# The Value of a Coordination Game

Willemien Kets* Wouter Kager† Alvaro Sandroni‡

January 26, 2022

**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.

*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. 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), 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.

This paper takes a first step in this research program by providing comparative statics on the value of coordination games (a subclass of supermodular games). We start from the observation that changing payoffs affects the value both directly (the value changes even if behavior does not) and indirectly (the payoff changes affect the way the game is played). Because these two effects can operate in opposite directions, the value need not be monotone in payoffs even when behavior changes in a monotone way. Moreover, the indirect strategic effects can be subtle: A change in payoffs that makes it less likely that players coordinate on a Pareto-dominated pure

[1] A notable exception is Crawford and Smallwood (1984), who 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.

Nash equilibrium (i.e., reduces *coordination failure*) may simultaneously make it more more likely that players fail to coordinate on one of the pure Nash equilibria (i.e., increase *miscoordination*). The net strategic effect then depends on the relative costs of miscoordination and coordination failure.

Motivated by this, we develop a novel theory of the value for symmetric $(2 \times 2)$ coordination games based on introspective equilibrium, a behavioral solution concept we developed in our earlier work (Kets and Sandroni, 2019, 2021). Unlike pure Nash equilibrium and its refinements, introspective equilibrium allows for miscoordination; and, unlike mixed Nash equilibrium, it has intuitive comparative statics on behavior (Proposition 2.1). This means that policies intended to incentivize particular behavior will have the intended effect on behavior; however, because the value depends on both direct and indirect effects, these policies may still make players worse off. Another important feature of introspective equilibrium is that it allows for non-economic factors to influence behavior. This is important because, in coordination games, the payoff structure may not fully pin down behavior. As a result, there is room for non-economic factors (e.g., salience of action labels) to shape behavior. Because this may affect the scope for miscoordination, this feature will be important for properly trading off the cost of miscoordination and coordination failure.

As a preliminary result, we completely characterize the value for any combination of payoff parameters when a researcher understands how non-economic factors impact behavior (Proposition 3.1). To be more precise, we model the effects of non-economic factors (e.g., salience) using type spaces (see Section 2.2).[2] If a researcher knows the type space, the characterization in Proposition 3.1 allows him to assess whether a policy that changes behavior in the desired direction also has positive welfare effects (i.e., increases the value). However, while this result is useful for settings where non-economic factors are easy to measure or to manipulate (as in, e.g., lab experiments), it is of limited use when the type space is not known, as is the case in many real-life settings. Our main results therefore derive testable comparative statics on the value, i.e., comparative statics that hold across type spaces.

We start by deriving testable predictions on the costs of miscoordination (Section 3.2). In the environments we consider there, the strategic effect is always negative. This is because the payoff changes we consider increase miscoordination without any offsetting positive strategic effects such as reducing coordination failure. We show that the negative strategic effect may dominate any positive direct effect. As a result, the value may fall even when all payoffs increase. We then move on to the key question of when miscoordination is more costly than coordination failure, i.e., under which conditions the negative strategic effect of increasing miscoordination

[2]In our model, non-economic factors are not directly payoff-relevant; they merely affect the likelihood that players are inclined to choose an action (e.g., because its label is salient). We therefore refer to these type spaces as *introspective* type spaces; see Section 2 for details.

dominates the positive effect of eliminating coordination failure (Section 3.3). We address this question in the context of two economic applications.

The first application studies when policies designed to stimulate investment improve welfare. Consider a setting where players can choose whether to invest or not, with both full investment and no investment being Nash equilibria, and where the Nash equilibrium with full investment is Pareto optimal. We consider the welfare effects of introducing an investment subsidy: Players who invest receive a subsidy regardless of whether the other player invests. This policy has only positive direct effects (as it (weakly) increases the payoffs to any action profile) and changes behavior in the desired direction (i.e., it (weakly) increases the probability of investment). Nevertheless, it can have a negative impact on welfare, as we show. Suppose that there is no investment in the absence of a subsidy, i.e., there is coordination failure. Then, if the subsidy stimulates investment but is not sufficient to induce full investment, it eliminates coordination failure but leads to miscoordination. We show that the subsidy has a negative welfare impact precisely when the payoffs under miscoordination are low and the subsidy is not very high (Theorem 3.4). Intuitively, when miscoordination payoffs are low, miscoordination is costly; and if the subsidy is not very high, then the positive strategic effect of eliminating coordination failure cannot compensate for this negative effect. Thus, a policy that has only positive direct effects and changes behavior in the desired direction can have negative welfare consequences when negative strategic effects dominate the positive ones. 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).

The second application considers the key question of whether firms benefit when collusion becomes easier to sustain. Some have 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 instead (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, firms are generally better off under miscoordination than if there is no collusion at all (Theorem 3.5). This suggests that industry associations have an incentive to lobby for changes that make collusion more attractive, such as improving the ease of detection or increasing the frequency of interaction (Ivaldi et al., 2003).

Because our main results provide comparative statics that hold across type spaces, our theory provides testable hypotheses on how the value changes with economic factors even when the non-economic factors that influence behavior are not known.[3] Our results demonstrate that the

[3]By contrast, much of the behavioral game theory literature estimates the relevant behavioral parameters from data (Eyster and Rabin, 2005; McKelvey and Palfrey, 1995; Nagel, 1995), with the notable exception of

value can change non-monotonically with payoffs even though behavior is monotone in payoffs. 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.

