct payoff effect is that for higher values of x, players who coordinate on s^2 receive a higher payoff. This has a positive effect on the value. The indirect strategic effect says that if action s^2 is more attractive in terms of payoffs (i.e., x close to $w - 1 > 0$), there can be more miscoordination. This has a negative effect on the value. Proposition 3 shows that the indirect strategic effect dominates if w is sufficiently large: while all payoff parameters u_{nm} are (weakly) higher when x is close to $w - 1$ than when $x = 0$, the value is strictly lower. Thus, the value can be non-monotonic in payoffs, even though equilibrium behavior changes with payoffs in a monotone way (Proposition 1). While it is well-known that the interaction between the payoff and strategic effect can lead to non-monotonicities in the value (e.g., Morris and Shin, 2003, p. 73), the novel insight is that this can be driven by the cost of miscoordination.

Social salience We discuss which kind of economic situations are adequately modeled by a type space that induces the effects in Propositions 2–3. A natural environment that fits the experiments discussed above is one where players' impulses are sensitive to social cues: Suppose either action may be "socially salient," in the sense that if an action is socially salient, then players are likely to have an impulse to choose that action. That is, suppose action s^1 is socially salient (denoted $\theta = s^1$) with probability $p \in (0, 1)$, while s^2 is socially salient with probability $1-p$ (denoted $\theta = s^2$). Conditional on action s being socially salient, each player j has an impulse to choose s with probability $q_j \geq \frac{1}{2}$. Thus, in some social contexts players are likely to have an

impulse to choose action s^1 (i.e., $\theta = s^1$) while in other contexts, they are likely to have an impulse to choose s^2 (i.e., $\theta = s^2$). The parameter q_j measures how sensitive player j is to social cues. For example, if q_j is close to 1, then player j is almost perfectly attuned to social cues; if q_j is close to $\frac{1}{2}$, he is fairly insensitive to social cues. In experiments, higher-order beliefs consistent with this introspective type space can be generated directly with a correlating device with this structure (Fehr et al., 2019), or more indirectly using salient action labels (Mehta et al., 1994) or random public announcements (Duffy and Fisher, 2005; Arifovic and Jiang, 2019).

If the parameters q_j are drawn from a continuous density $g(\cdot)$ with full support on $(\frac{1}{2}, 1)$ (independently across players) and if we identify the type of a player with impulse I_j and "sensitivity" q_j with its posterior belief $t = t(I_j, q_j)$ that s^1 is socially salient, then we obtain an introspective type space that satisfies our assumptions under mild assumptions; see Appendix B.1 for details.[19] Moreover, more "extreme" types are more sensitive to social cues, in that the sensitivity $q(t)$ of a type t increases in $|\tau^0 - t|$, with $q(\tau^0) = \frac{1}{2}$ and $q(1) = q(0) = 1$. Figure 2 presents the rank belief function for this type space (for specific choices of p and $g(\cdot)$). Because the rank belief function $F(t \mid t)$ in Figure 2 is decreasing near $t = \tau^0$, this type space satisfies (NMRB). To see the intuition why players' rank belief $F(t \mid t)$ decreases for t close to τ^0, note that an increase in t has two effects. First, higher types think it is more likely that s^1 is socially salient. Because it is less likely that the other player has a low type conditional on s^1 being socially salient, this has a negative impact on $F(t \mid t)$. Second, as a player's type t increases, there are simply more types $t' \leq t$. This has a positive effect on $F(t \mid t)$. "Extreme" types (i.e., t close to 0 or 1) are quite confident which action is socially salient (as $q(t)$ is close to 1). This means that the first effect is small, so $F(t \mid t)$ is increasing in t. On the other hand, "intermediate" types (i.e., t close to τ^0) become considerably more confident that s^1 is socially salient when t increases (as $q(t)$ is close to $\frac{1}{2}$). Hence, the first effect dominates and $F(t \mid t)$ is decreasing in t (under the mild assumption that $g(\frac{1}{2})$ is sufficiently small; see Appendix B.1). This introspective type space also has the natural interpretation that no action is strongly salient precisely when no action is highly likely to be socially salient (i.e., p sufficiently close to $\frac{1}{2}$).

This introspective type space is well-suited for modeling environments where the salience of an action depends on the social context (i.e., $p \in (0, 1)$) and on idiosyncratic factors (i.e., $q_j < 1$ with positive probability). This fits well with many experimental designs (e.g., Mehta et al., 1994; Duffy and Fisher, 2005; Fehr et al., 2019; Arifovic and Jiang, 2019) as well as situations where a shared culture leads players to have consistent expectations.[20] For these settings, our theory predicts that players can coordinate on an action profile even if the constituent actions

[19] The limiting case where $g(\cdot)$ converges to a point mass at some common q corresponds to the simple parametric model in Kets and Sandroni (2019, 2021).

[20] This introspective type space is is less well-suited for modeling settings where there are systematic differences in how people respond to social cues (as q_j is assumed to be at least $\frac{1}{2}$).

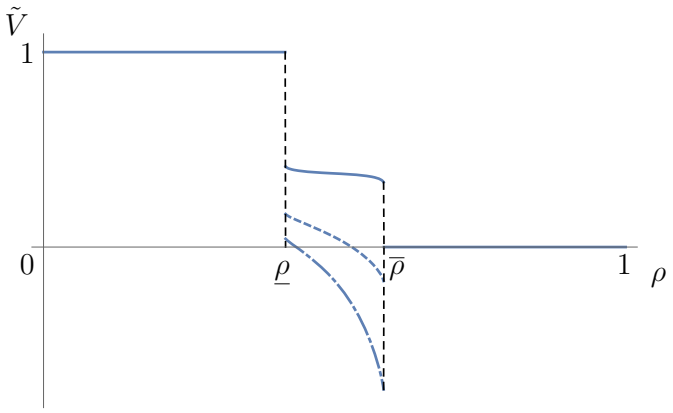

Figure 3: The normalized value $\tilde{V} := \frac{V-u_{22}}{u_{11}-u_{22}}$ as the investment subsidy s is varied, plotted as a function of $\rho = \rho(s)$. The solid, dashed, and dash-dotted line correspond to the game forms $\boldsymbol{u}^0 = (1, -\frac{3}{5}, \frac{3}{5}, 0)$, $\boldsymbol{u}^0 = (1, -3, -1, 0)$, and $\boldsymbol{u}^0 = (1, -6, -3, 0)$ at $s = 0$, respectively (so $\rho(0) = \frac{3}{5}$ in each case).

are not salient to them: A player may choose an action that is not salient to him if he thinks it likely that it is salient to the other player or that the other player thinks it is salient to the other player, and so on. However, because people are not always perfectly attuned to social cues (i.e., $q_j < 1$ with positive probability), there can be miscoordination if the payoff structure provides little guidance. Thus, increasing the payoffs to one of the actions can make players worse off if it creates miscoordination (Proposition 3) even if miscoordination is less costly than presumed under mixed Nash equilibrium (Proposition 2). As various experimental designs are well suited for implementing the social salience type space, these predictions can readily be tested.

3.2.2 Stimulating investment

This section considers a setting where a policy-maker can subsidize investment. We assume that investment is Pareto optimal, i.e., we identify action s^1 with investing and action s^2 with not investing. We study the effects of investment subsidies that are large enough to generate some investment but may not be sufficient to generate full investment (cf. Morris and Yildiz, 2019). In the language of our model, the investment subsidy eliminates coordination failure but may create miscoordination. Players who invest receive a subsidy s regardless of whether the other player invests. Formally, both u_{11} and u_{12} are increased by the same amount $s \geq 0$. The following result shows that while introducing an investment subsidy (weakly) increases the payoffs associated with each action profile, it may lead to a *reduction* in the value if miscoordination is more costly than coordination failure:

Theorem 2. [The Value of Investment Subsidies] *Fix an introspective type space with a positive probability of investment at $\overline{\rho}$ (i.e., $\overline{\tau} < 1$). Then, the value of a game with coordination failure in introspective equilibrium ($\rho > \overline{\rho}$) strictly decreases with the investment subsidy s as it induces miscoordination (i.e., the dominance parameter falls to $\overline{\rho}$) if and only if the off-diagonal*

payoffs are sufficiently small and ρ is not too high: There is $\rho^* \in (\overline{\rho}, 1)$ such that for all u_{11} and u_{22} with $u_{11} \geq u_{22}$ and for all $\rho > \overline{\rho}$, there exist u_{12}^* and u_{21}^* such that the following holds: for any game form $\boldsymbol{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ with dominance parameter ρ, as the subsidy s increases, the value falls below u_{22} at $\overline{\rho}$ if and only if $\rho < \rho^*$, $u_{12} < u_{12}^*$, and $u_{21} < u_{21}^*$.

Proof. Fix an introspective type space $\mathcal{T}$ with $\overline{\tau} < 1$ and let $(\overline{p_{11}}, \overline{p_{12}}, \overline{p_{21}}, \overline{p_{22}})$ be the probability distribution over action profiles in introspective equilibrium for $\mathcal{T}$ when the dominance parameter is $\overline{\rho}$. It will be convenient to define

$$\rho^* = \overline{\rho} + \frac{\overline{p_{21}}}{\overline{p_{11}} + \overline{p_{21}}}.$$

Note that ρ^* depends only on $\mathcal{T}$. Clearly, by (REG), $\rho^* > \overline{\rho}$. We also have that $\rho^* < 1$. To see this, note that $\rho^* < 1$ if and only if $\overline{p_{21}} < (1 - \overline{\rho})(\overline{p_{11}} + \overline{p_{21}})$. But this follows from the proof of Theorem 1 (using that $\overline{p_{21}} = \overline{p_{12}}$).

For $s \geq 0$, define $\boldsymbol{u}^s := (u_{11} + s, u_{12} + s, u_{21}, u_{22})$ and suppose that there is coordination failure in introspective equilibrium when $s = 0$, i.e., $\rho := \rho(\boldsymbol{u}^0) > \overline{\rho}$. As s increases, the dominance parameter decreases. Let $\overline{s}$ be the investment subsidy for which the dominance parameter attains the value $\overline{\rho}$. Then,

$$\rho = \frac{u_{22} - u_{12}}{u_{11} - u_{21} + u_{22} - u_{12}} \qquad \text{and} \qquad \overline{\rho} = \frac{u_{22} - u_{12} - \overline{s}}{u_{11} - u_{21} + u_{22} - u_{12}}. \tag{2}$$

The difference in value between the games with investment subsidy $\overline{s}$ (with miscoordination) and without an investment subsidy (i.e., $s = 0$) (with coordination failure) is

$$\Delta = \overline{p_{22}}\, u_{22} + \overline{p_{12}}\,(u_{12} + \overline{s}) + \overline{p_{21}}\, u_{21} + \overline{p_{11}}\,(u_{11} + \overline{s}) - u_{22}.$$

Using (2) and that $\overline{p_{21}} = \overline{p_{12}}$, we can rewrite this as follows:

$$\begin{aligned}\Delta &= (u_{11} - u_{22})(\overline{p_{11}} + \overline{p_{21}}) - \overline{p_{21}}\,(u_{11} - u_{21} + u_{22} - u_{12}) + (\overline{p_{11}} + \overline{p_{21}}), \overline{s} \\ &= (u_{11} - u_{22})(\overline{p_{11}} + \overline{p_{21}}) - \frac{u_{22} - u_{12}}{\rho}\left(p_{21} + (\overline{\rho} - \rho)(\overline{p_{11}} + \overline{p_{21}})\right).\end{aligned}$$

Using the definition of ρ^* and that $p_{12} + p_{11} > 0$, we find that $\Delta < 0$ if and only if

$$u_{11} - u_{22} < \frac{u_{22} - u_{12}}{\rho}(\rho^* - \rho).$$

As the left-hand side is non-negative, $\Delta < 0$ only if $\rho < \rho^*$. In that case, $\Delta < 0$ is equivalent to

$$\frac{u_{22} - u_{12}}{\rho} = \frac{u_{11} - u_{21}}{1 - \rho} > \frac{1}{\rho^* - \rho}(u_{11} - u_{22}),$$

where the equality follows from (2). □

Theorem 2 shows that introducing an investment subsidies may reduce welfare, even if there are no costs to the subsidy, the subsidy directly increases all players' payoffs when they invest and the subsidy does not decrease the payoffs of any player in any play of the game. Intuitively, while an investment subsidy has only positive direct effects, it can have a negative strategic effect in that it may create miscoordination even as it eliminates coordination failure. The net strategic effect is negative if miscoordination is more costly than coordination failure. Moreover, the negative strategic effect can dominate the positive direct effect. This is the case whenever two conditions are met: (1) the off-diagonal payoffs are low (i.e., $u_{21} < u_{21}^*$ and $u_{12} < u_{12}^*$); and (2) the risk of investing (in the absence of investment subsidies) is not too high (i.e., $\rho < \rho^*$). The intuition behind condition (1) is that miscoordination is particularly costly when the payoffs u_{12}, u_{21} players receive under miscoordination are low. This is illustrated in Figure 3: the value under miscoordination ($\rho \in (\underline{\rho}, \overline{\rho})$) lies below the value under coordination failure ($\rho > \overline{\rho}$) for low values of off-diagonal payoffs but not otherwise. Condition (2) says that introducing an investment subsidy can be particularly detrimental to welfare if investing is relatively attractive in terms of payoffs (though not sufficiently attractive to induce positive investment). In particular, if ρ is close to $\overline{\rho}$, then investment subsidies can reduce welfare. Intuitively, when investing is relatively attractive in terms of payoffs, a small investment subsidy suffices to induce (partial) investment. But since the subsidy is only small, it cannot offset the cost of miscoordination even though the subsidy offers some insurance to a player against the risk of miscoordination (since he receives the subsidy regardless of whether the other player invests). So, introducing an investment subsidy reduces the value even as it stimulates investment if miscoordination is more costly than coordination failure (i.e., $u_{21} < u_{21}^*$, $u_{12} < u_{12}^*$, and $\rho < \rho^*$). Thus, *for a policy to improve welfare, it is not sufficient that it changes behavior in the desired direction, it must also ensure that the costs of miscoordination are not too high.* These insights are robust: Appendix C shows that the same result obtains for an alternative policy to promote investment.

Animal spirits We discuss the kind of situations that can be modeled by a type space that induces the effects in Theorem 2. A natural environment that fits the present context is one where impulses may be driven by "animal spirits," that is, large shocks to public sentiment that affect individual impulses. To model this, we follow Morris and Yildiz (2019) and assume that each player's type is the sum of a "common shock" η that affects both players, and an idiosyncratic term ε_j that varies across players, i.e., $\tilde{t}_j = \eta + \varepsilon_j$. While in Morris and Yildiz's work, types reflect payoff-relevant information, types can alternatively encode players' impulses

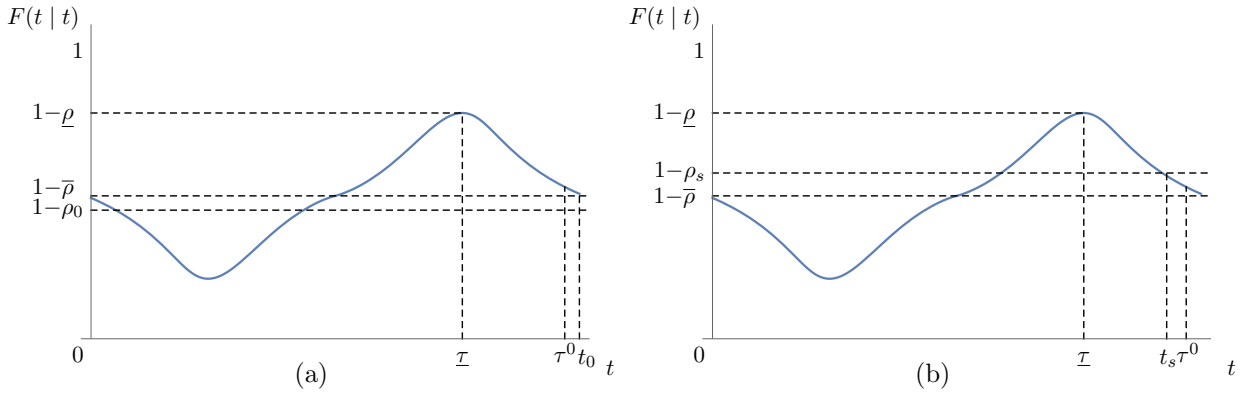

Figure 4: Introspective equilibria for the "animal spirits" type space: (a) without an investment subsidy; (b) with an investment subsidy.

and beliefs about others' impulses.[21] Under this interpretation, the common shock η reflects a shift in public sentiment while ε_j captures individual heterogeneity (with some types being more bullish or bearish than others). We are interested in the case where animal spirits are potentially important, i.e., where the shift η in public sentiment can be large. Following Morris and Yildiz, we therefore assume that the common shock η is drawn from a fat-tailed distribution (e.g., t-distribution) while the distribution of the idiosyncratic terms ε_j has thin tails (e.g., normal distribution). As detailed in Appendix B.2, our framework can accommodate this if we map each type $\tilde{t}_j \in \mathbb{R}$ into a type $t_j \in [0,1] = T$ in a way that preserves beliefs, and then associate each $t_j \in T$ with an impulse $\mathcal{I}_j(t_j)$ (cf. (MON-I)).

Figure 4 illustrates the rank belief function for the resulting introspective type space. Unlike in the social salience type space (Figure 2), the rank belief function is now decreasing for extreme types (i.e., t close to 0 or 1). Following Morris and Yildiz (2019, pp. 2831–2832), we can interpret this as follows. Suppose that there is limited individual heterogeneity (i.e., ε_j is drawn from the standard normal distribution) and that players are uncertain whether animal spirits are important, i.e., η is normally distributed but with unknown variance. Taking the reciprocal of the variance to be χ^2-distributed, this gives a t-distribution for the distribution of η. Under these assumptions, an increase in a player's type has two effects. First, when a player's type increases, he attributes some of the increase to idiosyncratic factors. This gives a *reversion to the mean effect*: The player's expectation of the common shock will be further from his own type as his type increases. This has a positive effect on the rank belief. Second, there is a *learning effect*: Extreme types think it is likely that the variance in public sentiment is higher. This has a negative effect on the rank belief. For extreme types, the learning effect dominates the reversion to the mean effect, while for intermediate types, the reverse is true. This gives the shape in Figure 4.

Our model predicts that investment may fail to take off when players have an impulse to

[21] In this sense, our model is related to the literature on sunspots (e.g., Cass and Shell, 1983). However, because sunspot models typically feature multiple equilibria, they do not allow for testable comparative statics on welfare.

invest only if they expect public sentiment to be strongly in favor of investing (i.e., τ^0 close to 1). This is the case illustrated in Figure 4. Panel (a) shows the case without investment subsidies (with equilibrium threshold t_0 and dominance parameter ρ_0), and panel (b) shows the case with an investment subsidy in place (with equilibrium threshold t_s and dominance parameter ρ_s). Without an investment subsidy (panel (a)), not investing is the unique best response for τ^0 (i.e., $F(\tau^0 \mid \tau^0) > 1 - \rho_0$); so, by our earlier argument, the probability that players invest in introspective equilibrium is lower than if they had followed their impulse (i.e., $t_s > \tau^0$). In this case, investing is so unattractive in terms of payoffs (i.e., $\rho_0 > \overline{\rho}$) that there is no investment in introspective equilibrium (i.e., $\tau = 1$). So, without any subsidies, there is coordination failure.

With an investment subsidy (panel (b)), investing is the unique best response for τ^0 (i.e., $F(\tau^0 \mid \tau^0) < 1 - \rho_s$). Moreover, as the introspective process continues, more types decide to invest (i.e., $\tau^k < \tau^{k-1}$): Types that are not too optimistic about public sentiment invest if they think it likely that other types are optimistic, or that other types think that other types are optimistic, and so on. However, while types close to τ^0 are optimistic that there has been a large shift in public sentiment (i.e., $F(t \mid t)$ is decreasing for t near τ^0), types become less and less confident of this as t decreases. Because investment is not too attractive in terms of payoffs in this example (i.e., $\rho_s < \underline{\rho}$), investment fails to take off fully (i.e., $t_s > 0$) even as some types invest ($t_s < 1$). So, there is miscoordination. Theorem 2 predicts introducing an investment subsidy makes players worse off if this induces miscoordination (as is the case here) and the investment subsidy is too small to compensate for the low payoffs players receive when they miscoordinate (i.e., $\rho < \rho^*$, $u_{21} < u_{21}^*$, $u_{12} < u_{12}^*$).[22]

3.2.3 Collusion

We next apply the model to study tacit collusion. We consider a simple model of price competition with product differentiation (Ross, 1992). There are two firms that produce a good. In each period $\tilde{t} = 0, 1, \ldots$, firms set their prices simultaneously. The goods are imperfect substitutes: The inverse demand function for firm $i \in \{1, 2\}$ in any given period $\tilde{t}$ is

$$p_i = a - bq_i - cq_{-i},$$

where $a > 0$ and $b > c > 0$ are constants, $p_i \geq 0$ is firm i's price, and $q_i, q_{-i} \geq 0$ are the demand for firm i's and the other firm's products, respectively. Hence, $r := \frac{c}{b} \in (0, 1)$ measures the degree of substitutability: In the limit $r \to 1$, the products are perfect substitutes; and in the limit $r \to 0$, the demand for each product is independent of the other's. For simplicity, the marginal cost

[22] A key difference between our belief-based approach with the payoff-based approach in Morris and Yildiz (2019) is that their approach does not select a unique (Bayesian) equilibrium. Their work therefore does not deliver testable comparative statics.

of each firm is taken to be 0. The collusive price p^* is the price that, when charged by both firms, maximizes joint profits; and the cheating price p^c is the price that maximizes a firm's profits if the other firm charges the collusive price. Let π^* be the "collusive" profit, i.e., a firm's (one-period) profit when both firms charge p^*; let π^N be the Bertrand-Nash profit, i.e., a firm's profit when both charge the Bertrand-Nash price p^N; let π^c be the "cheating" profit, i.e., a firm's profit when it charges p^c while the other firm charges p^*; and let π^v be the "victim" profit, i.e., a firm's profit when it charges p^* while the other firm charges p^c. In the appendix, we show that $\pi^c > \pi^* > \pi^N > \pi^v$, i.e., the (one-shot) price competition game has the structure of a prisoner's dilemma. We also calculate the "mutual cheating" profit π^m, i.e., the profit a firm earns when both charge p^c, and show that $\pi^* > \pi^m > \pi^N$.

Firms have a common discount factor $\delta \in (0,1)$ and their payoff is their expected discounted sum of profits. Following Spagnolo (2003), we assume that each firm chooses between a collusive (grim-trigger) strategy and a cheating strategy.[23] Under the *collusive strategy* (denoted by σ^*), in each period $\tilde{t} \geq 0$, a firm chooses the collusive price p^* provided that both firms have chosen the collusive price in all past periods $\tilde{t}' < \tilde{t}$; otherwise, it charges the Bertrand-Nash price p^N of the one-shot game. Under the *cheating strategy* (denoted by σ^c), a firm chooses the cheating price p^c in every period $\tilde{t} \geq 0$ as long as both firms have chosen the collusive price in all past periods; otherwise, it charges the Bertrand-Nash price p^N. Then, the firm's payoffs (expected discounted sum of profits) under various strategy combinations are given by

	σ^*	σ^c
σ^*	π^*	$(1-\delta)\,\pi^v + \delta\pi^N$
σ^c	$(1-\delta)\,\pi^c + \delta\pi^N$	$(1-\delta)\,\pi^m + \delta\pi^N$

where we have listed only the row player's payoff for expositional simplicity. To rule out trivialities, we assume that the discount factor δ is sufficiently high for both players choosing σ^* to be a (strict) subgame perfect equilibrium, i.e., we require that $\delta > (\pi^c - \pi^*)/(\pi^c - \pi^N)$.

If we identify s^1 with σ^* and s^2 with σ^c, we can view this as a coordination game. Coordination failure then means that no firm tries to collude; and miscoordination means that one firm tries to establish cooperation (i.e., chooses σ^*) but collusion breaks down because the other firm cheats (i.e., chooses σ^c). As firms become more patient (i.e., δ increases), collusion becomes more attractive in terms of payoffs (i.e., $\rho = \rho(\delta)$ decreases). The following result shows that firms tend to be better off if an increase in the discount factor allows them to avoid coordination failure even if that comes at the expense of miscoordination:

[23] Also see Blonski et al. (2011) and Dal Bó and Fréchette (2011). See Kets and Sandroni (2021) for microfoundations for this approach in the context of introspective equilibrium.

Theorem 3. [**The Value of Collusion**] *Fix an introspective type space and parameters a, b, c. Let $\delta, \delta' \in \left((\pi^c - \pi^*)/(\pi^c - \pi^N), 1 \right)$ be such that there is coordination failure in introspective equilibrium when the discount factor is δ (i.e, $\rho(\delta) > \overline{\rho}$) but not when the discount factor is δ' (i.e., $\rho(\delta') < \overline{\rho}$). Then, if $\delta' - \delta$ is either sufficiently small or sufficiently large, the value of the game with discount factor δ' is strictly larger than the value of the game with discount factor δ (with coordination failure).*

Proof. We start by deriving general conditions under which the cost of miscoordination exceeds the cost of coordination failure (for any game form $\boldsymbol{u}$).

Lemma 1. *Fix an introspective type space and let ρ, ρ' be such that $\underline{\rho} < \rho' < \overline{\rho} < \rho$, i.e., there is coordination failure in introspective equilibrium for the game with dominance parameter ρ and there is miscoordination for the game with dominance parameter ρ'. Let $(u_{11}, u_{12}, u_{21}, u_{22})$ be the payoffs such that the dominance parameter is $\rho > \overline{\rho}$, and $(u'_{11}, u'_{12}, u'_{21}, u'_{22})$ the payoffs such that the dominance parameter is ρ'. Also, let $p'_{11}, p'_{12}, p'_{21}, p'_{22}$ be the probabilities with which each action profile is played in introspective equilibrium when the dominance parameter is ρ'. Then, the difference Δ in value between the games with dominance parameters ρ' (with miscoordination) and ρ (with coordination failure) satisfies*

$$\Delta > (p'_{11} + p'_{12})(u'_{21} - u'_{22}) - (u_{22} - u'_{22}). \tag{3}$$

Proof. From

$$\frac{\rho'}{1 - \rho'} = \frac{u'_{22} - u'_{12}}{u'_{11} - u'_{21}}$$

it follows that

$$\rho'(u'_{11} - u'_{22}) = u'_{22} - u'_{12} + \rho'(u'_{12} + u'_{21} - 2u'_{22}). \tag{4}$$

Using the fact that $p'_{12} = p'_{21}$ and $p'_{11} + p'_{12} + p'_{21} + p'_{22} = 1$, the difference in value between the games with dominance parameters ρ' (with miscoordination) and ρ (with coordination failure) is

$$\begin{aligned} \Delta &= p'_{11}\, u'_{11} + p_{12}\, u'_{12} + p'_{21}\, u'_{21} + p'_{22}\, u'_{22} - u_{22} \\ &= p'_{11}\,(u'_{11} - u'_{22}) + p'_{12}\,(u'_{12} + u'_{21} - 2\,u'_{22}) - (u_{22} - u'_{22}). \end{aligned}$$

Using (4) we obtain

$$\Delta = \frac{p'_{11}}{\rho'}\,(u'_{22} - u'_{12}) + (p'_{11} + p'_{12})(u'_{12} + u'_{21} - 2\,u'_{22}) - (u_{22} - u'_{22}), \tag{5}$$

This, together with the fact that $p'_{11} > \rho'\,(p'_{11} + p'_{12})$ (proof of Theorem 1) then gives (3). □

We can now prove Theorem 3. The appendix shows that $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$. Since $\delta, \delta' > (\pi^c - \pi^*)/(\pi^c - \pi^N)$, we have $u_{11} - u_{21} > 0$; and since $\pi^m > \pi^v$, we have $u_{22} - u_{12} > 0$. Hence,

if we write ρ and ρ' for the dominance parameters for the games with discount factors δ and δ', respectively, then $\rho, \rho' \in (0,1)$. If $\delta' - \delta$ is sufficiently large, then $\rho' < \underline{\rho}$ (i.e., all players play the collusive strategy σ^* in introspective equilibrium). The result then follows immediately from the fact that $\pi^* > \pi^m, \pi^N$. So suppose $\rho' \in [\underline{\rho}, \overline{\rho})$ (i.e., $\delta' - \delta$ is not large). Then, (3) gives

$$\Delta > (p'_{11} + p'_{12})(1-\delta')(\pi^c - \pi^m) - (\delta' - \delta)(\pi^m - \pi^N).$$

Since the first term on the right is positive, it follows that $\Delta > 0$ for all $\delta < \delta'$ sufficiently close to δ'. □

Theorem 3 shows that firms have a strong incentive to avoid coordination failure. Clearly, firms are better off if they both collude than if neither colludes (i.e., $u_{11} > u_{22}$). More surprisingly, perhaps, is that firms may be better off even if there is miscoordination provided that the change in discount factor is not too large. That is, even conditions are such that collusion is not guaranteed (i.e., $\rho > \underline{\rho}$), firms are better off if there is some collusion (i.e., $\rho \in (\underline{\rho}, \overline{\rho})$) than if there is no collusion at all (i.e., $\rho > \overline{\rho}$). The intuition is that if the change in discount factor is not too large, then the positive indirect effect (a positive probability of collusion) outweighs the negative direct effect (a reduction in payoff u_{22} when no player colludes).

One important implication of Theorem 3 is that industry bodies have an incentive to lobby for changes that effectively increase the discount factor. Examples include increasing the frequency of interaction to improve the ease of detection (Stigler, 1964).[24] These predictions are in line with empirical evidence. For example, the US government's practice to buy vaccines in bulk helps to undo collusion by reducing the frequency of interaction (Scherer, 1980). As another example, an increase in price transparency in the Danish concrete industry made it easier for firms to detect defections and led to more collusion (Albæk et al., 1997). Relatedly, some trade associations frequently publish information on past prices, which can facilitate collusion (Kühn, 2001). Another implication of Theorem 3 is that focusing regulators' resources exclusively on detecting explicit collusion (as advocated by, e.g., Motta, 2004, p. 190) may allow many cases of collusion to go undetected. That is, even if the conditions for collusion are less than ideal and

[24] We follow Ivaldi et al. (2003). To see how changing the frequency of interaction affects the (effective) discount factor, suppose that goods are sold every T periods . Then, δ should be replaced by δ^T throughout in the calculations (e.g., $u_{22} = (1-\delta^T)\pi^m + \delta^T\pi^N$. Since $\delta^T < \delta$, reducing the frequency of interactions effectively lowers the discount factor. To see the effects of the ease of detection, suppose a firm cheating on a tacit agreement can earn profits for two periods before being detected and punished. Then, if both firms cheat, both receive

$$u_{22} = (1-\delta)\left(\pi^m(1+\delta) + \pi^N(\delta^2 + \delta^3 + \ldots\right) = (1-\delta^2)\pi^m + \delta^2\pi^N.$$

Similarly, $u_{12} = (1-\delta^2)\pi^v + \delta^2\pi^N$ and $u_{21} = (1-\delta^2)\pi^c + \delta^2\pi^N$. Thus, making it harder to detect collusion reduces the effective discount factor.

there is a positive probability that firms may fail to collude (i.e., $\rho > \underline{\rho}$ and thus $p_{11} < 1$), they may just be successful at initiating collusion (i.e., $p_{11} > 0$).[25] In Appendix D, we show that these conclusions do not hinge on the particulars of our model of collusion: similar results obtain for other commonly-used models.

Comparative statics on the value can thus give insight into the question of whether parties have an incentive to influence their environment. These results cannot be obtained with comparative statics on equilibrium behavior (which merely shows whether collusion can be sustained in equilibrium, not whether parties would want to). This suggests that regulators need to pay careful attention not just to whether policy changes make collusion viable in equilibrium but also to how these changes affect the value.

4 Dynamic application

This section studies whether the welfare effects we have identified thus far have long-term consequences, for example whether miscoordination can persist in the long run or whether policies that reduce the value have long-term detrimental welfare implications.

We have in mind situations where the social context may be influenced by past behavior (perhaps with noise). For example, an action can be salient due to historical precedent. To model this, we assume that there is a continuum of players and in each period $\tilde{t} = 0, 1, 2, \ldots$, all players are matched in pairs (at random) to play a game $\mathcal{G}_{\tilde{t}} = (\boldsymbol{u}, \mathcal{T}_{\tilde{t}})$. Thus, the game form $\boldsymbol{u}$ is fixed across periods but the social context (i.e., $\mathcal{T}_{\tilde{t}}$) may evolve over time. At each time $\tilde{t}$, each pair of players plays the introspective equilibrium of $\mathcal{G}_{\tilde{t}}$ (after their types have been realized).[26] At any time $\tilde{t}$, the social context is modeled by the introspective type space $\mathcal{T}_{\tilde{t}} = (F, \tau^0_{\tilde{t}})$, so the share of matches involving types $t_1, t_2 \in T$ is $f(t_1, t_2)$.[27] The key assumption is that players'

[25] Arguably, by restricting to two firms and to two strategies per firm, our model may understate the difficulties of collusion. Our model is thus mostly applicable to relatively simple situations where there is a clear focal collusive price, such as those studied by Carlton et al. (1997) and Knittel and Stango (2004). In other cases, identifying the appropriate collusive strategies may take time (e.g., Byrne and De Roos, 2019). In such cases, Theorem 3 suggests the striking conjecture that experimenting with strategies to identify appropriate collusive actions (which may create miscoordination) may carry little, if any, cost, relative to the baseline without collusion (coordination failure), implying that tacit collusion can be an important threat even in these more complex settings.

[26] Thus, players do not take into account that their actions today may influence the social context tomorrow. This seems reasonable for the current (large population) setting. We could alternatively assume that each period represents a generation, with limited externalities across generations.

[27] The assumption that the distribution function $F(\cdot)$ remains fixed is obviously restrictive. However, how the distribution function evolves over time likely depend on the specifics of the type space in a way that is difficult to generalize.

impulses at time $\tilde{t}$ are shaped by some combination of their original impulses and the most recent population play. That is, the level-0 threshold $\tau^0_{\tilde{t}}$ at time $\tilde{t}$ lies between the original level-0 threshold τ^0_0 and the equilibrium threshold $\tau_{\tilde{t}-1}$ at time $\tilde{t}-1$, with some noise $\varepsilon > 0$:

$$\begin{aligned} \tau^0_{\tilde{t}} &> \min\{\tau^0_0, \tau_{\tilde{t}-1}\} - \varepsilon; \\ \tau^0_{\tilde{t}} &< \max\{\tau^0_0, \tau_{\tilde{t}-1}\} + \varepsilon. \end{aligned} \tag{6}$$

The assumption that the original impulses (i.e., τ^0_0) may influence current impulses (i.e, $\tau^0_{\tilde{t}}$) captures the idea that there are factors extraneous to the game that have persistent effects on the social context.[28] The noise ε captures the idea that there can be slight shocks to the social context.[29] For example, past experiences are not always perfectly transmitted over time. We are agnostic about the extent to which players' current impulses are driven by their original impulses or past population play; we only require that they depend on some combination of these. That is, we allow for any dynamic $\{\tau^0_{\tilde{t}}\}_{\tilde{t}}$ that satisfies (6) (for given ε).[30]

The following result shows that introspective equilibrium can be viewed as the steady state of any such dynamic process:

Proposition 4. [Introspective Equilibrium as a Steady State] *If the noise is sufficiently small, then the introspective equilibrium remains largely unchanged over time: For every $\chi > 0$, there is $\bar{\varepsilon} > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon})$, for every pair of periods $\tilde{t}, \tilde{t}'$, the equilibrium thresholds $\tau_{\tilde{t}}, \tau_{\tilde{t}'}$ are within χ of each other, i.e., $|\tau_{\tilde{t}} - \tau_{\tilde{t}'}| < \chi$.*

Proposition 4 shows that introspective equilibrium can be viewed as the steady state of a process where the introspective process is not only shaped by but also shapes the social context. Proposition 4 implies that when players' impulses may be driven by past population play, then they will continue to behave similarly. Thus, in this sense, introspective equilibrium is rich enough to accommodate environments where the social context is partly shaped by people's behavior. Moreover, in such cases, introspective equilibrium reflects stable patterns of behavior. Intuitively, each introspective equilibrium has a "basin of attraction," such that as long as the level-0 threshold falls into that basin, the introspective equilibrium and the value remain unchanged. For example, in Figure 1(b), the basin of attraction for the introspective equilibrium $\tau \in (0,1)$ with miscoordination is $[\underline{\tau}, \bar{t})$: for any $\tau^0 \in [\underline{\tau}, \bar{t})$, the introspective process converges to the introspective equilibrium τ. Because the basin of attraction contains both the level-0 threshold and the equilibrium threshold, any introspective process that begins between these

[28] We could alternatively assume that the level-0 threshold $\tau^0_{\tilde{t}}$ at time $\tilde{t}$ lies between the thresholds $\tau^0_{\tilde{t}-1}$ and $\tau_{\tilde{t}-1}$ at time $\tilde{t}-1$ (with some noise $\varepsilon > 0$). Proposition 4 extends to that case, with slight modifications.