As noted above, we build on our earlier work (Kets and Sandroni, 2019, 2021) to develop this novel theory of the value. 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. As a result, none of our results on the value are suggested by or follow from our earlier work. 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 (see Appendix B.1). The current axiomatic approach not only makes the results stronger (since the results hold across all type spaces that satisfy our axioms), it also elucidates the main drivers behind the results (e.g., Lemma 2.2).

The outline of this paper is as follows. Section 2 introduces the model, and Section 3 presents the results. Section 4 presents a dynamic application. Section 5 discusses the model and connects our results to the related literature. Section 6 concludes. Proofs can be found in the appendix.

# 2 Model

## 2.1 Coordination

We consider symmetric $(2 \times 2)$ 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 $u_{11} \geq u_{22}$. All payoff parameters are common knowledge. Thus, both $(s^1, s^1)$ and $(s^2, s^2)$ are strict Nash equilibria and the equilibrium in which both players

Alaoui and Penta (2016) and Alaoui et al. (2020) who provide testable comparative statics for level-$k$ models and test their predictions experimentally.

choose $s^1$ is (weakly) Pareto dominant. If $u_{11} > u_{22}$ and players coordinate on $s^2$, there is *coordination failure*: While players play according to a (pure) Nash equilibrium, both would be better off if they could switch to the other pure Nash equilibrium. There is *miscoordination* if players choose the action profile $(s^1, s^2)$ or $(s^2, s^1)$ with positive probability.

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 the player assigns probability at least $\rho$ to the other player choosing $s^1$. That is, $(s^1, s^1)$ is $\rho$-dominant in the sense of Morris et al. (1995). We therefore refer to $\rho$ as the *dominance parameter*. This concept 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 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; but 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 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. Impulse functions are measurable, and players know their own impulse but not the other player's impulse. A player's first instinct is to follow his impulse. This defines his *level-0 strategy* $\sigma_j^0$ (i.e., $\sigma_j^0(t_j) = \mathcal{I}_j(t_j)$). For any $k > 0$, the *level-$k$ strategy* $\sigma_j^k$ for each player $j$ is a best response to the level-$(k-1)$ strategy $\sigma_{-j}^{k-1}$ of the other player.[4] A player's behavior is given by the limit $\sigma_j := \lim_{k\to\infty} \sigma_j^k$ of

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

the introspective process. If these limiting strategies exist, then $\sigma = (\sigma_j)_j$ is an *introspective equilibrium*.[5]

Introspective equilibrium depends in part on the impulse distribution (i.e., the distribution induced by the common prior $F(t_1, t_2)$ and the impulse functions $\mathcal{I}_j(t_j)$). In many settings of interest, players' impulses will be influenced by the social context (e.g., social cues, salient action labels). 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 can help players reach consistent expectations for the given economic environment (i.e., payoff parameters). 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.

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:

**Assumption 1 (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., $F(y, z) = F(z, y)$).*

**Assumption 2 (MON-I).** *Impulses are monotone in type: There is a threshold $\tau^0 \in (0, 1)$ such that for each player $j$, type $t$ has an impulse to choose $s^1$ (i.e., $\mathcal{I}_j(t) = s^1$) if and only if $t \geq \tau^0$.*

**Assumption 3 (MON-B).** *Beliefs are monotone in type: For every $\tau \in (0, 1)$, the conditional probability $F(\tau \mid t)$ that the other player has a type at most $\tau$ (given the player's type $t$) is strictly decreasing in $t$.*

**Assumption 4 (REG).** *The common prior $F(\cdot, \cdot)$ has a density $f$ that is continuous on $T \times T$ 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

[5] 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, but is not critical.

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 $s$ think it is likely that the other player has an impulse to choose $s$. This fits with the idea that the social context shapes beliefs. 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. Under these assumptions, we can summarize an introspective type space by its distribution function $F$ and the threshold $\tau^0$. We thus write $\mathcal{T} = (F, \tau^0)$ for an introspective type space.

Introspective equilibrium thus depends on both economic factors (the payoff parameters $\boldsymbol{u} := (u_{11}, u_{12}, u_{21}, u_{22})$) and non-economic factors (the type space). We will therefore refer to $\boldsymbol{u}$ 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* (though we use the term "coordination game" for both the game and the game form when no confusion can result).

We have the following preliminary result:[6]

**Proposition 2.1.** *Under Assumptions 1–4, every coordination game $\mathcal{G}$ has an introspective equilibrium, and it is essentially unique (i.e., introspective equilibrium uniquely determines behavior for all but a measure-0 set of types). Introspective equilibrium is monotone in type: There is a $\tau \in T$ such that all types $t > \tau$ choose $s^1$ and all types $t < \tau$ choose $s^2$. Moreover, introspective equilibrium is monotone in payoffs: When an action becomes more attractive in terms of payoffs, players are more likely to choose it (i.e., the equilibrium threshold $\tau$ increases with $\rho$).*

The existence result shows that, 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)), players can reach consistent expectations.[7] The result that introspective equilibrium is monotone in payoffs means that any non-monotonicities in the value we may find are driven by the trade-off between direct and indirect effects, not by non-monotonicities in behavior. The uniqueness result will be a critical first step towards deriving comparative statics results. However, while introspective equilibrium is unique for any given type space, it may vary across type spaces even if payoffs are held fixed. Thus, while uniqueness is helpful, deriving comparative statics that hold across type spaces still requires working with sets of equilibria.

The proof of Proposition 2.1 is mostly standard (Van Zandt and Vives, 2007). Suppose that, at level 1, action $s^1$ is a strict best response for type $\tau^0$, that is,

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

[6] Kets and Sandroni (2021) prove similar results for a related setting.

[7] 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 (i.e., assumptions on type spaces).

It is easy to check that this inequality holds if and only if $F(\tau^0 \mid \tau^0) < 1-\rho$. By (MON-B), there is a unique $\tau^1 < \tau^0$ such that, at level 1, each type $t$ chooses $s^1$ if $t > \tau^1$ and chooses $s^2$ if $t < \tau^1$ (i.e., $\tau^1$ solves $F(\tau^0 \mid \tau^1) = 1-\rho$ or $\tau^1 = 0$); moreover, by (REG), $F(\tau^1 \mid \tau^1) < 1-\rho$. So, players are more likely to choose $s^1$ at level 1 than at level 0. Because players have an incentive to coordinate, this increases the incentive to choose $s^1$ at higher levels: By a simple inductive argument, for each $k > 0$, there is a (unique) level-$k$ threshold $\tau^k$ such that types $t < \tau^k$ choose $s^2$ at level $k$, and types $t > \tau^k$ choose $s^1$ (i.e., $\tau^k$ solves $F(\tau^{k-1} \mid \tau^k) = 1-\rho$ or $\tau^k = 0$). Moreover, the thresholds decrease with $k$ (i.e., $\tau^k \leq \tau^{k-1}$). By the monotone convergence theorem, the sequence $\{\tau^k\}_k$ converges to a threshold $\tau$ such that a player of type $t$ chooses $s^1$ in introspective equilibrium if $t > \tau$ and $s^2$ if $t < \tau$, where the equilibrium threshold $\tau$ is the largest type $t$ smaller than $\tau^0$ such that $F(\tau \mid \tau) = 1-\rho$, if such a type exists, and $\tau = 0$ otherwise. By a similar argument, if action $s^2$ is a best response for $\tau^0$ at level 1, the introspective process again converges, and the equilibrium threshold $\tau$ is the smallest type $t$ not smaller than $\tau^0$ such that $F(\tau \mid \tau) = 1-\rho$ if such a type exists, and $\tau = 1$ otherwise. Hence, an introspective equilibrium exists. It is essentially unique: It pins down behavior for all types except the threshold type $\tau$. Behavior is monotone in type: Type $t$ chooses $s^1$ if $t > \tau$, and $s^2$ if $t < \tau$. If $\tau = 0$ (resp. $\tau = 1$), introspective equilibrium coincides with the Nash equilibrium where both players choose $s^1$ (resp. $s^2$); if $\tau \in (0,1)$, there is miscoordination in introspective equilibrium, with players choosing both actions with positive probability. To see why introspective equilibrium is monotone in payoffs, note that increasing the payoffs to action $s^1$ decreases $\rho$ and hence the level-1 threshold $\tau^1$, i.e., players are more likely to choose $s^1$ at level 1. By a simple inductive argument, the level-$k$ threshold falls for all $k > 0$. Hence, players are more likely to choose $s^1$ when its payoffs improve.

We will also use an additional assumption for some of our results. Given an introspective type space $\mathcal{T} = (F, \tau^0)$, let the *rank belief* of type $t$ 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 we say that the introspective type space $\mathcal{T}$ induces *non-monotone rank beliefs* if it satisfies the following condition:

**Assumption 5 (NMRB).** *There exists a $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$ or there exists a $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both).*

As its name suggests, this condition ensures that the rank belief function $F(t \mid t)$ is non-monotone (cf. Lemma C.1 in the appendix). Figure 1(a) illustrates the condition for an example type space. As the following result shows, Assumption (NMRB) is necessary and sufficient to obtain miscoordination for an open set of payoff parameters:

**Lemma 2.2.** *Under Assumptions 1–4, 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 Assumption 5).*

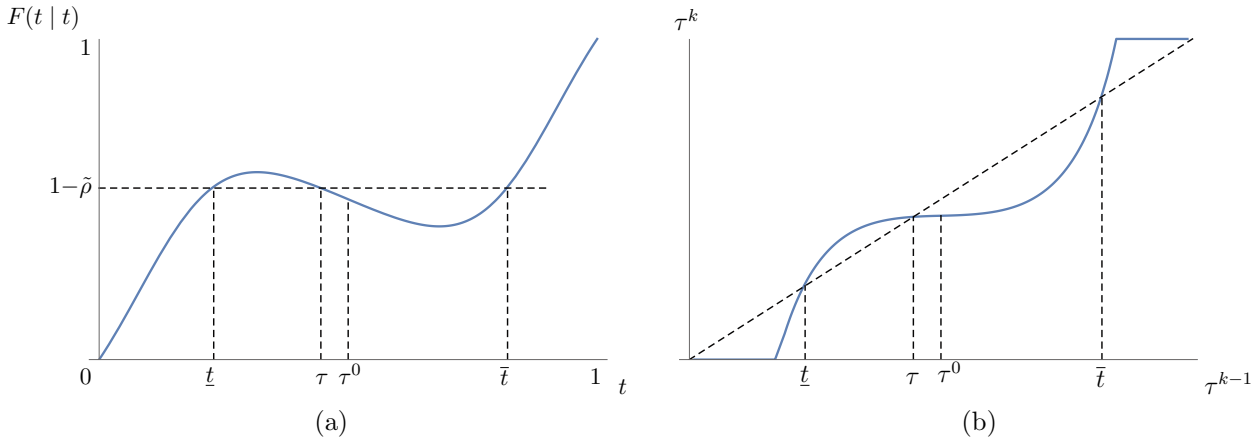


Figure 1: (a) The rank belief function $F(t \mid t)$ for an introspective type space that induces non-monotone rank beliefs. (b) The corresponding best-response function for dominance parameter $\tilde{\rho}$.