[29] This is not critical. Proposition 4 below also holds if there is no noise.

[30] For example, we take $\{\tau^0_{\tilde{t}}\}_{\tilde{t}}$ to be a deterministic process for expositional simplicity, but since our results hold for any such process, they extend to the case where $\{\tau^0_{\tilde{t}}\}_{\tilde{t}}$ is random.

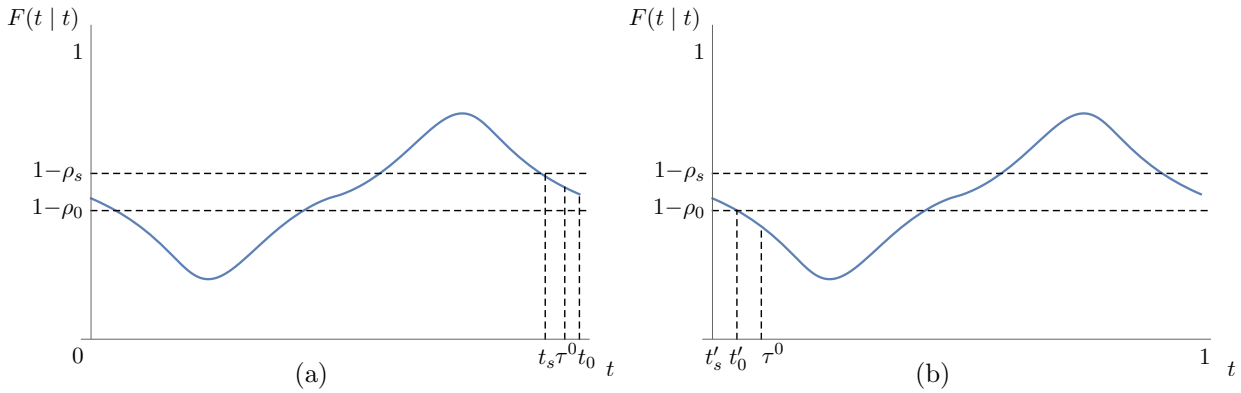

Figure 5: Introspective equilibria for the "animal spirits" type space with and without investment (with respective dominance parameters ρ_s and ρ_0): (a) when players have an impulse to invest only if they expect public sentiment to favor investment; (b) when players have an impulse to invest unless they expect public sentiment to be strongly against investing.

two thresholds converges to the same equilibrium threshold τ. Proposition 4 shows that this insight extends to the case where there can be shocks to the social context.

We apply Proposition 4 to our investment example in Section 3.2.2. Proposition 4 suggests that there can be *path dependence* in this environment. That is, depending on the initial social context (i.e., $\mathcal{T}_0$), a society may settle into different stable patterns of behavior. This is illustrated by Figure 5. Figure 5(a) shows the introspective equilibria with and without an investment subsidy (with respective thresholds t_s, t_0) under the assumption in Section 3.2.2 that players have an impulse to invest only if they expect public sentiment to strongly favor investment (i.e., τ^0 close to 1). In this case, there is no or little investment in equilibrium even with subsidies in place (i.e., $t_0 = 1$, t_s close to 1). Figure 5(b) shows the introspective equilibria t'_s, t'_0 for the same payoff parameters under the alternative assumption that players have an impulse to invest unless they expect public sentiment to be strongly against investing (i.e., τ^0 close to 0). In this case, there are high levels of investment in equilibrium even in the absence of subsidies (i.e., $t'_s = 1$, t'_0 close to 1). Thus, the initial social context can have persistent effects: Societies where players initially have no inclination to invest (Figure 5(a)) can get locked into a low-investment state for a very long time (i.e., $\tau_{\tilde{t}}$ close to 1 for $\tilde{t} \geq 0$) (Proposition 4). In particular, investment subsidies need not lead to a virtuous cycle even though investment decisions are strategic complements (i.e., $\tau_{\tilde{t}}$ close to 1 for all $\tilde{t}$). In fact, if investment is low, this is a sign that the social context does not favor investment (i.e., $\tau^0_{\tilde{t}}$ close to 1); and if that is the case, a subsidy is likely to be ineffective at promoting full investment (i.e., $\tau_{\tilde{t}}$ close to 1), unless the subsidy is very large. Moreover, low but nonzero levels of investment can be costlier than no investment at all (Theorem 2). Thus, *policies that do not account for the effects of the social context or that ignore the possibility of miscoordination can have long-term detrimental effects on welfare.*

While intuitive, these results are difficult to obtain with other approaches. Other approaches that deliver persistence or study shocks to public sentiment – whether belief-based (e.g., Cass and Shell, 1983; Diamond, 1982; Cooper and John, 1988) or payoff-based (Morris and Yildiz, 2019,

Sec. IV) – typically feature multiple equilibria. This makes it difficult to derive welfare implications or to make testable predictions. For example, some equilibria need not feature persistence or may not respond to shocks in public sentiment. And while the belief-based approach of Angeletos and La'O (2013) can generate boom-bust cycles, it delivers unique predictions only when there are significant information frictions. While information frictions can be important when there are frequent shocks, friction-based models seem less suitable for explaining the long-term persistence that we focus on here.

5 Discussion and related literature

5.1 Non-monotone rank beliefs

Our analysis reveals that miscoordination can have important welfare implications. For example, introducing subsidies that benefit everyone in the absence of strategic effects may reduce welfare if miscoordination is more costly than coordination failure. As we noted, miscoordination can arise in environments that induce non-monotone rank beliefs (i.e., satisfy (NMRB)) but not in other settings (generically). Thus, restricting attention to type spaces that fail (NMRB), as much of the literature has done so far, risks overlooking important welfare effects by allowing for coordination failure but ignoring miscoordination.

We have given two examples (Sections 3.2.1 and 3.2.2) of introspective type spaces that satisfy (NMRB). As these examples demonstrate, assumption (NMRB) covers a wide variety of settings. The examples differ not only in their assumptions on the underlying social context (social cues, animal spirits), but also in the rank beliefs they generate (Figure 2 vs. Figure 4). It is therefore difficult to draw any firm conclusions on what type of environments induce non-monotone rank beliefs. For example, while both examples involve some form of aggregate uncertainty (social salience in Section 3.2.1, animal spirits in Section 3.2.2), aggregate uncertainty is not sufficient to induce non-monotone rank beliefs. For example, the type spaces considered in the global games literature also feature aggregate uncertainty but do not induce non-monotone rank beliefs.[31] The question of which type of economic environments naturally induce non-monotone rank beliefs is a fascinating one that we leave for future research.

5.2 Point predictions

Throughout much of the paper, we have focused on providing testable *comparative statics*. However, it is also worth asking whether it is possible to provide testable *point predictions*, that is,

[31] The global games literature requires that it is (almost) common certainty that types have (approximately) uniform rank beliefs (i.e., $F(t \mid t)$ is arbitrarily close to $\frac{1}{2}$ for all t) (Morris et al., 2016).

predictions that hold across introspective type spaces for given payoff parameters. That is, if an analyst has no information on players' impulses or beliefs beyond those in Section 2.2, can he make any predictions about players' behavior or the value for a given game form $\boldsymbol{u}$?

Since the class of introspective type spaces consistent with our assumptions is quite general (cf. Section 5.1), requiring that predictions hold for all type spaces in this class is a stringent criterion. However, as illustrated by our results for the stylized game form G_w (Proposition 2), it is possible to make some progress on this question: For example, even if an analyst does not know the introspective type space, he knows that the value of G_w with $w = 1$ lies strictly between $\frac{1}{2}$ and 1. Rather than focusing on specific game forms as in Proposition 2, here we take a more abstract approach and focus on the question of what restrictions introspective equilibrium imposes in general games.

A first observation is that, because any introspective equilibrium induces a correlated equilibrium (Proposition 1), introspective equilibrium rules out any behavior that is not consistent with correlated equilibrium. This immediately implies that players' behavior is positively correlated in the sense that $\mu_{11}\,\mu_{22} \geq \mu_{12}\,\mu_{21}$, where μ_{nm} is the probability of action profile (s^n, s^m) in equilibrium (Calvó-Armengol, 2006).

We can say more, though: The following result shows that introspective equilibrium imposes restrictions on behavior over and above restrictions imposed by correlated equilibrium.[32]

Proposition 5. *For any coordination game $\boldsymbol{u}$, the set of correlated equilibria induced by an introspective equilibrium has Lebesgue measure 0 in the set of all correlated equilibria.*

Proposition 5 shows that the set of introspective equilibria is small (in the standard measure-theoretic sense) relative to the set of all correlated equilibria. Intuitively, any introspective equilibrium is symmetric (Proposition 1) and strict (for generic $\boldsymbol{u}$), while there are many asymmetric or non-strict correlated equilibria.[33,34]

[32] Proposition 5 relies on the assumptions we have made on the type space. Without any restrictions on the type space, it follows from a version of the revelation principle that any correlated equilibrium can be an introspective equilibrium: Given any correlated equilibrium of a game form $\boldsymbol{u}$, simply choose (F, τ^0) such that the action distribution induced by σ^0 coincides with that of the correlated equilibrium. This issue is of course not unique to introspective equilibrium. Battigalli and Siniscalchi (2003) show that a similar conundrum arises for Bayes-Nash equilibrium: If no restrictions are imposed on players' beliefs, then any rationalizable action can be played (also see Battigalli et al., 2011). In a similar vein, Bergin and Lipman (1996) show that if no restrictions are placed on mistake probabilities, then evolutionary dynamics cannot rule out any strict Nash equilibrium.

[33] Proposition 5 relies on (SYM). However, we conjecture that a similar result obtains if we allow for more general forms of symmetries that reflect the relationships between social roles (cf. Crawford and Haller, 1990).

[34] As an example of the type of behavior that introspective equilibrium rules out, note that the action distribution $(\mu_{11}, \mu_{22}, \mu_{12}, \mu_{21}) = (1/3, 1/3, 1/3, 0)$ is a correlated equilibrium of G_w for $w = 1$. However, since it is not symmetric or strict, it is not consistent with introspective equilibrium. It is not the case that any symmetric

To provide a full characterization of all action distributions consistent with introspective equilibrium for a given game form (by taking the union over all type spaces) seems challenging. For example, while the set of (symmetric, strict) correlated equilibria is convex, the set of introspective equilibria need not be.[35] The key difficulty is that while equilibrium incentives are described by players' rank beliefs $F(\tau \mid \tau)$, players' behavior is given by the joint distribution $F(t, t')$. So, characterizing the set of all action distributions that can arise in introspective equilibrium (for any introspective type space) requires translating conditions on conditional probabilities (viz., (MON-B)) and rank beliefs (viz., (NMRB)) to restrictions on joint distributions.[36] Without further restrictions on type spaces, it can thus be difficult to derive strong testable point predictions. Since reasonable restrictions on players' beliefs may vary by application, obtaining testable point predictions requires a better understanding what type of restrictions on beliefs are reasonable in different settings. Recently developed experimental protocols (e.g., Agranov et al., 2015) may aid with that.

5.3 Payoff-sensitive impulses

Thus far, we have assumed that economic and social factors can be perfectly separated in that economic factors are captured by the game form $\boldsymbol{u}$ while sociocultural factors are modeled by the introspective type space $\mathcal{T}$. While the assumption that impulses are driven entirely by social factors might not be unreasonable for decisions that have a strong cultural, moral, or ideological component, in other settings this assumption might perhaps be too strong. However, we can relax this assumption at least to some extent. Fix a game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$, where $\mathcal{T} = (F, \tau^0)$. Suppose there is a change in payoffs, i.e., the game form is now $\tilde{\boldsymbol{u}}$. A natural assumption is that when an action becomes more attractive in terms of payoffs, then players are more likely to choose it, i.e., the level-0 threshold increases with the dominance parameter.[37] That is, following the change in payoffs, the game is now $\widetilde{\mathcal{G}} = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}})$, with $\widetilde{\mathcal{T}} = (F, \tilde{\tau}^0)$ such that $\tilde{\tau}^0 < \tau^0$ if $\rho(\tilde{\boldsymbol{u}}) < \rho(\boldsymbol{u})$ and $\tilde{\tau}^0 \geq \tau^0$ otherwise. The following result shows that as long as the change in impulse distribution is not too large, it does not affect the introspective equilibrium or the value:

action distribution is a correlated equilibrium; for example, $(\mu_{11}, \mu_{22}, \mu_{12}, \mu_{21}) = (0, 0, 1/2, 1/2)$ is symmetric, but it is not a correlated equilibrium for any coordination game with two strict Nash equilibria.

[35] To see this, note that even if a subset of joint distributions form a convex set (as is the case for correlated equilibrium), the corresponding rank beliefs need not.

[36] In this respect, it is harder to obtain testable point predictions for introspective equilibrium than for solution concepts that predict uncorrelated play, as for these concepts, it is sufficient to consider a single feature of the distribution, the marginal. This is because for concepts that predict uncorrelated play, the rank belief $F(\tau \mid \tau)$ reduces to the marginal distribution $F(\tau)$ and the joint distribution is simply the product of marginals.

[37] This captures a wide variety of payoff effects. For example, this can model situations where players are more likely to have an impulse to choose an action with the highest minimum or maximum payoff.

Corollary 1. *Let $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$ and $\widetilde{\mathcal{G}} = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}})$ be as defined above. Then, if $|\tau^0 - \tilde{\tau}^0|$ is not too large, the introspective equilibrium (resp. value) of $\widetilde{\mathcal{G}}$ coincides with the introspective equilibrium (resp. value) of the game $\widetilde{\mathcal{G}}' = (\tilde{\boldsymbol{u}}, \mathcal{T})$ with the original type space.*

The proof is a straightforward adaptation of the proof of Proposition 4, so we refer to this result as a corollary. As before, the intuition is that as long as the level-0 threshold remains in the "basin of attraction" of the original introspective equilibrium, then this will not affect the introspective equilibrium or the value. As the example in Figure 1(b) illustrates, the introspective equilibrium and value remain unchanged even if the change $|\tau^0 - \tilde{\tau}^0|$ in level-0 thresholds is quite substantial (unless the original level-0 threshold is close to the boundary of a basin of attraction). So, in this sense, our results are robust to relaxing the assumption that impulses are independent of payoffs, for any game that satisfies our assumptions in Section 2.

For specific applications, it may be possible to extend this robustness result. In particular, for a particular class of introspective type spaces, there could be intuitive ways in which the distribution function $F(\cdot)$ changes with payoffs. For example, in the social salience type space, an action may be more likely to be socially salient when its payoffs increase. While such substantive changes to the type space may change some of the more quantitative predictions (e.g., $\underline{\rho}, \overline{\rho}$, the exact equilibrium threshold τ), our more qualitative insights – and thus our testable implications – largely extend. Intuitively, belief changes that are aligned with payoff changes reinforce each other in monotone supermodular games like ours. That makes our predictions robust to changes in beliefs that align with payoff changes like the ones considered here.

5.4 Relation to adaptive dynamics

This section discusses the methodological connection between the introspective process and the adaptive processes studied in the literature on evolution and learning in games. The introspective process is most closely related to the (myopic) *best response dynamic* (or: Cournot tatônnement). The best response dynamic assumes that in each time period $\tilde{t}$, each player j chooses a best response to the opponent's action in period $\tilde{t}$–1, much like players at level k choose a best response to the opponent's level-$(k-1)$ strategy in our model. The key difference is that introspective players form beliefs about others by considering their own mental state, while under the best response dynamic agents do not rely on their mental state to form beliefs about others. More precisely, introspective players form beliefs by conditioning on their (private) type: The action of a type t at level k is a best response to the type's posterior belief $\mu_{-j}^{k-1}(\cdot \mid t)$ about the other player's action at level $k-1$ (where $\mu_{-j}^{k-1}(s^2 \mid t) = F(\tau^{k-1} \mid t)$). By contrast, under the best response dynamic, players do not condition their beliefs on any private information. This may appear to be a small difference, but the consequences are profound. For example, for our class

of games, the standard best response dynamic either does not converge or it converges to one of the pure Nash equilibria.[38] By contrast, the introspective process always converges and allows for miscoordination (i.e., miscoordination can be attracting; Proposition 1). In fact, even if we consider stronger notions of stability, miscoordination remains stable. In particular, refinements based on asymptotic stability do not eliminate miscoordination. To formalize this claim, fix a game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$, where $\mathcal{T} = (F, \tau^0)$, and let τ be the equilibrium threshold for the introspective equilibrium. That is, τ is the limit of the level-k thresholds τ^k as $k \to \infty$. For every $t \in T$ and $\varepsilon > 0$, let $B_\varepsilon(t) := \{t' \in T : |t - t'| < \varepsilon\}$ be the ε-ball around t. Then, introspective equilibrium is *asymptotically stable* if it is both attracting and Lyapunov stable, i.e., the equilibrium threshold τ satisfies the following conditions:

(ATTR) There is $\varepsilon > 0$ such that for any $\tilde{\tau}^0 \in B_\varepsilon(\tau)$, the introspective process $\{\tilde{\tau}^k\}_k$ starting at $\tilde{\tau}^0$ converges to τ, i.e., $\lim_{k\to\infty} \tilde{\tau}^k = \tau$; and

(LYAP) For every $\eta > 0$, there is $\delta > 0$ such that if $\tilde{\tau}^0 \in B_\delta(\tau)$, then the introspective process $\{\tilde{\tau}^k\}_k$ starting at $\tilde{\tau}^0$ remains in $B_\eta(\tau)$, i.e., $\tilde{\tau}^k \in B_\eta(\tau)$ for all $k \geq 0$.

The following result shows that introspective equilibrium is asymptotically stable for generic payoff parameters:

Proposition 6. [**Asymptotic Stability**] *Introspective equilibrium is generically asymptotically stable in coordination games.*

Combined with Proposition 1, Proposition 6 implies that miscoordination is asymptotically stable whenever the payoff structure of the game provides little guidance (i.e., ρ intermediate) and (NMRB) holds.[39] Intuitively, introspective equilibria, including interior equilibria with miscoordination, have an open neighborhood (basin of attraction) such that, following a perturbation, the introspective process remains in that neighborhood and converges back to original equilibrium. Standard adaptive dynamics cannot capture this: under the best response dynamic and a wide variety of other dynamics, miscoordination is not asymptotically stable (Echenique and Edlin, 2004).