To see the intuition behind Lemma 2.2, consider the best-response function in Figure 1(b). The best-response function maps the level-$(k-1)$ threshold $\tau^{k-1}$ into the level-$k$ threshold $\tau^k$ for the dominance parameter $\tilde{\rho}$ in Figure 1(a). By Assumption (MON-B), the best-response function is increasing; moreover, by our discussion of Proposition 2.1 above, we have $\tau^k < \tau^{k-1}$ whenever it is the case that $\tau^{k-1} \in (\tau, \bar{t}\,)$. So, the best-response function lies below the diagonal for $\tau^{k-1} \in (\tau, \bar{t}\,)$. 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 for $\tilde{\rho}$ intersects the diagonal from above at $\tau$, and $\tau$ is a stable (attracting) fixed point. If we increase $\tilde{\rho}$ a little, the best-response function shifts up by a little (Proposition 2.1), but it continues to have a stable interior fixed point. Hence, there is miscoordination for an open set of payoff parameters.

Assumption (NMRB) is more novel and therefore less well understood than the other conditions. We therefore first illustrate its properties in the context of concrete applications (Sections 3.2–3.3) before discussing it in more abstract terms in Section 5.

## 2.3 Value

The value of a game is the ex ante expected payoff for a player in introspective equilibrium. An ex ante definition seems well suited for assessing the welfare implications of policies (as in our applications).[8] Formally, given a game form $\boldsymbol{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ and introspective type

[8] 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 influence his and other players' behavior.

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 $\sigma = (\sigma_1, \sigma_2)$ when he has type $t_1$ and the other player has type $t_2$. By Proposition 2.1, the value of a game is well-defined: The uniqueness of introspective equilibrium implies that the expected payoff for each player is well-defined. Moreover, because introspective equilibrium is symmetric, each player receives the same expected payoff.

# 3 Results

## 3.1 Characterization when the type space is known

To provide insights into the determinants of the value, we first consider the case where the type space is known. The next section provides testable comparative statics that hold across type spaces.

The following result fully characterizes the value of coordination games for a given introspective type space:

**Proposition 3.1 (The Value of a Coordination Game for a Given Type Space).** *For any introspective type space $\mathcal{T}$ that satisfies Assumptions 1–5, there exist $\underline{\rho}, \overline{\rho}$ with $0 < \underline{\rho} < \overline{\rho} < 1$ such that for any game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$ with dominance parameter $\rho$,*

*(a) if $\rho < \underline{\rho}$, both players choose action $s^1$ and the value is equal to $u_{11}$;*

*(b) if $\rho > \overline{\rho}$, both players choose action $s^2$ and the value is equal to $u_{22}$;*

*(c) if $\rho \in (\underline{\rho}, \overline{\rho})$, then the value is generically not equal to the expected payoff in any of the Nash equilibria (pure or mixed). In fact,*

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

*where $F(t) := F(t, 1)$ is the marginal distribution function of a player's type and $\tau \in (0, 1)$ is the equilibrium threshold.*

Proposition 3.1 completely characterizes the value for any given game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$. The characterization uses that, for any given game, introspective equilibrium is unique (Proposition 2.1). However, the characterization depends on the details of the type space $\mathcal{T}$. For example, the value under miscoordination in Eq. (1) depends on the type space through the common prior $F$ as well

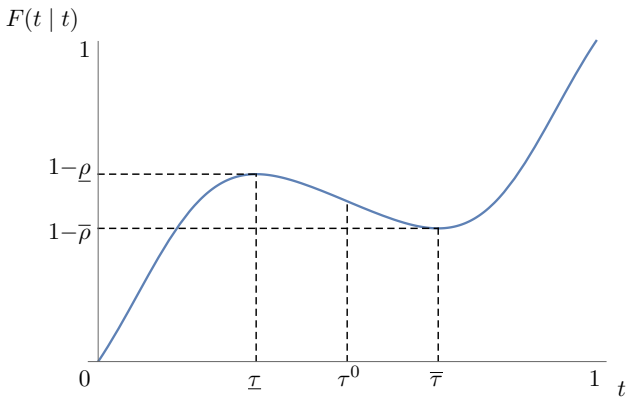


Figure 2: The rank belief function $F(t \mid t)$ from Figure 1 with the bounds $\underline{\rho}, \overline{\rho}$ that mark the regime with miscoordination.

as the equilibrium threshold $\tau$. Moreover, the dominance parameters $\underline{\rho}, \overline{\rho}$ that mark the regime with miscoordination depend on the rank belief function (and $\tau^0$): The proof of Proposition 3.1 shows that

$$1 - \underline{\rho} = \max\{F(t \mid t) : t \in [0, \tau^0]\};$$
$$1 - \overline{\rho} = \min\{F(t \mid t) : t \in [\tau^0, 1]\}.$$

We denote the corresponding equilibrium thresholds by $\underline{\tau}$ and $\overline{\tau}$, respectively. That is,

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

see Figure 2 for an illustration.