[38] An example where the best response dynamic does not converge is the following: At $t = 1$, one player assigns probability 1 to action s^1 while the other player assigns probability 1 to action s^2 (i.e., $\mu_j^0(s^1) = \mu_{-j}^0(s^2) = 1$). One way to ensure convergence is to assume that the best response dynamic starts from an "extremal" strategy profile (Vives, 1990, Thm. 5.1). However, this approach rules out miscoordination. Our approach of imposing restrictions on the introspective type space (while allowing the level-0 strategies to be "non-extremal") ensures that the introspective process converges yet allows for miscoordination.

[39] Miscoordination is not asymptotically stable (or even attracting) when (NMRB) does not hold. When (NMRB) does not hold, then there is miscoordination in introspective equilibrium only in the knife-edge case that $F(\tau^0 \mid \tau^0) = 1 - \rho$ (Proposition 1(c) and Theorem 1).

Miscoordination can be stable in richer adaptive models, but the predictions of these models are otherwise fundamentally different than the ones we obtain. This is the case, for example, for models that feature incomplete information about payoffs (but maintain the assumption that players do not use their private information to form beliefs about the other player).[40] In these models, mixed Nash equilibrium can be asymptotically stable (Sandholm, 2007).[41] Thus, miscoordination can be asymptotically stable in payoff-based extensions of standard dynamics. However, because mixed Nash equilibrium predicts that players choose an action less when its payoffs improve, these models do not always deliver intuitive comparative statics even for stable outcomes. By contrast, both the stable and (nongeneric) unstable outcomes of introspective equilibrium deliver intuitive comparative statics (Proposition 1(d)).[42] Other adaptive models that predict miscoordination are silent on which outcome is selected (e.g., Foster and Vohra, 1997; Fudenberg and Levine, 1999; Hart and Mas-Colell, 2000; Hart, 2005; Metzger, 2018). These models therefore cannot address the key question of how strategic behavior and the value change when payoffs are varied. So, introspective equilibrium is the only approach that both provides intuitive comparative statics (on behavior and the value) and predicts that miscoordination can be stable.

We can also go a step further and analyze the introspective process as an adaptive process. That is, rather than assuming that the levels in the introspective process are merely constructs in a player's mind, as we have done so far, we could alternatively assume that the process takes place over time. One way to model this is to view each introspective type as an agent, and the distribution $f(\cdot)$ as a local interaction network that governs the interactions between agents.[43] A novel question that arises in this context under what conditions an action can spread to the entire population (cf. Morris, 2000). An immediate implication of Theorem 1 is that contagion is likely to occur when the payoff structure makes one of the actions more attractive than the other (i.e., ρ close to 0 or 1). A novel and more subtle prediction is that if agents with a high label (i.e., t large) assign a lower weight than agents with a low label (i.e., t small) to agents having a lower label than themselves (i.e., (NMRB) is satisfied), then it may be more difficult to

[40] That is, expected payoffs are calculated using unconditional beliefs (i.e., not conditioned on the player's payoff type). This could be (though need not be) because payoff types are independent.

[41] However, whether this is the case depends on the class of perturbations being considered (Ellison and Fudenberg, 2000; Echenique and Edlin, 2004; Sandholm, 2007).

[42] This may seem to conflict with the observation that mixed Nash equilibrium can be consistent with introspective equilibrium (namely, if $F(\tau^0 \mid \tau^0) = 1 - \rho$). However, if mixed Nash equilibrium is consistent with introspective equilibrium, it must be at an isolated value of ρ. Hence, its comparative statics properties do not affect the comparative statics of introspective equilibrium. A similar comment applies to potential other unstable introspective equilibria.

[43] For papers that consider the links with local interaction games, see Mailath et al. (1997), Morris (1997, 2000), and Kets (2011); see Cripps (1991) and Lenzo and Sarver (2006) for a discussion in the context of evolution.

get contagion (i.e., $\tau \in (0,1)$). Which natural properties of networks ensure that this condition holds is a tantalizing question we leave for future research.

5.5 Related literature

This section summarizes related work not discussed elsewhere in the paper. The idea that strategic uncertainty can give rise to both miscoordination and coordination failure has a long history in experimental economics (e.g., Van Huyck et al., 1990, pp. 235–236). There is also ample evidence that social factors are a central determinant of behavior in games with multiple equilibria (Bacharach and Bernasconi, 1997) and that there can be a nontrivial interaction between payoff considerations and non-economic factors (Crawford et al., 2008). We connect these ideas by developing a theoretical model that delivers testable hypotheses on how the relative strength of economic and non-economic factors affects the value, by how they influence the scope for miscoordination and coordination failure.

The predictions we obtain are intuitive yet difficult to obtain using other approaches. Nash and correlated equilibrium cannot capture the intuition that there can be a qualitative change in behavior even when the equilibrium set remains the same (as in G_w and $\widetilde{G}_x$ when w or x is changed). Equilibrium refinements often emphasize the risk of coordination failure but ignore miscoordination. This includes risk dominance (Harsanyi and Selten, 1988), the global games selection (Carlsson and van Damme, 1993; Morris and Shin, 2003), and learning-based refinements (Kandori et al., 1993; Young, 1993).[44,45] On the other hand, mixed Nash equilibrium predicts miscoordination but cannot account for coordination failure. And while some behavioral models, such as level-k models (Crawford et al., 2013) and quantal response equilibrium (McKelvey and Palfrey, 1995) can model both miscoordination and coordination failure, which of the two arises depends on the assumptions on players' rationality: if players are fully rational (as they are in our framework), these models allow for coordination failure but cannot capture miscoordination; and if players are boundedly rational, they allow for miscoordination but cannot capture coordination failure. Thus, these models are silent on how the scope for miscoordination and coordination failure varies with payoff parameters when the assumptions on players' rationality are held fixed, as in our model.

[44] Other prominent evolutionary models select the efficient equilibrium (Robson and Vega-Redondo, 1995) or select different equilibria depending on the economic environment (Binmore and Samuelson, 1997), features of the learning process (Crawford, 1995), or on initial conditions (Samuelson, 2002). Also see Section 5.4 for a discussion of the relation with adaptive dynamics beyond the question of equilibrium selection.

[45] There are also equilibrium refinements that select the efficient outcome, such as payoff dominance (Harsanyi and Selten, 1988) or refinements that require predictions to be robust to perturbing the assumption that the extensive form is common knowledge (Penta and Zuazo-Garin, 2021).

The idea that social factors can be an important determinant of coordination has motivated a literature that shows how players can exploit salient action labels (e.g., Bacharach, 1993; Sugden, 1995) or precedent (Crawford and Haller, 1990) to improve coordination. However, this literature largely abstracts away from coordination failure. Another important distinguishing feature of our approach relative to this literature is that while the existing literature takes great care in modeling how particular non-economic factors influence behavior, our approach is largely "detail-free" in that it is agnostic as to which particular social factors drive behavior. Instead, we impose general assumptions on the introspective type space (i.e., (SYM)–(REG)) and show that our results for any type space that satisfies those assumptions. While this means we lose some of the richness of more detailed models, it has the advantage that it allows us to derive testable hypotheses that are independent of the details of the relevant non-economic factors.

Our work is also very different from the literature that posits that social factors can act as an equilibrium selection device. This prominent approach, which goes back to the seminal work of Schelling (1960), can help explain why societies that are essentially identical in all payoff-relevant effects may behave very differently.[46] However, because these models have multiple equilibria, they do not deliver testable comparative statics on the value. By contrast, our approach delivers testable hypotheses on how the value changes with economic primitives and delivers new policy implications.

6 Conclusions

This paper develops a novel theory of the value of coordination games. Relative to the existing literature, which focuses on comparative statics on equilibrium behavior, this theory delivers testable comparative statics on welfare (i.e., the value). While equilibrium behavior is monotone in payoffs, welfare need not be. As a result, policies that change behavior in the desired direction can reduce welfare. Likewise, policies that do not have any apparent downside in that they increase everyone's payoffs may make everyone worse off. These novel effects arise because policies generally have both direct payoff effects and indirect strategic effects: A policy that increases the payoffs to one of the actions (leaving other payoffs unchanged) has a positive direct payoff effect (everyone gets a (weakly) higher payoff assuming behavior remains unchanged) but can have negative indirect strategic effects (i.e., the policy changes how the game is played, in a way that reduces welfare). This tradeoff does not feature in comparative statics on equilibrium behavior but is critical for the current results. Moreover, because the theory provides a full characterization of when negative strategic effects dominate positive direct effects, the theory

[46] This approach has been applied widely in economics, see, e.g., Kreps (1990) on corporate culture, Greif (1994) on economic history, Ray (2004) on development, and Cass and Shell (1983) on sunspots.

also demonstrates how policies can be designed to avoid detrimental welfare effects. While we focus primarily on static settings, our results indicate that the effects we identify can be persistent. Thus, policies that ignore the key tradeoff between direct payoff effects and indirect strategic effects can have long-term detrimental welfare implications.

Our analysis also demonstrates that comparative statics on the value requires a more complex analyses than comparative statics on equilibrium behavior: While it suffices to consider a single "summary statistic" ρ of the payoffs to do comparative statics on equilibrium behavior, comparative statics on the value requires considering all payoff parameters. As the applications in Section 3.2 illustrate, this delivers a much richer picture: Changes in payoff parameters that lead to the same change in the summary statistic ρ can have different effects on the value. While the applications in Section 3.2 obviously do not exhaust the economic settings where comparative statics on the value can provide new insights,[47] they showcase the breadth of economic results that can be obtained. Moreover, the techniques developed there can be used to derive testable implications for other applications of interest.

A key feature of our approach is that behavior is driven by both economic and non-economic factors (e.g., salience, shared culture, animal spirits). Non-economic factors matter because they shape people's instinctive responses to a game. For example, a player may have an impulse to choose an action that has a salient label. However, players need not act on instinct: Before taking an action, they take the other player's perspective, and this may lead them to adjust their initial response. This approach is well-suited to model any setting where the payoff structure of the game does not fully pin down behavior and there is scope for the social context to influence play. As we show, the social context can help players reach consistent expectations when players face the same decision problem and have a shared understanding of the social context. An important question for future research is how the theory can be augmented to account for asymmetries in payoffs or social expectations, and how this affect the value of a game.

Appendix A The value under miscoordination

This appendix characterizes the value in the regime with miscoordination. Consider a coordination game with payoffs $\boldsymbol{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ and fix an introspective type space $\mathcal{T} = (F, \tau^0)$. Recall that, by Proposition 1, the introspective equilibrium is characterized by an equilibrium threshold $\tau \in [0, 1]$, which depends on the payoffs only through the dominance parameter $\rho = \rho(\boldsymbol{u})$ associated with $\boldsymbol{u}$. If the equilibrium threshold τ is differentiable with respect to

[47] For example, simply considering how the value changes with each of the 4 payoff parameters u_{nm} does not exhaust the possibilities as in many economic applications, a change in policy or economic primitives affect multiple payoff parameters simultaneously (e.g., Sections 3.2.2–3.2.3).

ρ, then the change in value with payoffs is given by

$$\nabla_{\boldsymbol{u}} V(\boldsymbol{u}; \mathcal{T}) = \frac{\partial V}{\partial \tau} \frac{\partial \tau}{\partial \rho} \nabla_{\boldsymbol{u}} \rho + \boldsymbol{p}(\tau), \tag{7}$$

where $\nabla_{\boldsymbol{u}} \phi$ denotes the gradient of a function ϕ with respect to the payoffs $\boldsymbol{u}$, and $\boldsymbol{p}(\tau) = (p_{11}(\tau), p_{12}(\tau), p_{21}(\tau), p_{22}(\tau))$ gives the probabilities $p_{nm}(\tau)$ that each action profile (s^n, s^m) is played in introspective equilibrium.[48] To see this, note that the total effect on the value of a change in payoffs can be decomposed into a direct payoff effect and indirect strategic effects. The direct effect is straightforward: If the payoff u_{nm} to action profile (s^n, s^m) is increased (keeping other payoffs fixed), then, conditional on players choosing (s, s'), each player receives a higher payoff. Hence, the change in value due to the direct effect of a change in u_{nm} is just the probability $p_{nm}(\tau)$ that players choose (s^n, s^m) in introspective equilibrium. This is the second term in (7). There are also indirect effects which arise because the equilibrium threshold τ changes with ρ (Proposition 1). These indirect effects are given by the first term in (7).

While the direct effect is always positive, the indirect effects may have either sign, as we show next. If there is no miscoordination in introspective equilibrium ($\rho < \underline{\rho}$ or $\rho > \overline{\rho}$), then the equilibrium threshold does not change with ρ (Theorem 1), so there are no strategic effects (i.e., $\nabla_{\boldsymbol{u}} V(\boldsymbol{u}; \mathcal{T}) = \boldsymbol{p}(\tau)$). So, suppose there is miscoordination in introspective equilibrium ($\rho \in (\underline{\rho}, \overline{\rho})$ and $\tau \neq 0, 1$). By Theorem 1 (Eq. (1)), we have

$$V(\boldsymbol{u}; \mathcal{T}) = u_{11} + (u_{21} + u_{12} - 2u_{11}) F(\tau) + \frac{u_{11} - u_{21}}{1 - \rho} F(\tau, \tau).$$

Note that by symmetry,

$$F(\tau, \tau) = 2 \int_0^\tau F(t \mid t) f(t) \, \mathrm{d}t.$$

Hence, using that $F(\tau \mid \tau) = 1 - \rho$, the partial derivative of the value with respect to τ is given by

$$\frac{\partial V}{\partial \tau} = (u_{21} + u_{12} - 2u_{11}) f(\tau) + 2(u_{11} - u_{21}) f(\tau) = (u_{12} - u_{21}) f(\tau).$$

We conclude that

$$\nabla_{\boldsymbol{u}} V(\boldsymbol{u}; \mathcal{T}) = (u_{12} - u_{21}) f(\tau) \frac{\partial \tau}{\partial \rho} \nabla_{\boldsymbol{u}} \rho + \boldsymbol{p}(\tau).$$

As $F(\tau \mid \tau) = 1 - \rho$, we have

$$\frac{\partial \tau}{\partial \rho} = \left(- \frac{\mathrm{d} F(\tau \mid \tau)}{\mathrm{d} \tau} \right)^{-1},$$

[48] This is with some abuse of notation; strictly speaking, the vectors $\boldsymbol{u}$ and $\boldsymbol{p}$ are column vectors, not row vectors.

provided that $F(t \mid t)$ is differentiable at $t = \tau$. Hence, the comparative statics of the value under miscoordination is given by

$$\nabla_{\boldsymbol{u}} V(\boldsymbol{u}; \mathcal{T}) = (u_{12} - u_{21}) f(\tau) \left(-\frac{\mathrm{d}F(\tau \mid \tau)}{\mathrm{d}\tau} \right)^{-1} \nabla_{\boldsymbol{u}} \rho + \boldsymbol{p}(\tau).$$

The sign of the indirect effects thus depends on how ρ changes with payoffs (e.g., ρ increases with u_{22} but decreases with u_{11}) and the relative magnitude of u_{12} and u_{21}.

Appendix B Example type spaces

This appendix shows that the social salience type space (Section 3.2.1) and the "animal spirits" type space (Section 3.2.2) satisfy our conditions.

B.1 Social salience

This section formally defines the introspective type space in Section 3.2.1 and shows that it satisfies conditions (SYM)–(REG). We also characterize the conditions under which the type space satisfies (NMRB).

The idea is that an action may be "socially salient" in the sense that players are likely to have an impulse to choose that action. More precisely, there is a (latent) state $\theta \in \{s^1, s^2\}$ representing which action is "socially salient," where the probability that action s^1 is socially salient ($\theta = s^1$) is $p \in (0, 1)$. Conditional on $\theta = s^n$, each player j has an impulse to choose action s^n with probability $q_j \in (\frac{1}{2}, 1)$, where the parameter q_j is a measure of how sensitive player j is to social cues and is drawn from a continuous density $g(\cdot)$ with full support on $[\frac{1}{2}, 1)$ independently across players.

We define an introspective type space as follows: We identify the type of a player with impulse I_j and parameter q_j with its posterior belief $t_j = t_j(I_j, q_j)$ that $\theta = s^1$. More precisely, depending on whether $I_j = s^1$ or $I_j = s^2$, the introspective type of the player is given by

$$t_j(s^1, q_j) = \mathbb{P}(\theta = s^1 \mid I_j = s^1, q_j) = \frac{p\, q_j}{p\, q_j + \hat{p}\, \hat{q}_j},$$
$$t_j(s^2, q_j) = \mathbb{P}(\theta = s^1 \mid I_j = s^2, q_j) = \frac{p\, \hat{q}_j}{p\, \hat{q}_j + \hat{p}\, q_j};$$

where we have introduced the notation $\hat{x} := 1 - x$ for a given variable x. We denote this introspective type space by $\mathcal{T}^*$; note that it is parameterized by p and $g(\cdot)$.

This introspective type space satisfies the conditions in Section 2 for any p and $g(\cdot)$:

Proposition 7. *The introspective type space $\mathcal{T}^*$ satisfies conditions (SYM), (MON-I), (MON-B), and (REG).*

Proof. A first observation is that for given q_j,

$$\mathbb{P}(I_j = s^1, \theta = s^1 \mid q_j) = p\, q_j, \quad \mathbb{P}(I_j = s^2, \theta = s^1 \mid q_j) = p\, \hat{q}_j,$$
$$\mathbb{P}(I_j = s^2, \theta = s^2 \mid q_j) = \hat{p}\, q_j, \quad \mathbb{P}(I_j = s^1, \theta = s^2 \mid q_j) = \hat{p}\, \hat{q}_j,$$

Clearly, condition (SYM) is satisfied. We next show that condition (MON-I) holds. It is easy to verify that for $I_j = s^2$, the type is strictly decreasing in q_j and takes values between 0 and p, while for $I_j = s^1$, the type is strictly increasing in q_j and takes values between p and 1. Every introspective type $t_j \neq p$ is therefore associated with a unique pair (I_j, q_j); moreover, types $t_j > p$ have an impulse to choose s^1, while types $t_j < p$ have an impulse to choose s^2. Hence, the introspective type space satisfies (MON-I) with threshold $\tau^0 = p$.