Proposition 3.1 shows that a key determinant of the value is the relative strength of economic and non-economic factors: When one of the actions, say $s^m$, stands out in terms of payoffs (i.e., $\rho < \underline{\rho}$ or $\rho > \overline{\rho}$), then both players choose it, and the value is equal to the payoff $u_{mm}$ in the corresponding Nash equilibrium. So, 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 ($u_{11} > u_{22}$). When the payoff structure provides little guidance (i.e., $\rho \in (\underline{\rho}, \overline{\rho})$), players' decisions may depend on non-economic factors: Some types choose $s^1$ while others choose $s^2$ (i.e., $\tau \in (0, 1)$). Thus, there is miscoordination, and the value is not equal to the payoff in any of the pure Nash equilibria. The value is also not equal to the expected payoff in the mixed Nash equilibrium. Intuitively, in introspective equilibrium, players are more likely to coordinate than in mixed Nash equilibrium (by (MON-B)). This captures the idea from Schelling (1960) that non-economic factors such as salience can facilitate coordination.

An important implication of Proposition 3.1 is that the comparative statics on the value are driven by the interplay between direct and indirect strategic effects. When one of the actions

stands out in terms of payoffs (i.e., $\rho < \underline{\rho}$ or $\rho > \overline{\rho}$), there are only direct effects as a small change in payoffs does not affect how the game is played. However, when the payoff structure provides little guidance (i.e., $\rho \in (\underline{\rho}, \overline{\rho})$) or when a change in payoffs induces miscoordination (i.e., when $\rho$ crosses one of the bounds $\underline{\rho}, \overline{\rho}$), there can be strategic effects. Whether the direct or indirect effect dominates typically depends on the details of the type space.

To illustrate, consider a game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$. Suppose that there is miscoordination (i.e., $\tau \in (0,1)$), so the value is given by Eq. (1). An increase in a payoff parameter, say $u_{22}$, has both a direct payoff effect and an indirect strategic effect. The net impact on the value depends on the relative magnitude of each effect. It turns out that the direct and indirect effects can be separated when there is miscoordination: Appendix A shows that, if the rank belief function is differentiable at $\tau$, the change in value as a function of $u_{22}$ is given by

$$\frac{\partial V}{\partial u_{22}} = p_{22}(\tau) + (u_{12} - u_{21})\, f(\tau)\, \frac{\partial \tau}{\partial u_{22}}, \tag{2}$$

where $f(t) := F'(t)$ is the marginal probability density of a player's type and $p_{22}(\tau)$ is the probability that both players choose $s^2$ in introspective equilibrium; similar expressions obtain for changes to the other payoff parameters $u_{nm}$. The first term is the direct payoff effect: keeping players' behavior fixed, the change in value when $u_{22}$ increases is directly proportional to the probability that players coordinate on $(s^2, s^2)$. The second term in Eq. (2) is the indirect strategic effect; see Appendix A for details. Which effect dominates obviously depends on the payoff parameters, but also on the details of the type space, such as the probability $p_{22}(\tau)$ and the effect $\partial\tau/\partial u_{22}$ of a change in $u_{22}$ on the equilibrium threshold $\tau$. Appendix A shows that the latter can be tied directly to the derivative of the rank belief function $F(t \mid t)$ with respect to $t$ at the equilibrium threshold $t = \tau$.

Thus, how the value changes with payoffs can depend on the details of the type space in intricate ways. The next sections show that it is nevertheless possible to provide comparative statics on the value even when the type space is not known or difficult to measure. Section 3.2 considers the costs of miscoordination, and Section 3.3 compares the costs of miscoordination and coordination failure. Since the value depends on all payoff parameters (e.g., Eq. (1)), not just the "summary statistic" $\rho$, we focus on specific applications throughout.

## 3.2 The costs of miscoordination

This section derives testable predictions on the costs of miscoordination. For some of the results in this section, we make the additional assumption that *no action is strongly salient*, i.e., $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$. This assumption captures the idea that, when the payoff structure provides no guidance (i.e., $\rho = \frac{1}{2}$), players have no ground to choose one action over the other, i.e., there is miscoordination ($\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$). Note that this assumption on the type space implies non-monotone

rank beliefs (Assumption 5); at the end of this section, we discuss a class of type spaces that naturally satisfies this condition.

We start with the following simple coordination game, denoted $\boldsymbol{u}_w$:

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

$w \geq 1$

The following result provides testable comparative statics on how the value of $\boldsymbol{u}_w$ varies with $w$:

**Proposition 3.2 (The Value of $\boldsymbol{u}_w$).** *Under Assumptions 1–5:*

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

*(b) If $w = 1$ and no action is strongly salient, then the value of $\boldsymbol{u}_w$ 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 (a) yields the intuitive prediction that for $w$ sufficiently large, the value of $\boldsymbol{u}_w$ equals $w$.[9] Part (b) 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 is consistent with experimental evidence that subjects' payoffs in the coordination game $\boldsymbol{u}_w$ with $w = 1$ generally exceeds that in the mixed Nash equilibrium yet is less than that in pure Nash equilibrium. For example, Mehta et al. (1994, p. 668) shows experimentally that if one of the actions has a salient label, then the value of $\boldsymbol{u}_w$ with $w = 1$ is 0.76.