We next show that condition (MON-B) holds. Consider a player with introspective type t and corresponding parameter q. By inverting the relations above between a player's type and the parameter q, we find that the parameter $q = q(t)$ that corresponds to type t is

$$q(t) = \begin{cases} r(t) & t \in [0, p); \\ \hat{r}(t) & t \in [p, 1]; \end{cases} \tag{8}$$

where

$$r(t) = p\,\hat{t}\,(p\,\hat{t} + \hat{p}\,t)^{-1}.$$

We use this to calculate the density of introspective types in terms of the density $g(q)$. We need to take into account that the relation between the parameter q and the type t is not one-to-one: An introspective type $t < p$ is associated (uniquely) with the pair $(I_j = s^2, q_j = q(t))$ and a type $t > p$ is associated with the pair $(I_j = s^1, q_j = q(t))$. So, the density $f(t)$ of $t < p$ is thus $g(q(t))|q'(t)|$ times the conditional probability that $I_j = s^2$ given that $q_j = q(t)$, and analogously for introspective types $t > p$ (where $q'(t)$ is the derivative of $q(t)$ with respect to t). If we define $I(t) = s^2$ for $t < p$ and $I(t) = s^1$ for $t \geq p$, then, for all $t \in [0, 1]$,

$$\mathbb{P}\big(I_j = I(t) \mid q_j = q(t)\big) = p\hat{r}(t) + \hat{p}r(t).$$

Hence, the probability density of the introspective type of a player is given by

$$f(t) = (p\hat{r}(t) + \hat{p}r(t))\, g(q(t))\, |q'(t)| = p\hat{p}\,(p\hat{t} + \hat{p}t)^{-1}\, g(q(t))\, |q'(t)|$$

for all $t \in [0,1]$. Similarly, the joint probability density[49] of types t and u for the two players is given by

$$f(t,u) = (p\hat{r}(t)\hat{r}(u) + \hat{p}r(t)r(u))\, g(q(t))\, g(q(u))\, |q'(t)|\, |q'(u)|,$$

where $t, u \in [0,1]$. Note that $f(t)$ and $f(t,u)$ are well-defined at $t = p$ and $u = p$, since $\lim_{t \downarrow p} q'(t) = \lim_{t \uparrow p} -q'(t) = 1/(4p(1-p))$. Dividing the joint density by $f(t)$ yields

$$f(u \mid t) = (t\hat{r}(u) + \hat{t}r(u))\, g(q(u))\, |q'(u)|.$$

By integrating with respect to u we obtain

$$F(\tau \mid t) = \int_{q(\tau)}^{1} (t\hat{q} + \hat{t}q) g(q)\, \mathrm{d}q$$

for $\tau \in [0,p)$, while for $\tau \in [p,1]$,

$$1 - F(\tau \mid t) = \int_{q(\tau)}^{1} (tq + \hat{t}\hat{q}) g(q)\, \mathrm{d}q.$$

In particular, observe that for all $\tau \in (0,1)$ the derivative of $F(\tau \mid t)$ with respect to t is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\tau \mid t) = -\int_{q(\tau)}^{1} (q - \hat{q}) g(q)\, \mathrm{d}q = -\int_{q(\tau)}^{1} (2q-1) g(q)\, \mathrm{d}q.$$

Since the density g has full support on $[\frac{1}{2}, 1)$, it follows that the introspective type space satisfies (MON-B).

We next show that condition (REG) is satisfied. As a first step, we rewrite the expressions for the rank belief function as

$$F(t \mid t) = \begin{cases} \displaystyle\int_{q(t)}^{1} (q - t(2q-1)) g(q)\, \mathrm{d}q & t \in [0,p); \\ \displaystyle 1 - \int_{q(t)}^{1} ((1-q) + t(2q-1)) g(q)\, \mathrm{d}q & t \in [p,1]. \end{cases} \tag{9}$$

Since $\lim_{t \downarrow 0} q(t) = \lim_{t \uparrow 1} q(t) = 1$, it follows that $\lim_{t \downarrow 0} F(t \mid t) = 0$ and $\lim_{t \uparrow 1} F(t \mid t) = 1$. Together with the expression for $f(t,u)$ above, this shows that the introspective type space satisfies (REG). □

We next characterize the conditions under which the introspective type space $\mathcal{T}^*$ satisfies (NMRB):

[49] This is with some abuse of notation as we are using the same symbol f with different meanings; however, it should be clear from the arguments of the function whether we mean the density of a single (introspective) type, the joint density of two types, or the conditional density of one type given another type.

Proposition 8. *The introspective type space $\mathcal{T}^* = \mathcal{T}^*(p, g)$ satisfies condition (NMRB) if and only if $g(\frac{1}{2}) < 8p(1-p)\mathbb{E}(2q-1)$.*

Proof. Recall that condition (NMRB) says that there is $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$, or there is $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both). We show that the rank belief function $F(\cdot \mid \cdot)$ for $\mathcal{T}^*$ is differentiable. Then, (NMRB) holds if and only if the derivative of the rank belief function is negative at $\tau^0 = p$. We first calculate the derivative of the rank belief function. Using that $r(t) = q(t)$ for $t < p$ and $r(t) = 1 - q(t)$ for $t > p$, we see that for all $t \neq p$,

$$\frac{\mathrm{d}}{\mathrm{d}t} F(t \mid t) = -r(t)\, g(q(t))\, r'(t) + t(2q(t)-1)\, g(q(t))\, q'(t) - \int_{q(t)}^{1} (2q-1) g(q)\, \mathrm{d}q.$$

In particular, since $\lim_{t \to p} q(t) = \frac{1}{2}$ and $\lim_{t \uparrow p} r'(t) = \lim_{t \downarrow p} r'(t) = -1/(4p(1-p))$, it follows that $F(t \mid t)$ is also differentiable at $t = p = \tau^0$, and

$$\frac{\mathrm{d}}{\mathrm{d}t} F(t \mid t)\Big|_{t=p} = \frac{g(\frac{1}{2})}{8p\,(1-p)} - \mathbb{E}(2q-1).$$

Hence, $\mathcal{T}^*$ satisfies (NMRB) whenever $g(\frac{1}{2}) < 8p\,(1-p)\,\mathbb{E}(2q-1)$. □

Thus, under the assumption in Section 3.2.1 that $q_j > \frac{1}{2}$ (with probability 1, Assumption (NMRB) holds.

Finally, a sufficient condition under which no action is strongly salient in $\mathcal{T}^*$ is that $p = \frac{1}{2}$:

Proposition 9. *Suppose $\mathcal{T}^*$ satisfies (NMRB). Then no action is strongly salient whenever $p = \frac{1}{2}$.*

Proof. Recall that $\tau^0 = \frac{1}{2}$ for $p = \frac{1}{2}$. By (NMRB), the interval $(\underline{\rho}, \overline{\rho})$ is nonempty, and by (9), $F(\frac{1}{2} \mid \frac{1}{2}) = \frac{1}{2}$. □

By continuity, introspective type spaces for which p is sufficiently close to $\frac{1}{2}$ also satisfy the condition that no action is strongly salient.

The introspective type space $\mathcal{T}^*$ was used to generate Figures 1–3, where we have taken $p = \frac{1}{2}$ and $g(\cdot)$ is the truncated normal distribution with mean 7/8 and variance 1/64. By Propositions 7–9, this introspective type space satisfies (SYM)–(REG) and (NMRB), and has the property that no action is strongly salient.

B.2 Animal spirits

This section shows that the animal spirits type space, which is derived from the type space in Morris and Yildiz (2019), is a special case of our framework. That is, under the assumptions in Morris and Yildiz (2019), the introspective type space in Section 3.2.2 satisfies (SYM)–(REG) as well as (NMRB).

We follow the exposition Morris and Yildiz (2019, Sec. I). Each player's type is the sum of a common shock η that affects both players, and an idiosyncratic noise term ε_j that varies across players. That is, the type for player j is

$$\tilde{t}_j = \eta + \varepsilon_j,$$

where ε_j and η are drawn independently across players from distributions $\widetilde{F}$ and $\widetilde{G}$, respectively. The distributions $\widetilde{F}$ and $\widetilde{G}$ are assumed to have positive continuous densities $\tilde{f}$ and $\tilde{g}$ everywhere on $\mathbb{R}$. The densities $\tilde{f}$ and $\tilde{g}$ are taken to be symmetric around zero, i.e., $\tilde{f}(\varepsilon) = \tilde{f}(-\varepsilon)$ and $\tilde{g}(\eta) = \tilde{g}(-\eta)$. Moreover, both densities are weakly decreasing on $(0, \infty)$. By symmetry both the idiosyncratic and the common shock have zero mean. The distribution of idiosyncratic shocks is taken to be log-concave (i.e., $\log \tilde{f}$ is concave). The distribution of common shocks has regularly-varying tails, that is, for all $\eta, \eta' \in (0, \infty)$,

$$\lim_{\lambda \to \infty} \frac{\tilde{g}(\lambda\eta)}{\tilde{g}(\lambda\eta')} \in (0, \infty).$$

Together, $\tilde{f}$ and $\tilde{g}$ define a joint distribution $F_{shocks}(\tilde{t}_1, \tilde{t}_2)$ on $(-\infty, +\infty) \times (-\infty, +\infty)$ with corresponding joint density $f_{shocks}(\tilde{t}_1, \tilde{t}_2)$.

Because the densities $\tilde{f}$ and $\tilde{g}$ have full support on $\mathbb{R}$, we need to apply a (continuous order-preserving) transformation $\tilde{t} \mapsto t$ from $(-\infty, \infty)$ to $(0, 1)$ to ensure that each player's type lies between 0 and 1, as in our model. The particular choice of transformation is immaterial; however, given that the distribution of common shocks has fat tails, some care must be taken that the resulting density on $[0, 1]$ is integrable. Given an appropriate transformation, we take the set of introspective types to be $T = [0, 1]$, as before, with the joint distribution $F(t_1, t_2)$ derived from the original densities $\tilde{f}$ and $\tilde{g}$ by applying the transformation. In particular, if the transformation maps $\tilde{t}_1$ and $\tilde{t}_2$ into t_1 and t_2, respectively, then $f(t_1, t_2) = f_{shocks}(\tilde{t}_1, \tilde{t}_2)$. Because the transformation is continuous, f is a continuous density; by construction, it has full support on the interior of $T \times T$. Pick some $\tau^0 \in (0, 1) = T^\circ$ and define the function $\mathcal{I} : T \to \{s^1, s^2\}$ by $\mathcal{I}(t) = s^2$ if $t < \tau^0$ and $\mathcal{I}(t) = s^1$ for $t \geq \tau^0$.

Clearly, this introspective type space satisfies (SYM) and (MON-I). It is also not hard to check that it satisfies (MON-B) (by the log concavity of $\tilde{f}$). By Lemma 1 in Morris and Yildiz (2019), the introspective type space satisfies (REG); moreover, if τ^0 is sufficiently close to 0 or 1, it satisfies (NMRB) (cf. Figure 4).

This type space was used to generate Figures 4–5, using the following specifications: The common shock η has a Student's-t distribution with parameter $n = 4$ while the idiosyncratic shocks have a standard normal distribution.[50] For the transformation, it will be (numerically)

[50]The choice of n is partly governed by the choice of transformation from $(-\infty, \infty)$ to $(0, 1)$: For our choice

convenient to define a mapping from $[0,1]$ to $[-\infty,\infty]$ and then use its inverse to calculate the rank belief functions. We use the following mapping: We first use $t \mapsto 2t-1$ to map the type in $T=[0,1]$ to a type in $[-1,1]$, and then apply the map $t \mapsto t(1-|t|^{\alpha})^{-1}$ to map the type to $[-\infty,\infty]$, where the parameter α controls the shape of the transformation; we use $\alpha = 1/4$. To avoid numerical problems, we use a slight modification of this transformation where extreme types (beyond a cutoff R) are mapped into $\pm(1+(\alpha R)^{-1})$; we use $R=200$.

Appendix C Investment

This appendix shows that our results in Section 3.2.2 continue to hold if the investment subsidy is replaced by an investment "bonus" that players receive when they successfully invest. That is, we consider an increase in the payoff u_{11} that players receive if they both invest. The following result characterizes the conditions miscoordination is more costly than coordination failure in the sense that the value decreases as we move from the regime with coordination failure to the regime with miscoordination (i.e., as ρ falls to $\overline{\rho}$).

Theorem 4. [Investment Bonus] *Fix an introspective type space with a positive probability of investment at $\overline{\rho}$ (i.e., $\overline{\tau}<1$). Then, the value of a game with coordination failure in introspective equilibrium ($\rho>\overline{\rho}$) strictly decreases with the investment payoff u_{11} as it induces miscoordination (i.e., the dominance parameter falls to $\overline{\rho}$) if and only if the off-diagonal payoffs are sufficiently small and ρ is not too high: There is $\rho^c \in (\overline{\rho},1)$ such that for all u_{11} and u_{22} with $u_{11} \geq u_{22}$ and $\rho > \overline{\rho}$, there exist u_{12}^* and u_{21}^* such that the following holds: for any game form $\boldsymbol{u}=(u_{11},u_{12},u_{21},u_{22})$ with dominance parameter ρ, as u_{11} increases, the value falls below u_{22} at $\overline{\rho}$ if and only if $\rho<\rho^c$, $u_{12}<u_{12}^*$, and $u_{21}<u_{21}^*$.*

Proof. Let $(\overline{p_{11}},\overline{p_{12}},\overline{p_{21}},\overline{p_{22}})$ be the probability distribution over action profiles in introspective equilibrium for the given introspective type space when the dominance parameter is $\overline{\rho}$. Recall the definition of ρ^c in the proof of Theorem 2; as noted there, $\rho^c \in (\overline{\rho},1)$. Let $\boldsymbol{u}=(u_{11},u_{12},u_{21},u_{22})$ be a game with coordination failure in introspective equilibrium, i.e., $\rho=\rho(\boldsymbol{u})>\overline{\rho}$. As u_{11} increases, the dominance parameter decreases. Let $\overline{u_{11}}$ be the investment payoff for which the dominance parameter attains the value $\overline{\rho}$. Then,

$$\frac{1-\rho}{\rho}=\frac{u_{11}-u_{21}}{u_{22}-u_{12}} \qquad \text{and} \qquad \frac{1-\overline{\rho}}{\overline{\rho}}=\frac{\overline{u_{11}}-u_{21}}{u_{22}-u_{12}}. \tag{10}$$

of transformation (described below), we need $n>3$ to prevent the joint distribution of types from blowing up in the corners $(0,0)$ and $(1,1)$ of $T\times T$.

The difference in value between the games with investment payoffs u_{11} (with coordination failure) and $\overline{u_{11}}$ (with miscoordination) is

$$\begin{aligned}\Delta &= \overline{p_{11}}\,\overline{u_{11}} + \overline{p_{12}}\,u_{12} + \overline{p_{21}}\,u_{21} + \overline{p_{22}}\,u_{22} - u_{22} \\ &= \overline{p_{11}}\,(\overline{u_{11}} - u_{21}) - (\overline{p_{11}} + \overline{p_{12}})(u_{11} - u_{21}) + (\overline{p_{11}} + \overline{p_{12}})(u_{11} - u_{22}) \\ &\qquad - \overline{p_{12}}\,(u_{22} - u_{12}),\end{aligned}$$

where we have used $\overline{p_{12}} = \overline{p_{21}}$ and $\overline{p_{11}} + \overline{p_{12}} + \overline{p_{21}} + \overline{p_{22}} = 1$. Applying (10) to the factors $\overline{u_{11}} - u_{21}$ and $u_{11} - u_{21}$ (and reordering terms) gives

$$\Delta = (\overline{p_{11}} + \overline{p_{12}})(u_{11} - u_{22}) - \Big(\frac{\overline{p_{11}} + \overline{p_{12}}}{\rho} - \frac{\overline{p_{11}}}{\overline{\rho}}\Big)(u_{22} - u_{12}).$$

Hence, $\Delta < 0$ if and only if

$$\frac{u_{11} - u_{22}}{u_{22} - u_{12}} < \frac{1}{\rho} - \frac{\overline{p_{11}}}{\overline{\rho}\,(\overline{p_{11}} + \overline{p_{12}})} = \frac{\rho^c - \rho}{\rho\,\rho^c}.$$

As the left-hand side is non-negative, $\Delta < 0$ only if $\rho < \rho^c$. In that case, $\Delta < 0$ is equivalent to

$$\frac{u_{22} - u_{12}}{\rho} = \frac{u_{11} - u_{21}}{1 - \rho} > \frac{\rho^c}{\rho^c - \rho}\,(u_{11} - u_{22}),$$

where the equality follows from (10). □

Appendix D Collusion

We consider two other commonly-used models of collusion. In addition, because models of collusion have the structure of a social dilemma, we also study the classic repeated prisoner's dilemma. The results demonstrate that our results in Section 3.2.3 are robust.

D.1 Tourists and natives

This appendix studies the effects of a change in the competitiveness of the market as measured by the fraction of consumers who buy from the firm with the lowest price. We consider a simple model of price insensitive "tourists" and best-price shopping "natives", or, closer to most applications, loyal buyers and switchers (Salop and Stiglitz, 1977). There are two firms, labeled by $i \in \{1, 2\}$. In each period $\tilde{t} = 0, 1, \ldots$, firms choose a price $p \in \{H, L\}$, with $H > L > 0$ and a mass of consumers (of measure 1) decides which firm to buy from. A fraction $m_s \in (0, 1)$ of consumers are *switchers*: they buy from the firm with the lowest price. In addition, each firm

i has a mass $m_\ell \in (0, \frac{1}{2})$ of *loyal buyers* (i.e., $m_s + 2m_\ell = 1$). Firms' marginal cost is equal to 0, and we normalize by setting $L = 1$. Then, the payoffs in the one-shot game are given by

	H	L
H	$(m_\ell + \frac{1}{2}m_s)H, (m_\ell + \frac{1}{2}m_s)H$	$m_\ell H, m_\ell + m_s$
L	$m_\ell + m_s, m_\ell H$	$m_\ell + \frac{1}{2}m_s, m_\ell + \frac{1}{2}m_s$

As in Section 3.2.3, we assume that the one-shot game has the structure of a prisoner's dilemma (i.e., the low price L is a strictly dominant strategy for both firms). This is the case if and only if $H < 2(1 - m_\ell)$. Thus, in the absence of repetition, firms compete for the market.

For the repeated game, we again consider a collusive strategy and a cheating strategy. Under the collusive strategy, the firm chooses the high price H in every period as long as both firms chose the high price in all past periods; otherwise, it charges the low price L. Under the cheating strategy, the firm chooses the low price L in every period. Again, we identify the collusive strategy with s^1 and the cheating strategy with s^2. Then, using that $m_\ell + \frac{1}{2}m_s = \frac{1}{2}$ and $m_\ell + m_s = 1 - m_\ell$, the payoffs in the repeated game are given by

	s^1	s^2
s^1	$\frac{1}{2}H$	$(1-\delta)\, m_\ell H + \frac{1}{2}\delta$
s^2	$(1-\delta)(1-m_\ell) + \frac{1}{2}\delta$	$\frac{1}{2}$

where we have listed only the row player's payoffs. We again assume that collusion can be sustained as a subgame perfect equilibrium, i.e., $\delta > (2(1 - m_\ell) - H)/(1 - 2m_\ell)$. Then, the repeated game can be viewed as a coordination game, and coordination failure corresponds to both firms choosing the low price while miscoordination corresponds to one firm choosing the collusive strategy and the other firm choosing the cheating strategy.

The following result shows that, starting from a game with coordination failure, any change in payoff parameters that makes collusion less risky (i.e., decreases ρ) leads to a strict increase in the value even if it induces miscoordination:

Theorem 5. [Collusion: Tourists & Natives] *Fix an introspective type space and a game form such that there is coordination failure in introspective equilibrium. Then any change in payoff parameters that makes the dominance parameter smaller than $\overline{\rho}$ strictly increases the value of the game.*

Proof. As in Section 3.2.3, we consider a change of payoff parameters such that the dominance parameter decreases from $\rho > \overline{\rho}$ to $\rho' < \overline{\rho}$. Since $u_{11} > u_{22}$, the result clearly holds if the change in payoff parameters is such that firms choose the high price in introspective equilibrium (i.e., $\rho' < \underline{\rho}$). So suppose $\rho' \in (\underline{\rho}, \overline{\rho})$, and let δ and δ' be the discount factors in the game with

dominance parameters ρ and ρ', respectively. Similarly, let m_ℓ and m'_ℓ be the fraction of loyal buyers for a firm in the game with dominance parameters ρ and ρ', respectively. Then, the inequality (3) in Lemma 1 gives

$$\Delta > (p'_{11} + p'_{12})(1 - \delta')(\tfrac{1}{2} - m'_\ell).$$

The result then follows by noting that the right-hand side is positive. □

Examples of changes in the dominance parameter that make collusion less risky include an an increase in the fraction m_ℓ of loyal buyers, an increase in the discount factor δ, and an increase in the high price H. Theorem 5 shows that any of these changes make firms better off.

D.2 Homogeneous goods

This appendix studies the limiting case where the two firms produce identical goods (i.e., $b = c$). To avoid problems with the nonexistence of a pure Nash equilibrium, we assume that firms can undercut each other only by a fixed amount $\varepsilon > 0$ (taken to be small, i.e., $\varepsilon < \frac{1}{3}a$). The model is otherwise the same as in Section 3.2.3: In the repeated game, the collusive strategy σ^* is to choose the monopoly price p^* in every period as long as both firms chose the monopoly price in all past periods; and to choose the competitive price $p^N = 0$ otherwise. The cheating strategy σ^* is to choose $p^c := p^* - \varepsilon$ in every period as long as both firms chose the monopoly price in all past periods, and to choose the competitive price $p^N = 0$ otherwise. We normalize and set $b = 1$. We also define $\eta := 2\varepsilon$. Then, if we again identify the collusive and the cheating strategy with s^1 and s^2, the payoffs in the repeated game are given by

	s^1	s^2
s^1	$\frac{1}{8}a^2$	0
s^2	$(1-\delta)\,\frac{1}{4}(a^2-\eta^2)$	$(1-\delta)\,\frac{1}{8}(a^2-\eta^2)$

where we have listed only the row player's payoffs. We again assume that collusion can be sustained as a subgame perfect equilibrium, i.e., $\delta > (a^2 - 2\eta^2)/(2a^2 - 2\eta^2)$. As before, the repeated game can be viewed as a coordination game, and coordination failure corresponds to both firms choosing the cheating strategy while miscoordination corresponds to one firm choosing the collusive strategy and the other firm choosing the cheating strategy.

The following result considers the effects of an increase in the discount factor δ.

Theorem 6. [Collusion: Homogeneous Goods] *Fix an introspective type space and let* $\delta, \delta' \in \big((a^2 - 2\eta^2)/(2a^2 - 2\eta^2), 1\big)$ *be such that there is coordination failure in introspective equilibrium when the discount factor is* δ *(i.e,* $\rho(\delta) > \overline{\rho}$*) but not when the discount factor is* δ' *(i.e.,* $\rho(\delta') < \overline{\rho}$*).*

Then, if $\delta' - \delta$ is either sufficiently small or sufficiently large, the value of the game with discount factor δ' is strictly larger than the value of the game with discount factor δ (with coordination failure)

Proof. Define $\rho := \rho(\delta)$ and $\rho' := \rho(\delta')$ to be the dominance parameters for the games with discount factors δ and δ', respectively. Again, the result clearly holds if δ' is much smaller than δ so that firms choose the collusive strategy in introspective equilibrium (i.e., $\rho' < \underline{\rho}$). So suppose $\rho' \in (\underline{\rho}, \overline{\rho})$ Substituting the payoffs into (5) in Lemma 1 gives

$$8\Delta = \frac{p'_{11}}{\rho'}(1-\delta')(a^2-\eta^2) - (\delta'-\delta)(a^2-\eta^2).$$

The result then follows by noting that $p'_{11} > 0$ and $a > \eta$. □

D.3 Prisoner's dilemma

This appendix studies the classic prisoner's dilemma. In each period $\tilde{t} = 0, 1, \ldots$, players can either cooperate (play c) or defect (play d). The payoffs in the one-shot game are given by

	c	d
c	C, C	S, T
d	T, S	D, D

where $T > C > D > S$. In the repeated game, players choose between a cooperative strategy and always defect. Under the cooperative (grim trigger) strategy, the player cooperates in every period as long as both players cooperated in all past periods; otherwise, he defects. Under always defect, the player defects in every period. Again, we identify the cooperative strategy with s^1 and always defect with s^2. Then, the payoffs in the repeated game are given by

	s^1	s^2
s^1	C	$(1-\delta)S + \delta D$
s^2	$(1-\delta)T + \delta D$	D

where have listed only the row player's payoffs. We again assume that cooperation can be sustained as a strict subgame perfect equilibrium, i.e., $\delta > (T-C)/(T-D)$. We can then view the repeated game as a coordination game, where coordination failure corresponds to both players choosing to always defect while miscoordination corresponds to one player trying to initiate cooperation (i.e., choosing s^1) and the other player choosing to always defect. It will be convenient to write (δ, C, S, T, D) for the game form $\boldsymbol{u}$ and to denote the corresponding dominance parameter by $\rho(\delta, C, S, T, D)$.

We consider the effect of two types of changes: An increase in the discount factor δ and a decrease in the payoff D when cooperation breaks down. Both types of changes make cooperation less risky (i.e., ρ decreases with δ and increases with D). Hence, if we start from a game with coordination failure, both these changes can induce miscoordination. The following result shows that, again, players are better off when there is miscoordination than if there is coordination failure:

Theorem 7. [**Prisoner's Dilemma**] *Fix an introspective type space and a game form* $\boldsymbol{u} = (\delta, C, S, T, D)$ *with* $\delta \in \big((T-C)/(T-D), 1\big)$, *and let* $\delta' \in (\delta, 1)$ *and* $D' < D$.

(a) Suppose that there is coordination failure in introspective equilibrium when the discount factor is δ *but not when it is* δ' *(i.e.,* $\rho(\delta, C, S, T, D) > \overline{\rho}$ *and* $\rho(\delta', C, S, T, D) < \overline{\rho}$*). Then, the value of the game with the discount factor* δ' *is strictly higher than that of the game with the discount factor* δ.

(b) Suppose that there is coordination failure in introspective equilibrium when the defection payoff is D *but not when it is* D' *(i.e.,* $\rho(\delta, C, S, T, D) > \overline{\rho}$ *and* $\rho(\delta, C, S, T, D') < \overline{\rho}$*). Then the value of the game when the defection payoff is* D' *is strictly higher than when the defection payoff is* D *whenever* $D - D'$ *is sufficiently small or sufficiently large.*

Proof. It will be convenient to combine the proofs of (a) and (b) by considering the game forms $\boldsymbol{u} = (\delta, C, S, T, D)$ and $\boldsymbol{u}' = (\delta', C, S, T, D')$. Then, with some abuse of notation, we can take $\delta' > \delta$ and $D' = D$ for proving (a), and $\delta' = \delta$ and $D' < D$ for proving (b). Write ρ and ρ' for the dominance parameters associated with $\boldsymbol{u}$ and $\boldsymbol{u}'$, respectively. As before, if the change in payoff parameters is sufficiently large (i.e., $\rho' < \underline{\rho}$), the result follows from the fact that $u_{11} > u_{22}$. So suppose $\rho' \in (\underline{\rho}, \overline{\rho})$. Then, we can again apply Lemma 1. In this case, (3) gives

$$\Delta > (p'_{11} + p'_{12})(1-\delta')(T-D') - (D-D').$$

So, if $\delta' > \delta$ and $D' = D$, we have $\Delta > 0$, proving (a); and if $\delta' = \delta$ and $D' < D$, then $\Delta > 0$ provided D is sufficiently close to D', proving (b). □

Appendix E Omitted proofs

E.1 Proof of Proposition 1

(a) We start by showing that every introspective equilibrium is a correlated equilibrium. We prove the result for general finite games and also do not require Assumptions (SYM)–(REG).

Let $\mathcal{G} = \langle N, \{S_j\}_{j\in N}, \{u_j\}_{j\in N}\rangle$ be a finite game, where N is the (finite) player set and for each player $j \in N$, S_j is the (finite) set of actions and $u_j : S_j \times S_{-j} \to \mathbb{R}$ is the payoff function. Fix an introspective type space, that is, a set T_j of introspective types and an impulse function $\mathcal{I}_j$ for each player $j \in N$ as well as a common prior on the set $\prod_j T_j$ of type profiles. We require that for each player $j \in N$, the type set T_j is a closed subset of the real line and that the impulse function $\mathcal{I}_j$ is measurable with respect to the σ-algebra $\mathscr{B}(T_j)$ on T_j induced by the Borel σ-algebra on $\mathbb{R}$. For each player $j \in N$, let Σ_j be the set of (pure) *strategies*, i.e., measurable functions $\sigma_j : T_j \to S_j$. For simplicity, we write $\sigma_{-j}(t_{-j})$ for $(\sigma_i(t_i))_{i\neq j}$. It will also be convenient to represent the common prior by its cumulative distribution function $F(\cdot)$.

The first step is to show that the level-k strategies are, in fact, strategies:

Lemma 2. *Let $j \in N$. Then, for every k, σ_j^k is measurable.*

Proof. For $k = 0$, the result follows from the assumption that the impulse functions are measurable. We prove the result for $k > 0$ by showing the following claim: for every player $j \in N$, tie-breaking rule ψ_j, and profile $\sigma_{-j} \in \Sigma_{-j}$ for the other player, the tie-breaking rule yields a strategy $\sigma_j \in \Sigma_j$ such that for every $t_j \in T_j$, $\sigma_j(t_j)$ is a best response to σ_{-j}. Given that the level-0 strategies are measurable for all players, it then follows that for each player $j \in N$, σ_j^1 is measurable. Iterating this argument gives that σ_j^k is measurable for all $j \in N$ and $k = 0, 1, \ldots$

It remains to prove the claim. Fix a player $j \in N$ and a strategy profile $\sigma_{-j} \in \Sigma_{-j}$. Then, for $s_j \in S_j$, the function mapping introspective type $t_j \in T_j$ into its interim expected payoff

$$V_j(s_j, \sigma_{-j}; t_j) := \int_{T_{-j}} u_j(s_j, \sigma_{-j}(t_{-j}))\, \mathrm{d}F(t_{-j} \mid t_j)$$

is measurable (e.g., Aliprantis and Border, 2006, Thm. 15.13). Let $\varphi_j(\cdot, \sigma_{-j}) : T_j \twoheadrightarrow S_j$ be the best-response correspondence (given σ_{-j}), i.e., $\varphi_j(t_j, \sigma_{-j})$ is the set of actions that maximize the interim expected payoff $V_j(\cdot, \sigma_{-j}; t_j)$ for t_j. By the Measurable Maximum Theorem (e.g., Aliprantis and Border, 2006, Thm. 18.19), $\varphi_j(\cdot, \sigma_{-j})$ is measurable. That is, for every collection C_j of subsets of S_j,

$$\{t_j \in T_j : \varphi_j(t_j, \sigma_{-j}) \in C_j\} \in \mathscr{B}(T_j).$$

Since S_j is finite, it now follows immediately that for every subset $B_j \subset S_j$ of actions,

$$\{t_j \in T_j : \varphi_j(t_j, \sigma_{-j}) = B_j\} \in \mathscr{B}(T_j).$$

Fix a tie-breaking rule, i.e., a function ψ_j that maps each nonempty subset $B_j \subset S_j$ into an element s_j of B_j. Then, $\psi_j \circ \varphi_j(\cdot, \sigma_{-j}) : T_j \to S_j$ is measurable. This proves the claim.

Hence, for every player $j \in N$, tie-breaking rule ψ_j, and $k > 0$, σ_j^k, defined by $\sigma_j^k(t_j) = \psi_j(\varphi_j(t_j, \sigma_{-j}^{k-1}))$ for $t_j \in T_j$, is measurable. □

Because the (pointwise) limit of a sequence of measurable functions is measurable, we have that, for each player j, the limit $\lim_{k\to\infty}\sigma_j^k$ of the level-k strategies is measurable. Hence, if $\sigma = (\sigma_j)_{j\in N}$ is an introspective equilibrium, then for each player $j \in N$, σ_j is a strategy.

It remains to show that if $\sigma = (\sigma_j)_{j\in N}$ is an introspective equilibrium, then for each player $j \in N$ and $t_j \in T_j$,

$$\int u_j(\sigma_j(t_j), \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) \geq \int u_j(s_j, \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) \tag{11}$$

for $s_j \in S_j$. By Lemma 2, the integrals in (11) are well-defined. Fix $j \in N$ and $t_j \in T_j$. By a standard integration to the limit result,

$$\lim_{k\to\infty}\int u_j(\sigma_j^k(t_j), \sigma_{-j}^{k-1}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) = \int u_j(\sigma_j(t_j), \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j);$$

likewise, for every $s_j \in S_j$,

$$\lim_{k\to\infty}\int u_j(s_j, \sigma_{-j}^{k-1}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) = \int u_j(s_j, \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j).$$

(Again, the integrals are well-defined.) The result then follows from a standard continuity argument.

(b)–(c) Say that a strategy σ_j is a *switching strategy* with threshold $t^* \in T$ if introspective types $t \in T$ with $t < t^*$ choose s^2 (i.e., $\sigma_j(t) = s^2$), and introspective types $t \in T$ with $t > t^*$ choose s^1 (i.e., $\sigma_j(t) = s^1$). (The introspective type t^* may choose either action.) At level 0, types follow their impulse. By (MON-I) and (SYM), the level-0 strategy σ_j^0 for each player j is a switching strategy with (common) threshold τ^0. Suppose that, at level 1, type τ^0 has a strict best response to choose s^1, i.e.,

$$(1 - F(\tau^0 \mid \tau^0))u_{11} + F(\tau^0 \mid \tau^0)u_{12} > (1 - F(\tau^0 \mid \tau^0))u_{21} + F(\tau^0 \mid \tau^0)u_{22},$$

or, equivalently, $F(\tau^0 \mid \tau^0) < 1 - \rho$. Let τ^1 be the largest introspective type $\tau^1 \leq \tau^0$ such that $F(\tau^0 \mid \tau^1) \geq 1 - \rho$ if such a type exists; otherwise let $\tau^1 = 0$ (i.e., all types choose s^1). Then, the level-1 strategy is a switching strategy with threshold τ^1: By (MON-B) and (SYM), action s^1 is a strict best response for introspective types $t > \tau^1$ against the belief that the other player follows the level-0 strategy, and action s^2 is a strict best response for introspective types $t < \tau^1$. Moreover, $F(t \mid t) < 1 - \rho$ for all $t \in [\tau^1, \tau^0]$, as, by (REG), $F(\tau \mid t)$ is strictly increasing in τ.

For $k > 1$, suppose, inductively, that for each player, the level-$(k-1)$ strategy is a switching strategy with threshold τ^{k-1}, and that, furthermore, $F(t \mid t) < 1 - \rho$ for all $t \in [\tau^{k-1}, \tau^0]$. Define τ^k to be the largest introspective type $\tau^k \leq \tau^{k-1}$ such that $F(\tau^{k-1} \mid \tau^k) \geq 1 - \rho$ if such a type

exists, or set $\tau^k = 0$ otherwise. Then, by a similar argument as before, the level-k strategy is a switching strategy with threshold τ^k, and $F(t \mid t) < 1 - \rho$ for all $t \in [\tau^k, \tau^0]$.

The sequence $\tau^0, \tau^1, \ldots$ of level-k thresholds, being a monotone sequence in a compact space, converges to some equilibrium threshold $\tau \in T$. Moreover, this equilibrium threshold is the largest $\tau \leq \tau^0$ such that $F(\tau \mid \tau) \geq 1 - \rho$ if such a type exists, or $\tau = 0$ otherwise. A similar argument shows that if action s^2 is a strict best response to the switching strategy with threshold τ^0, the equilibrium threshold $\tau = \lim_{k \to \infty} \tau^k$ is the smallest $\tau \geq \tau^0$ such that $F(\tau \mid \tau) \leq 1 - \rho$ if such a type exists, and $\tau = 1$ otherwise. Finally, if type τ^0 is indifferent between s^1 and s^2, i.e., $F(\tau^0 \mid \tau^0) = 1 - \rho$, then, by (MON-B), $F(\tau^0 \mid t) < 1 - \rho$ for $t > \tau^0$ and $F(\tau^0 \mid t) > 1 - \rho$ for $t < \tau^0$. Hence, the equilibrium threshold τ is just τ^0. So, in introspective equilibrium, each player follows a switching strategy with threshold τ. The equilibrium is essentially unique: it pins down the behavior for all introspective types $t \neq \tau$, and this set has probability 1.

(d) Note that the dominance parameter decreases (resp. increases) when the payoffs to action s^1 (resp. action s^2) are increased (holding other payoff parameters fixed). Hence, it suffices to consider how the equilibrium threshold varies with the dominance parameter. Fix an introspective type space and dominance parameters $\rho, \tilde{\rho}$ such that $\tilde{\rho} > \rho$. Denote the games with dominance parameters ρ and $\tilde{\rho}$ by $\mathcal{G}$ and $\widetilde{\mathcal{G}}$, respectively, and let τ and $\tilde{\tau}$ be the respective equilibrium thresholds.

First suppose that s^1 is a strict best response for the level-0 threshold type in game $\mathcal{G}$, but not in $\widetilde{\mathcal{G}}$, that is, $1 - \tilde{\rho} \leq F(\tau^0 \mid \tau^0) < 1 - \rho$. Then, by the proof of part (b), it follows immediately that $\tau < \tau^0 \leq \tilde{\tau}$.