We next consider the following variant of $\boldsymbol{u}_w$, denoted $\tilde{\boldsymbol{u}}_x$:

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

$1 \leq 1 + x < w;\ -c < 1 + x$

The following result shows that, in settings like this, players can be worse off if the payoffs to an initially unplayed Nash equilibrium increase:

**Proposition 3.3 (The Value of $\tilde{\boldsymbol{u}}_x$).** *If no action is strongly salient, then for $w > 1$ sufficiently large, the value of $\tilde{\boldsymbol{u}}_x$ is strictly lower when $x$ is close to $w - 1$ than when $x$ equals 0. That is, for every introspective type space $\mathcal{T}$ that satisfies Assumptions 1–5, if $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$, there is a $w^*$ such that for $w > w^*$, $\limsup_{x \uparrow w-1} V((w, -c, 0, 1 + x); \mathcal{T}) < V((w, -c, 0, 1); \mathcal{T})$.*

[9] We are not aware of experimental papers that study coordination games like $\boldsymbol{u}_w$ for $w > 1$. However, this could be due to a selection effect: If it is obvious that the value is $w$ when $w$ is large, there is no reason to study the game experimentally. See Schmidt et al. (2003, p. 285) for comments along these lines.

Proposition 3.3 shows that increasing payoffs can reduce the value: While all payoff parameters $u_{nm}$ are (weakly) higher when $x$ is close to $w-1$ than when $x$ equals 0, the value is strictly lower (provided that $w$ is sufficiently high). Intuitively, there is a tension between the direct payoff effect and the indirect strategic effect of an increase in $x$. 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 as action $s^2$ becomes more attractive in terms of payoffs (i.e., $x$ approaches $w-1>0$), there is more miscoordination. This has a negative effect on the value. Proposition 3.3 shows that the indirect strategic effect dominates for $w$ sufficiently high. Thus, increasing the payoff $u_{22}$ to the Pareto-dominated equilibrium can reduce the value. While simple, this insight is critical for understanding why policy changes that create direct benefits may ultimately reduce welfare. For example, if labor supply decisions are strategic complements, as in Lindbeck et al. (1999), an increase in unemployment benefits may make workers worse off.

Propositions 3.2–3.3 show the intuitive result that miscoordination is costly: The value under miscoordination is strictly lower than in the benchmark case where players coordinate on the Pareto-dominant Nash equilibrium. While intuitive, this insight is difficult to formalize using traditional game-theoretic models. For example, an increase in $x$ does not change the set of pure Nash equilibria of $\tilde{\boldsymbol{u}}_x$, while mixed Nash equilibrium predicts it *reduces* the likelihood that players choose $s^2$. Standard equilibrium refinements are also unable to capture this result. For example, both payoff dominance and risk dominance predict that the value of $\tilde{\boldsymbol{u}}_x$ is (weakly) higher when $x=w-1>0$ than when $x=0$. We discuss the predictions from alternative solution concepts in Section 5.5.

Propositions 3.2–3.3 are testable: they hold across type spaces. To better understand which kind of economic situations are adequately modeled by type spaces that induce the effects in Propositions 3.2–3.3, it is nevertheless valuable to consider specific type spaces. Appendix B.1 discusses a class of introspective type spaces that naturally fits experimental settings and that induces a rank belief function as in Figure 1. We refer to this type space as the *social salience type space*. The basic idea is that, with some probability $p\in(0,1)$, action $s^1$ is "socially salient" in the sense that both players are likely to have an impulse to choose $s^1$; with the remaining probability $1-p$, action $s^2$ is socially salient (i.e., both players are likely to have an impulse to choose $s^2$). When both actions are equally likely to be socially salient (i.e., $p=\frac{1}{2}$), this type space satisfies the condition that no action is strongly salient. In experiments, higher-order beliefs consistent with this introspective type space can be generated directly using a correlating device (Fehr et al., 2019), or more indirectly using salient action labels (Mehta et al., 1994) or public announcements (Duffy and Fisher, 2005). For example, in Mehta et al.'s experiments, subjects play a coordination game like $\boldsymbol{u}_w$ with $w=1$, but with salient action labels (e.g., $s^1$ and $s^2$ are replaced by "heads" and "tails"). Ex ante, each of the two actions is equally likely to

be salient (i.e., $p = \frac{1}{2}$); once the action labels are assigned, subjects are likely to have an impulse to choose the action with the more salient action label.

## 3.3 Miscoordination versus coordination failure

This section derives testable predictions on when miscoordination is more costly than coordination failure (i.e., the value under miscoordination is lower than the value under coordination failure). We do so in the context of two applications.

### 3.3.1 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, consider the game form $\boldsymbol{u}^s = (u_{11} + s, u_{12} + s, u_{21}, u_{22})$, where $s \geq 0$. The following result characterizes the conditions under which an investment subsidy *decreases* the value as it eliminates coordination failure but induces miscoordination (i.e., the dominance parameter falls from $\rho > \overline{\rho}$ to $\overline{\rho}$):

**Theorem 3.4 (The Value of Investment Subsidies).** *Fix an introspective type space that satisfies Assumptions 1–5 and is such that there is a positive probability of investment at $\overline{\rho}$ (i.e., $\overline{\tau} < 1$). Suppose there is coordination failure if there is no investment subsidy (i.e., $\rho > \overline{\rho}$ when $s = 0$). Then, the value strictly decreases with the 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 $\rho$ is not too high. That is, there is a $\rho^* \in (\overline{\rho}, 1)$ 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 $\boldsymbol{u}^s = (u_{11}+s, u_{12}+s, u_{21}, u_{22})$ with dominance parameter $\rho$ at $s = 0$, 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 $u_{21} < u_{21}^*$.*