Next, suppose that s^1 is a best response for the level-0 threshold type in $\widetilde{\mathcal{G}}$ (and hence also in $\mathcal{G}$), that is, $F(\tau^0 \mid \tau^0) < 1 - \tilde{\rho} < 1 - \rho$. Then, by the proof of part (b), either $\tau = 0 \leq \tilde{\tau}$ or otherwise

$$\tau = \sup\{t \leq \tau^0 \colon F(t \mid t) \geq 1 - \rho\} \leq \sup\{t \leq \tau^0 \colon F(t \mid t) \geq 1 - \tilde{\rho}\} = \tilde{\tau}.$$

By a similar argument, if s^1 is not a best response for the level-0 threshold type in $\mathcal{G}$ (and hence s^2 is a best response in $\widetilde{\mathcal{G}}$), then the proof of part (b) shows that either $\tilde{\tau} = 1 \geq \tau$ or

$$\tilde{\tau} = \inf\{t \geq \tau^0 \colon F(t \mid t) \leq 1 - \rho\} \geq \inf\{t \leq \tau^0 \colon F(t \mid t) \leq 1 - \tilde{\rho}\} = \tau.$$

So in either case, $\tau \leq \tilde{\tau}$.

(e) To prove the necessity of (NMRB), suppose that for some $\rho^0 \in (0, 1)$, $F(t \mid t) \leq 1 - \rho^0$ for all $t < \tau^0$ and $F(t \mid t) \geq 1 - \rho^0$ for all $t > \tau^0$. Then, by the argument above, $\tau = 0$ for $\rho < \rho^0$ and $\tau = 1$ for $\rho > \rho^0$. The sufficiency of the condition follows from the proof of Theorem 1 below. □

E.2 Proof of Theorem 1

Fix a type space $\mathcal{T} = (F, \tau^0)$. Recall the definition of $\underline{\rho}$ and $\overline{\rho}$, and let

$$\underline{\tau} = \sup\{t \in [0, \tau^0] \colon F(t \mid t) \geq 1 - \underline{\rho}\};$$
$$\overline{\tau} = \inf\{t \in [\tau^0, 1] \colon F(t \mid t) \leq 1 - \overline{\rho}\};$$

be the introspective types "closest" to τ^0 whose rank beliefs attain the relevant values; see Figure 2 for an illustration. We start with a few auxiliary results:

Lemma 3. $\lim_{t \downarrow 0} F(t \mid t) \leq 1/2$ *and* $\lim_{t \uparrow 1} F(t \mid t) \geq 1/2$.

Proof. Suppose, by contradiction, that $\lim_{t \downarrow 0} F(t \mid t) > 1/2$. This implies that there exist $\alpha > 1/2$ and $\delta > 0$ such that $F(t \mid t) \geq \alpha$ for all $t \in (0, \delta)$. Because

$$F(t \mid t) = \frac{\int_0^t f(x,t)\,\mathrm{d}x}{\int_0^1 f(x,t)\,\mathrm{d}x},$$

we have that for $t \in (0, \delta)$,

$$\int_0^t f(x,t)\,dx \geq \alpha \int_0^1 f(x,t)\,dx$$

and therefore

$$\int_0^\delta \int_0^t f(x,t)\,\mathrm{d}x\,\mathrm{d}t \geq \alpha \int_0^\delta \int_0^1 f(x,t)\,\mathrm{d}x\,\mathrm{d}t \geq \alpha \int_0^\delta \int_0^\delta f(x,t)\,\mathrm{d}x\,\mathrm{d}t.$$

But by (SYM),

$$\int_0^\delta \int_0^\delta f(x,t)\,\mathrm{d}x\,\mathrm{d}t = 2 \int_0^\delta \int_0^t f(x,t)\,\mathrm{d}x\,\mathrm{d}t.$$

Hence, $\alpha \leq 1/2$, contradicting our assumptions. The proof that $\lim_{t \uparrow 1} F(t \mid t) \geq 1/2$ is similar and thus omitted. □

Lemma 4. *For any introspective type space,* $< 0 \underline{\rho} \leq \overline{\rho} < 1$*, where we have a strict inequality if and only if the type space induces non-monotone rank beliefs. Moreover, we cannot have both* $\underline{\tau} = 0$ *and* $\overline{\tau} = 1$*, i.e., at least one of these types must lie in the interior of* T*.*

Proof. We start with the first claim. It is immediate from the definitions that $\underline{\rho} \leq \overline{\rho}$ and that (NMRB) implies that $\underline{\rho} < \overline{\rho}$. To show the converse, consider a type space such that $\underline{\rho} < \overline{\rho}$. Then, there is $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$ or there is $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both). But this is just (NMRB). To conclude our proof of the first claim, it remains to show that $\underline{\rho} > 0$ and $\overline{\rho} < 1$. By (REG), $F(t \mid t)$ is continuous on $[0, 1]$ (and thus has a maximum on

$[0,\tau^0]$ and a minimum on $[\tau^0,1]$) and $F(t \mid t) \in (0,1)$ for $t \in (0,1)$. It thus remains to show that $\sup\{F(t \mid t) : t \in [0,\tau^0]\} < 1$ and $\inf\{F(t \mid t) : t \in [\tau^0,1]\} > 0$. The claim now follows from Lemma 3.

We next prove the second claim.By Lemma 3 and the definition $\underline{\tau} = \sup\{t \in [0,\tau^0] : F(t \mid t) \geq 1-\underline{\rho}\}$ it follows that $\underline{\tau} = 0$ implies $F(t \mid t) < \frac{1}{2}$ for all $t \in (0,\tau^0]$. Similarly, $\overline{\tau} = 1$ implies $F(t \mid t) > \frac{1}{2}$ for all $t \in [\tau^0,1)$. Since $F(t \mid t)$ is continuous on $[0,1]$, it follows that we cannot have both $\underline{\tau} = 0$ and $\overline{\tau} = 1$. □

We are now ready to prove Theorem 1. By the proof of Proposition 1(b)–(c), the equilibrium threshold is $\tau = 0$ for $\rho < \underline{\rho}$ and $\tau = 1$ for $\rho > \overline{\rho}$, proving (a) and (b).

To prove (c), we first show that the value is generically not equal to the expected payoff in one of the Nash equilibria. Recall that (NMRB) implies that $\underline{\rho} < \overline{\rho}$. Moreover, for $\rho \in (\underline{\rho},\overline{\rho})$, the equilibrium threshold τ lies strictly between $\underline{\tau}$ and $\overline{\tau}$. To see that behavior in introspective equilibrium is not consistent with Nash equilibrium if $\rho \in (\underline{\rho},\overline{\rho})$, first note that introspective equilibrium is not consistent with pure Nash equilibrium, as players choose both actions with strictly positive probability (by (REG)). To prove that behavior is not consistent with mixed Nash equilibrium, note that in any (strictly) mixed Nash equilibrium, the probability $p_{11}+p_{22}$ that players coordinate on one of the strict Nash equilibria equals $\rho^2+(1-\rho)^2 = 1-2\rho(1-\rho)$. Fix $\rho \in (\underline{\rho},\overline{\rho})$ and let τ be the corresponding equilibrium threshold. Denote by $p_{nm}(\tau)$ the probability that players play according to the action profile (s^n,s^m) in introspective equilibrium. Since $F(\tau \mid \tau) = 1-\rho$ (proof of Proposition 1), we have

$$p_{11}(\tau) = \int_\tau^1 (1-F(\tau \mid t))\,\mathrm{d}F(t) > \rho\,(1-F(\tau)) = \rho\,\big(p_{11}(\tau)+p_{12}(\tau)\big);$$
$$p_{22}(\tau) = \int_0^\tau F(\tau \mid t)\,\mathrm{d}F(t) > (1-\rho)\,F(\tau) = (1-\rho)\,\big(p_{22}(\tau)+p_{21}(\tau)\big);$$

where the inequalities follow from (MON-B) and (REG). Hence, $\rho\,p_{22}(\tau) > (1-\rho)p_{21}(\tau)$ and thus $\rho\,(p_{22}(\tau)+p_{21}(\tau)) > p_{21}(\tau)$. Likewise, $(1-\rho)p_{11}(\tau) > \rho\,p_{12}(\tau)$ and thus $(1-\rho)(p_{11}(\tau)+p_{12}(\tau)) > p_{12}(\tau)$. We thus have

$$\begin{aligned}1 =& p_{11}(\tau)+p_{12}(\tau)+p_{21}(\tau)+p_{22}(\tau)\\ >& \left(\frac{1}{1-\rho}+\frac{1}{\rho}\right)\frac{p_{12}(\tau)+p_{21}(\tau)}{2}\\ =& \frac{1}{2\rho\,(1-\rho)}\,\big(p_{12}(\tau)+p_{21}(\tau)\big),\end{aligned}$$

where we have used that $p_{12}(\tau) = p_{21}(\tau)$. Hence, $p_{12}(\tau)+p_{21}(\tau) < 2\rho\,(1-\rho)$, or, equivalently,

$$p_{11}(\tau)+p_{22}(\tau) > 1-2\rho(1-\rho). \tag{12}$$

We thus conclude that behavior in introspective equilibrium is not consistent with mixed Nash equilibrium. To show that the value in introspective equilibrium is not equal to the expected payoff in mixed Nash equilibria for generic payoff parameters (i.e., for a set of payoff parameters with Lebesgue measure 1), note that there exist $\alpha_\rho > 0$ (dependent on ρ) such that $p_{12}(\tau) = p_{21}(\tau) = \rho(1-\rho)(1-\alpha_\rho)$. Then, there is β_ρ (which could be positive or negative) such that $p_{11}(\tau) = \rho^2 + \rho(1-\rho)(\alpha_\rho - \beta_\rho)$ and $p_{22}(\tau) = (1-\rho)^2 + \rho(1-\rho)(\alpha_\rho + \beta_\rho)$. So, the difference between the expected payoff V_{MNE} in mixed Nash equilibrium and the value V is

$$V_{MNE} - V = \rho(1-\rho)\big(\alpha_\rho u_{12} + \alpha_\rho u_{21} - (\alpha_\rho - \beta_\rho)\, u_{11} - (\alpha_\rho + \beta_\rho)\, u_{22}\big).$$

Hence, we have that $V_{MNE} = V$ if and only if

$$\alpha_\rho(u_{11} - u_{21} + u_{22} - u_{12}) = \beta_\rho(u_{11} - u_{22}).$$

Fixing $u_{11} - u_{21}$ and $u_{22} - u_{12}$ fixes $u_{11} - u_{21} + u_{22} - u_{12}$ and ρ (and thus α_ρ and β_ρ), but does not pin down $u_{11} - u_{22}$. We thus conclude that the value in introspective equilibrium is equal to the expected payoff in mixed Nash equilibrium only for a set of payoff parameters of Lebesgue measure 0.

It remains to derive (1). Note that the value of a game can be expressed as

$$V = V(\boldsymbol{u}; \mathcal{T}) = p_{11}(\tau)\, u_{11} + p_{12}(\tau)\, u_{12} + p_{21}(\tau)\, u_{21} + p_{22}(\tau)\, u_{22},$$

where $p_{nm}(\tau)$ is the probability that the action profile (s^n, s^m) is played in introspective equilibrium. Since in introspective equilibrium, an introspective type t chooses s^1 if $t > \tau$ and chooses s^2 if $t < \tau$, we have

$$\begin{aligned} p_{22}(\tau) &= F(\tau, \tau), \\ p_{12}(\tau) &= p_{21}(\tau) = F(\tau) - F(\tau, \tau), \\ p_{11}(\tau) &= 1 - 2F(\tau) + F(\tau, \tau), \end{aligned}$$

where $F(t)$ is the (marginal) distribution function of a player's type. Hence, we can rewrite the value of the game as

$$\begin{aligned} V &= u_{11} + (u_{21} + u_{12} - 2u_{11})F(\tau) + (u_{11} + u_{22} - u_{21} - u_{12})F(\tau, \tau) \\ &= u_{11} + (u_{21} + u_{12} - 2u_{11})F(\tau) + \frac{u_{11} - u_{21}}{1 - \rho} F(\tau, \tau). \end{aligned}$$

This proves (1). □

Remark E.1. *The proof of Theorem 1 can be simplified if we strengthen (REG) to require that the density $f(\cdot)$ has full support on the whole of $T \times T$ (not just its interior): with this stronger*

assumption, it follows directly that $F(t \mid t)$ goes to 0 and to 1 as t approaches 0 and 1, respectively (which is obviously stronger than Lemma 3). A disadvantage of adopting a stronger version of (REG) is that it rules out some potentially interesting introspective type spaces such as versions of the social salience type space (Section 3.2.1) with $g(1) = 0$ or the animal spirits type space (Section 3.2.2).

E.3 Proof of Proposition 2

Part (p1) follows directly from Theorem 1 by noting that $\rho = 1/(w+1)$. To prove (p2), fix an introspective type space $\mathcal{T}$ with $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$. Since $1 - \rho = \frac{1}{2}$, players choose both actions with positive probability in introspective equilibrium (i.e., $\tau \in (0,1)$). The value is given by

$$V = p_{11}(\tau) + p_{22}(\tau) = 1 - 2F(\tau) + 2F(\tau, \tau).$$

As $\tau \in (0,1)$, by (REG), $F(\tau,\tau) < F(\tau)$ and thus $V < 1$. That the value is strictly greater than $\frac{1}{2}$ follows directly from the fact that $V = p_{11}(\tau) + p_{22}(\tau) > 1 - 2\rho(1-\rho) = \frac{1}{2}$ (by Eq. (12)). □

E.4 Proof of Proposition 3

Let $\mathcal{T}$ be an introspective type space with $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$. For $w > 0$ and $x \in [0, w-1)$, write $\rho(w,x)$ for the dominance parameter of $\widetilde{G}_x$. Then, there is $\underline{w}$ such that for $w > \underline{w}$, $\rho(w,0) < \underline{\rho}$ and thus, by Theorem 1, $V((w,-c,0,1);\mathcal{T}) = w$. Moreover, for any w, $\lim_{x \uparrow w-1} \rho(w,x) = (w+c)/(2w+c)$. So, as $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$, there is $\underline{w}'$ such that for $w > \underline{w}'$, $\lim_{x \uparrow w-1} \rho(w,x) \in (\underline{\rho}, \overline{\rho})$. The result then follows by choosing $w^* = \max\{\underline{w}, \underline{w}'\}$. □

E.5 Proof of Theorem 3 (cntd)

We start by calculating the per-period profits. First note that the inverse demand functions

$$p_i = 1 - q_i - r\, q_{-i} \qquad (0 < r < 1) \tag{13}$$

and

$$\tilde{p}_i = a - b\, \tilde{q}_i - c\, \tilde{q}_{-i} \qquad (a > 0, b > c > 0) \tag{14}$$

are equivalent in the sense that the transformations $\tilde{p}_i = a\, p_i$, $\tilde{q}_i = \frac{a}{b} q_i$, $r = \frac{c}{b}$ provide a one-to-one correspondence between the solutions of (13) and (14). So, it suffices to consider (13). Solving for q_i gives

$$q_i = \frac{1 - r - (p_i - r p_{-i})}{1 - r^2}.$$

Hence,

$$\pi_i = p_i q_i = \frac{1 - r - (p_i - r p_{-i})}{1 - r^2} p_i.$$

The Bertrand-Nash price maximizes a firm's profit given the other firm's price. Hence, $p^N = \frac{1-r}{2} + \frac{r}{2} p^N$, which yields $p^N = \frac{1-r}{2-r}$ and

$$\pi^N = \left(\frac{1-r}{2-r}\right)^2 \frac{1}{1-r^2}.$$

The collusive price p^* maximizes $\frac{(1-r)(1-p^*)}{1-r^2} p^*$. Hence, $p^* = \frac{1}{2}$, and the corresponding profit is

$$\pi^* = \frac{1}{4(1+r)}.$$

The other profit terms depend on whether the constraint $q_{-i} \geq 0$ is binding. The constraint is not binding if $r < \sqrt{3} - 1$, and binds if $r \geq \sqrt{3} - 1$. First suppose the constraint is not binding, i.e., $r < \sqrt{3} - 1$. The cheating price p^c maximizes $(1 - \frac{1}{2} r - p^c) p^c$. Hence, $p^c = \frac{1}{2} - \frac{r}{4}$ and the corresponding profit is

$$\pi^c = \frac{1 - r + \frac{r^2}{4}}{4(1 - r^2)}.$$

It follows that $q^v = \frac{1 - r - r^2/2}{2(1-r^2)}$. Hence,

$$\pi^v = \frac{1 - r - \frac{r^2}{2}}{4(1 - r^2)}.$$

Finally,

$$\pi^m = \frac{(1-r)(1-p^c)}{1-r^2} p^c = \frac{(1-r)\left(1 - \frac{r^2}{4}\right)}{4(1-r^2)}.$$

It will be convenient to define $\psi := 4(1 - r^2)\pi$. Thus,

$$\psi^c = 1 - r + \tfrac{r^2}{4}; \qquad \psi^* = 1 - r;$$
$$\psi^m = (1-r)\left(1 - \tfrac{r^2}{4}\right); \qquad \psi^v = 1 - r - \tfrac{r^2}{2};$$
$$\psi^N = \frac{(1-r)^2}{1 - r + \frac{r^2}{4}}.$$

Clearly, $\psi^c > \psi^* > \psi^m$. Moreover,

$$\frac{\psi^m}{\psi^N} = \frac{1 - r + \frac{r^3}{4}\left(1 - \frac{r}{4}\right)}{1 - r} > 1;$$
$$\frac{\psi^N}{\psi^v} = \frac{(1-r)^2}{(1-r)^2 - \frac{r^2}{4}\left(1 - r - \frac{r^2}{2}\right)} > 1.$$

Hence, $\psi^c > \psi^* > \psi^m > \psi^N > \psi^v$ and thus $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$.

Next suppose that the constraint binds, i.e., $r \geq \sqrt{3} - 1$. We thus have $q^v = 0$. This yields $p^c = 1 - \frac{1}{2r}$ and thus

$$\pi^c = \frac{(2 - \frac{1}{r})(\frac{1}{r} - r)}{4(1 - r^2)}; \qquad \pi^m = \frac{(1 - r)(2 - \frac{1}{r})\frac{1}{r}}{4(1 - r^2)}; \qquad \pi^v = 0.$$

Using the notation $\psi := 4(1 - r^2)\pi$ again, we have

$$\psi^c = (2 - \tfrac{1}{r})(1 + \tfrac{1}{r})(1 - r); \qquad \psi^m = \tfrac{1}{r}(2 - \tfrac{1}{r})(1 - r); \qquad \psi^v = 0.$$

Let $x := \frac{1}{r}$, so $1 < x \leq \frac{1}{2}(\sqrt{3} + 1)$. Then, $(2 - x)(1 + x) \geq (2 - \frac{1}{2}(\sqrt{3} + 1))(1 + \frac{1}{2}(\sqrt{3} + 1)) = \frac{3}{2}$. Hence, $\psi^c > \psi^* = 1 - r$. Also, $x(2 - x) < 1$, and thus $\psi^* > \psi^m$. Moreover, $\psi^m > \psi^N$ if and only if $(2 - \frac{1}{r})(\frac{1}{r} - \frac{1}{2}) > \frac{1-r}{1-r/2}$, which is equivalent to $(2 - x)(x - \frac{1}{2}) > \frac{x-1}{x-1/2} = 1 - \frac{1}{2x-1}$. Since $(2 - x)(x - \frac{1}{2}) \geq (2 - 1)(1 - \frac{1}{2}) = \frac{1}{2}$ while $1 - \frac{1}{2x-1} \leq 1 - \frac{1}{\sqrt{3}} < \frac{1}{2}$, we have $\psi^m > \psi^N$. It follows that $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$.

We next derive the payoffs for the repeated game. If firm i chooses strategy $\sigma_i \in \{\sigma^*, \sigma^c\}$ and the other firm chooses strategy $\sigma_{-i} \in \{\sigma^*, \sigma^c\}$, then, if the common discount factor is $\delta \in (0, 1)$, the (normalized) expected discounted sum of profits for firm i is

$$(1 - \delta) \sum_{\tilde{t}=0}^{\infty} \delta^{\tilde{t}} \, \mathbb{E}_{(\sigma_i, \sigma_{-i})}\big[\pi_i^{\tilde{t}}\big],$$

where $\pi_i^{\tilde{t}}$ is firm i's profit in period $\tilde{t}$ and $\mathbb{E}_{(\sigma_i, \sigma_{-i})}[\cdot]$ is the expectation operator induced by the strategy profile (σ_i, σ_{-i}). This yields the payoff matrix in the main text. By the assumption that $\delta > (\pi^c - \pi^*)/(\pi^c - \pi^N))$ and given that $\pi^m > \pi^v$, this is a coordination game, i.e., $\rho \in (0, 1)$. □

E.6 Proof of Proposition 4

Fix a game form $\boldsymbol{u}$ (with corresponding dominance parameter ρ) and fix the introspective type space $\mathcal{T}_0 = (F, \tau_0^0)$ at time 0. We assume that $F(\tau_0^0 \mid \tau_0^0) \neq 1 - \rho$ and that the rank belief function does not attain a local extremum at the equilibrium threshold τ^0 at time 0, i.e., $F(\tau^0 \mid \tau^0)$ is not a local maximum or minimum. Since there are at most countably many values for ρ such that $F(\tau_0^0 \mid \tau_0^0) = 1 - \rho$ or that $1 - \rho$ is the value that $F(t \mid t)$ attains at a local extremum, proving the claim for this case establishes that the claim holds for generic $\boldsymbol{u}$.

Fix a dynamic $\{\tau_{\tilde{t}}^0\}_{\tilde{t}}$ satisfying (6). Notice that this is a deterministic process (as there is a continuum of players). For $\tilde{t} \geq 0$, let $\tau_{\tilde{t}}$ be the introspective equilibrium for $\mathcal{G}_t = (\boldsymbol{u}, \mathcal{T}_t)$, where $\mathcal{T}_t = (F, \tau_{\tilde{t}}^0)$. By the triangle inequality, it suffices to prove the following: For every $\chi > 0$, there is $\bar{\varepsilon}_\chi > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon}_\chi)$ and for every period $\tilde{t}$, we have $|\tau_{\tilde{t}} - \tau_0| < \chi$. That is, for

$\varepsilon \in (0, \bar{\varepsilon}_\chi)$, the introspective equilibrium in period $\tilde{t}$ is within χ the introspective equilibrium at time 0 (uniformly in $\tilde{t}$). In fact, we show a stronger result: There is $\bar{\varepsilon} > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon})$ and for every period $\tilde{t}$, we have $\tau_{\tilde{t}} = \tau_0$. Thus, if the noise is sufficiently small, players play according to the same introspective equilibrium in every period.

To show this, note that, by assumption, $F(\tau_0^0 \mid \tau_0^0) \neq 1 - \rho$. So, $\tau_0 \neq \tau_0^0$. Also, note that by (REG), the rank belief function $F(t \mid t)$ is continuous in t. We prove the result for the case that $\tau_0^0 > \tau_0$; the proof for the case that $\tau_0^0 < \tau_0$ is similar and thus omitted. By the proof of Proposition 6 (showing that introspective equilibrium satisfies (ATTR)), there is $\bar{\varepsilon}' > 0$ such that for any $\varepsilon \in (0, \bar{\varepsilon}^1)$, for any $\tau_{\tilde{t}}^0 \in (\tau_0 - \varepsilon, \tau_0^0] \cap T$, the introspective process $\{\tau_{\tilde{t}}^k\}_k$ converges to τ_0. The proof then follows if we show that there is $\bar{\varepsilon}'' > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}'')$, for any $\tau_{\tilde{t}}^0 \in [\tau_0, \tau_0^0 + \varepsilon) \cap T$, the introspective process $\{\tau_{\tilde{t}}^k\}_k$ converges to τ_0 (and set $\bar{\varepsilon} = \min\{\bar{\varepsilon}', \bar{\varepsilon}''\}$). But this follows from the continuity of $F(t \mid t)$ and the fact that $F(\tau_0^0 \mid \tau_0^0) < 1 - \rho$. □

E.7 Proof of Proposition 5

By Proposition 1, introspective equilibrium is a symmetric correlated equilibrium, i.e., $\mu_{12} = \mu_{21}$. By Proposition 1 in Calvó-Armengol (2006), the set $CE(\rho)$ of correlated equilibria of a coordination game with dominance parameter ρ is a polytope in $\mathbb{R}^4$ with 5 vertices (all distinct) given by

$$\begin{aligned}
\mu^1 &= (1, 0, 0, 0);\\
\mu^2 &= (0, 1, 0, 0);\\
\mu^3 &= \big(\rho^2, (1-\rho)^2, \rho(1-\rho), \rho(1-\rho)\big);\\
\mu^4 &= \left(\frac{\rho^2}{1-\rho(1-\rho)}, \frac{(1-\rho)^2}{1-\rho(1-\rho)}, \frac{\rho(1-\rho)}{1-\rho(1-\rho)}, 0\right);\\
\mu^5 &= \left(\frac{\rho^2}{1-\rho(1-\rho)}, \frac{(1-\rho)^2}{1-\rho(1-\rho)}, 0, \frac{\rho(1-\rho)}{1-\rho(1-\rho)}\right);
\end{aligned}$$

where μ^1 and μ^2 are the action distributions corresponding to the pure Nash equilibria, and μ^3 corresponds to the mixed Nash equilibrium. Hence, the set of correlated equilibria $CE(\rho)$ is a full-dimensional subset of the three-dimensional simplex. The set of symmetric correlated equilibria is a two-dimensional subset of $CE(\rho)$. Hence, the set of symmetric correlated equilibria has Lebesgue measure 0 in the set of correlated equilibria. Moreover, introspective equilibrium is a strict correlated, except perhaps when one of the following holds: (1) $F(\tau^0 \mid \tau^0) = 1 - \rho$; or (2) $F(\tau \mid \tau)$ is a local extremum. By the proof of Proposition 6, these cases are non-generic. We can thus conclude that the set of correlated equilibria induced by an introspective equilibrium is a subset of a set that has Lebesgue measure 0 in the set of all correlated equilibria, and generically a strict subset thereof. Hence, it has Lebesgue measure 0. □

E.8 Proof of Proposition 6

We show that introspective equilibrium satisfies (ATTR) and (LYAP) except when $F(\tau^0 \mid \tau^0) = 1-\rho$ or $F(\tau \mid \tau)$ is a local extremum. Since there are at most countably many values for ρ such that $F(\tau^0 \mid \tau^0) = 1-\rho$ or that $1-\rho$ is the value that $F(t \mid t)$ attains at a local extremum, this establishes that introspective equilibrium is asymptotically stable for generic $\boldsymbol{u}$. So suppose that $F(\tau^0 \mid \tau^0) \neq 1-\rho$ and that $F(\tau \mid \tau)$ is not a local maximum or minimum. We start with (ATTR). First suppose $\tau = 0$; the proof for $\tau = 1$ is similar and thus omitted. Since $F(\tau \mid \tau) \neq 1-\rho$, we have $\tau \neq \tau^0$ and thus $\tau^0 > 0$. Then, for any level-0 threshold $\tilde{\tau}^0 \in B_{\tau^0}(\tau)$, the introspective process converges to τ (by the proof of Proposition 1); so, the claim holds with $\varepsilon = \tau^0$. Next suppose $\tau \in (0,1)$. We will construct a nonempty open interval (τ_{min}, τ_{max}) such that, starting from any level-0 threshold $\tilde{\tau}^0 \in (\tau_{min}, \tau_{max})$, the introspective process converges to τ; this shows that the claim holds with $\varepsilon := \tau_{max} - \tau_{min} > 0$. We prove the result for the case where $F(\tau^0 \mid \tau^0) < 1-\rho$; the proof for the result when $F(\tau^0 \mid \tau^0) > 1-\rho$ is similar and thus omitted. Again, $\tau^0 > \tau$, and the level-k thresholds $\{\tau^k\}$ form a decreasing sequence that converges to τ; so, we can set $\tau_{max} = \tau^0$, where we note that $\tau_{max} > \tau$. We next construct $\tau_{min} < \tau$. By the proof of Proposition 1, there is $t \in [0, \tau]$ such that $F(t \mid t)$ is a local maximum. Moreover, since $F(\tau \mid \tau)$ is not a local extremum,

$$\tau^* := \sup\{t \in [0,\tau] : F(t \mid t) \text{ is a local maximum}\}$$

is strictly smaller than τ. By the proof of Proposition 1, $F(\tau^* \mid \tau^*) > 1-\rho$, and, by a similar argument as before, any introspective process starting at $\tilde{\tau}^0 \in (\tau^*, \tau]$ converges to τ. Hence, we can set $\tau_{min} = \tau^*$.

The proof that introspective equilibrium satisfies (LYAP) now follows immediately. Fix $\eta > 0$, and let $\varepsilon > 0$ be as constructed in the proof of property (ATTR). By the argument above, for any $\tilde{\tau}^0 \in B_\varepsilon(\tau)$, the introspective process $\{\tilde{\tau}^k\}_k$ starting at $\tilde{\tau}^0$ has the property that $|\tau - \tilde{\tau}^k|$ decreases with k. Hence, the claim holds for $\delta = \min\{\eta, \varepsilon\}$. □

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.

Alaoui, L., K. A. Janezic, and A. Penta (2020). Reasoning about others' reasoning. *Journal of Economic Theory 189*, 105091.

Alaoui, L. and A. Penta (2016). Endogenous depth of reasoning. *Review of Economic Studies 83*(4), 1297–1333.

Albæk, S., P. Møllgaard, and P. B. Overgaard (1997). Government-assisted oligopoly coordination? A concrete case. *Journal of Industrial Economics 45*(4), 429–443.

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

Angeletos, G.-M. and J. La'O (2013). Sentiments. *Econometrica 81*(2), 739–779.

Angeletos, G.-M. and A. Pavan (2004). Transparency of information and coordination in economies with investment complementarities. *American Economic Review 94*(2), 91–98.

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

Arifovic, J. and J. Jiang (2019). Strategic uncertainty and the power of extrinsic signals: Evidence from an experimental study of bank runs. *Journal of Economic Behavior and Organization 167*, 1–17.

Athey, S. (2002). Monotone comparative statics under uncertainty. *Quarterly Journal of Economics 117*(1), 187–223.

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

Bacharach, M. (1993). Variable universe games. In K. Binmore, A. Kirman, and P. Tani (Eds.), *Frontiers of Game Theory*. MIT Press.

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

Battigalli, P., A. Di Tillio, E. Grillo, and A. Penta (2011). Interactive epistemology and solution concepts for games with asymmetric information. *The BE Journal of Theoretical Economics 11*(1).

Battigalli, P. and M. Siniscalchi (2003). Rationalization and incomplete information. *The BE Journal of Theoretical Economics 3*(1).

Bergin, J. and B. L. Lipman (1996). Evolution with state-dependent mutations. *Econometrica*, 943–956.

Bicchieri, C. and E. Dimant (2019). Nudging with care: The risks and benefits of social information. *Public choice*, 1–22.

Bicchieri, C. and H. Mercier (2014). Norms and beliefs: How change occurs. In M. Xenitidou and B. Edmonds (Eds.), *The Complexity of Social Norms*, pp. 37–54. Springer.

Binmore, K. and L. Samuelson (1997). Muddling through: Noisy equilibrium selection. *Journal of Economic Theory 74*, 235–265.

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

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

Calvó-Armengol, A. (2006). The set of correlated equilibria of (2×2) games. Working paper.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics 119*(3), 861–898.

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

Carlton, D. W., R. H. Gertner, and A. M. Rosenfield (1997). Communication among competitors: Game theory and antitrust. *George Mason Law Review 5*, 423–440.

Cass, D. and K. Shell (1983). Do sunspots matter? *Journal of Political Economy 91*, 193–228.

Cooper, R. and A. John (1988). Coordinating coordination failures in Keynesian models. *Quarterly Journal of Economics 103*(3), 441–463.

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

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

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

Crawford, V. P., U. Gneezy, and Y. Rottenstreich (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Economic Review 98*, 1443–1458.

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

Crawford, V. P. and D. E. Smallwood (1984). Comparative statics of mixed-strategy equilibria in non-cooperative two-person games. *Theory and Decision 16*, 225–232.

Cripps, M. (1991). Correlated equilibria and evolutionary stability. *Journal of Economic Theory 55*, 428–434.

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

Diamond, P. (1982). Aggregate demand equilibrium in search equilibrium. *Journal of Political Economy 90*, 881–894.

Dixit, A. K. (2004). *Lawlessness and Economics: Alternative Modes of Governance*. Princeton University Press.

Duffy, J. and E. O. Fisher (2005). Sunspots in the laboratory. *American Economic Review 95*, 510–529.

Echenique, F. (2002). Comparative statics by adaptive dynamics and the correspondence principle. *Econometrica 70*, 833–844.

Echenique, F. and A. S. Edlin (2004). Mixed equilibria are unstable in games of strategic complements. *Journal of Economic Theory 118*, 61–79.

Ellison, G. and D. Fudenberg (2000). Learning purified mixed equilibria. *Journal of Economic Theory 90*(1), 84–115.

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

Fehr, D., F. Heinemann, and A. Llorente-Saguer (2019). The power of sunspots: An experimental analysis. *Journal of Monetary Economics 103*, 123–136.

Foster, D. P. and R. V. Vohra (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior 21*, 40–55.

Fudenberg, D. and D. K. Levine (1999). Conditional universal consistency. *Games and Economic Behavior 29*, 104–130.

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

Grimmett, G. and D. Stirzaker (2020). *Probability and random processes*. Oxford university press.

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

Hart, S. (2005). Adaptive heuristics. *Econometrica 73*, 1401–1430.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*(5), 1127–1150.

Ivaldi, M., B. Jullien, P. Rey, P. Seabright, and J. Tirole (2003). The economics of tacit collusion. *Report for DG Competition, European Commission*.

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

Kets, W. (2011). Robustness of equilibria in anonymous local games. *Journal of Economic Theory 146*, 300–325.

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

Kets, W. and A. Sandroni (2021). A theory of strategic uncertainty and cultural diversity. *Review of Economic Studies 88*, 287–333.

Knittel, C. R. and V. Stango (2004). Price ceilings as focal points for tacit collusion: Evidence from credit cards. *American Economic Review 93*, 1703–1729.

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.

Kühn, K.-U. (2001). Fighting collusion by regulating communication between firms. *Economic Policy 16*(32), 168–204.

Lenzo, J. and T. Sarver (2006). Correlated equilibrium in evolutionary models with subpopulations. *Games and Economic Behavior 56*(2), 271–284.

Lindbeck, A., S. Nyberg, and J. W. Weibull (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics 114*, 1–35.

Mailath, G. J., L. Samuelson, and A. Shaked (1997). Correlated equilibria and local interactions. *Economic Theory 9*(3), 551–556.

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.

Metzger, L. P. (2018). Evolution and correlated equilibrium. *Journal of Evolutionary Economics 28*(2), 333–346.

Milgrom, P. and J. Roberts (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica 58*, 1255–1277.

Milgrom, P. and C. Shannon (1994). Monotone comparative statics. *Econometrica*, 157–180.

Morris, S. (1997). Interaction games: A unified analysis of incomplete information, local interaction, and random matching. Working paper.

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

Morris, S., R. Rob, and H. Shin (1995). p-dominance and belief potential. *Econometrica 63*, 145–157.

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

Morris, S., H. S. Shin, and M. Yildiz (2016). Common belief foundations of global games. *Journal of Economic Theory 163*, 826–848.

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

Motta, M. (2004). *Competition Policy: Theory and Practice*. Cambridge University Press.

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

Penta, A. and P. Zuazo-Garin (2021). Rationalizability, observability and common knowledge. *Review of Economic Studies*. Forthcoming.

Ray, D. (2004). What's new in development economics? In M. Szenberg and L. Ramrattan (Eds.), *New Frontiers in Economics*, Chapter 10. Cambridge University Press.

Robson, A. J. and F. Vega-Redondo (1995). Efficient equilibrium selection in evolutionary games with random matching. *Journal of Economic Theory 70*, 65–92.

Ross, T. W. (1992). Cartel stability and product differentiation. *International Journal of Industrial Organization 10*(1), 1–13.

Salop, S. and J. Stiglitz (1977). Bargains and ripoffs: A model of monopolistically competitive price dispersion. *Review of Economic Studies 44*(3), 493–510.

Samuelson, L. (2002). Evolution and game theory. *Journal of Economic Perspectives 16*(2), 47–66.

Sandholm, W. H. (2007). Evolution in Bayesian games II: Stability of purified equilibria. *Journal of Economic Theory 136*(1), 641–667.

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

Scherer, F. M. (1980). *Industrial market structure and economic performance*.

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.

Selten, R. (1995). An axiomatic theory of a risk dominance measure for bipolar games with linear incentives. *Games and Economic Behavior 8*(1), 213–263.

Spagnolo, G. (2003). Divide et impera: Optimal deterrence mechanisms against cartels and organized crime. Working paper.

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.

Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy 72*(1), 44–61.

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

Sugden, R. (1995). A theory of focal points. *Economic Journal 105*, 533–550.

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.

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

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

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

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