Theorem 3.4 characterizes the conditions under which miscoordination is more costly than coordination failure for all type spaces that satisfy our conditions. An important implication of Theorem 3.4 is that introducing an investment subsidy 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. Theorem 3.4 shows that this happens precisely when 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)

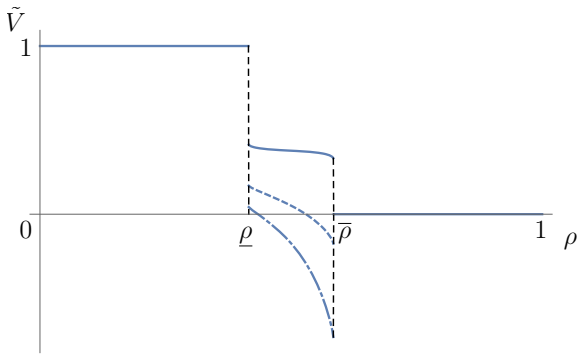


Figure 3: The normalized value $\tilde{V} := \frac{V-u_{22}}{u_{11}-u_{22}}$ as the investment subsidy $s$ is varied, plotted as a function of $\rho = \rho(s)$. The solid, dashed, and dash-dotted lines correspond to the game forms $\boldsymbol{u}^0 = (1, -\frac{3}{5}, \frac{3}{5}, 0)$, $\boldsymbol{u}^0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

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 the 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), i.e., $\rho$ is close to $\overline{\rho}$. 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. Thus, *for a policy to improve welfare, it is not sufficient that it changes behavior in the desired direction (i.e., increases investment), it must also ensure that the costs of miscoordination are not too high.* These insights are robust: The online appendix shows that the same result obtains for an alternative policy to promote investment.

To better understand Theorem 3.4, notice that, for any given type space $\mathcal{T}$, the difference between the value under miscoordination (with subsidy $\overline{s} > 0$, $\rho = \overline{\rho}$) and the value under coordination failure (without a subsidy, $\rho > \overline{\rho}$) is given by

$$\Delta^{\mathcal{T}} = \overline{p_{11}}\,(u_{11} + \overline{s}\,) + \overline{p_{12}}\,(u_{12} + \overline{s}\,) + \overline{p_{21}}\,u_{21} + \overline{p_{22}}\,u_{22} - u_{22}, \tag{3}$$

where $\overline{p_{nm}}$ is the probability that players play according to $(s^n, s^m)$ at $\overline{\rho}$. Unfortunately, this expression for $\Delta^{\mathcal{T}}$ does not provide testable predictions on when miscoordination is more costly than coordination failure. For example, while the value under miscoordination is higher when $u_{12}$ and $u_{21}$ are higher, ceteris paribus, we cannot conclude that the costs of miscoordination are low compared to the costs of coordination failure when $u_{12}$ or $u_{21}$ are high.[10] This is because the

[10]Indeed, this does not follow from Theorem 3.4: The conditions on $u_{12}$ and $u_{21}$ in Theorem 3.4 depend on

subsidy $\overline{s}$ has both a direct payoff effect (which is positive) as well as an indirect strategic effect (i.e., it eliminates coordination failure at the expense of inducing miscoordination). Importantly, both the subsidy $\overline{s}$ that induces miscoordination (which drives the direct effect) as well as the action distribution $(\overline{p_{11}}, \overline{p_{12}}, \overline{p_{21}}, \overline{p_{22}})$ under miscoordination (which drives the indirect effect) depends on the type space. This means that the trade-off between direct and indirect effects depends on the details of the type space (cf. Proposition 3.1). As a result, we cannot easily separate the direct and indirect effects if we want our predictions to hold across type spaces, unlike in the case where the type space is known (Section 3.1). As part of the proof, we therefore rewrite the expression for $\Delta^{\mathcal{T}}$ so as to separate out the terms involving the type space from those involving the underlying payoff parameters (i.e., $u_{11}, u_{12}, u_{21}, u_{22}$ and thus $\rho$). This delivers testable predictions: The proof shows that for any given payoffs $u_{11}, u_{22}$ and dominance parameter $\rho$ less than some bound $\rho^*$ that depends only on the type space, miscoordination is more costly than coordination failure if and only if a term that decreases in the off-diagonal payoffs $u_{12}, u_{21}$ is larger than a term that depends only on the type space, through the bound $\rho^*$ (Eq. (12) in the appendix). To see why this prediction is testable despite the fact that the precise bounds $u_{12}^*, u_{21}^*, \rho^*$ depend on the type space, note that, as we vary the relevant payoff parameters along the directions in Theorem 3.4, we move between the regime where miscoordination is more costly than coordination failure to the regime where the opposite is true (cf. Figure 3).

**Animal spirits** While the predictions in Theorem 3.4 hold for any type space that satisfies our conditions, it will be instructive to consider which kind of situations can be modeled by a type space that induces the effects in Theorem 3.4. 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 individuals' impulses. To model this, we follow Morris and Yildiz (2019) and assume that each player's type $\tilde{t}_j$ is the sum of a common shock $\eta$ that affects both players, and an idiosyncratic term $\varepsilon_j$ 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 and beliefs about others' impulses.[11] Under this interpretation, the common shock $\eta$ reflects a shift in public sentiment while $\varepsilon_j$ 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 $\eta$ in public sentiment can be large. Following Morris and Yildiz, we therefore assume that the common shock $\eta$ is drawn from a fat-tailed distribution

other parameters such as $u_{11}$ and $u_{22}$.

[11] The payoff-based approach in Morris and Yildiz (2019) predicts multiple equilibria when the payoff structure provides little guidance; the same is true for sunspot models (e.g., Cass and Shell, 1983), an important belief-based approach. The predictions from these approaches are therefore difficult to test (even when the type space is known).

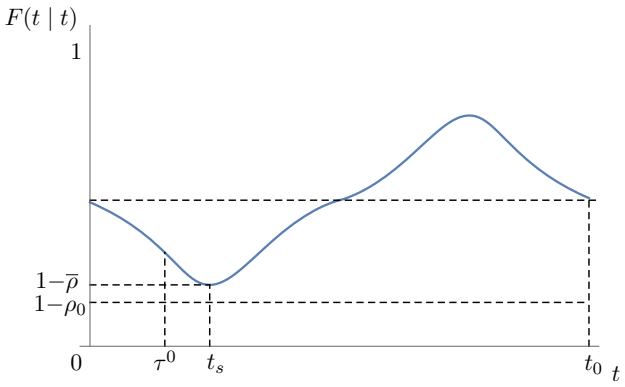


Figure 4: Introspective equilibria for the "animal spirits" type space: without an investment subsidy (dominance parameter $\rho_0$, equilibrium threshold $t_0 = 1$) and with an investment subsidy $s > 0$ just sufficient to induce miscoordination (dominance parameter $\overline{\rho}$, equilibrium threshold $t_s = \overline{\tau}$).

(e.g., a $t$-distribution) while the distribution of the idiosyncratic terms $\varepsilon_j$ has thin tails (e.g., a 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 $\mathcal{I}_j(t_j)$. Figure 4 illustrates the rank belief function for the resulting introspective type space; notice that, unlike in the social salience type space in Figure 2, the rank belief function is now decreasing for extreme types (i.e., $t$ close to 0 or 1); see Morris and Yildiz (2019, pp. 2831–2832) for a discussion of the intuition behind this shape.

Our model predicts that welfare may be lower under partial investment than if there is no investment at all. Figure 4 shows a particular example. In this example, there is no investment when there is no subsidy: When the dominance parameter is $\rho_0$ (corresponding to subsidy $s = 0$), the equilibrium threshold is $t_0 = 1$, i.e., there is coordination failure in introspective equilibrium. Introducing an investment subsidy $s > 0$ increases the probability of investment (i.e., the equilibrium threshold $t_s$ decreases with $s$), and as the associated dominance parameter $\rho_s$ crosses $\overline{\rho}$, there is a positive probability of investment in introspective equilibrium. For example, when the investment subsidy $s$ is such that $\rho_s = \overline{\rho}$, the equilibrium threshold is $t_s = \overline{\tau} \in (0,1)$. Theorem 3.4 shows that, unless the subsidy is so large that full investment is guaranteed (i.e., $\rho_s$ crosses $\underline{\rho}$ and $t_s = 0$), players may be worse off with an investment subsidy than without one. This happens precisely when the costs of miscoordination exceed the costs of coordination failure. While Theorem 3.4 focuses on predictions that hold across type spaces, its proof (especially Eq. (12)) can be used to obtain sharper predictions for any given type space (such as the one in Figure 4).

### 3.3.2 Collusion

We next consider the problem of tacit collusion: Firms benefit from keeping prices high, but cannot explicitly exchange information and achieve agreement about coordinating their actions. This means that there is both scope for coordination failure (i.e., a failure to collude even though it would make both firms better off) and miscoordination (one firm attempts to initiate collusion, but the other fails to do so).

We consider a simple model of price competition with product differentiation (Ross, 1992). There are two firms that each 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

$$p_i = a - b\, q_i - c\, q_{-i},$$

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, $r := c/b \in (0, 1)$ measures the degree of substitutability: In the limit $r \to 1$, the products are perfect substitutes, and in the limit $r \to 0$, the demands for the two products are independent. For simplicity, the marginal cost 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 profit if the other firm charges the collusive price. A firm's per-period profit can then be one of the following: the "collusive" profit $\pi^*$ if both firms charge $p^*$; the Bertrand–Nash profit $\pi^N$ if both firms charge the Bertrand–Nash price $p^N$; the "cheating" profit $\pi^c$ if the firm's price is $p^c$ while it's competitor charges $p^*$; the "mutual cheating" profit $\pi^m$ if both firms cheat and charge $p^c$; or the "victim" profit $\pi^v$ if the firm charges $p^*$ and the other firm cheats. Lemma C.4 in the appendix shows that $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$. This implies that the (one-shot) price competition game has the structure of a prisoner's dilemma.

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.[12] Under the *collusive strategy* (denoted by $\sigma^*$), 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 $\sigma^c$), 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

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

discounted sum of profits) under the various strategy combinations are given by

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

where we have listed only the row player's payoff for simplicity. To rule out trivialities, we assume that the discount factor $\delta$ is sufficiently high for both players choosing $\sigma^*$ 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 $\sigma^*$ and $s^2$ with $\sigma^c$, 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 $\sigma^*$) but collusion breaks down because the other firm cheats (i.e., chooses $\sigma^c$). As firms become more patient (i.e., $\delta$ 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:

**Theorem 3.5 (The Value of Collusion).** *Fix an introspective type space that satisfies Assumptions 1–5, and fix the 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 $\delta$ (i.e, $\rho(\delta) > \overline{\rho}$) but not when the discount factor is $\delta'$ (i.e., $\rho(\delta') < \overline{\rho}$). Then, if either $\delta'$ is so large that $\rho(\delta') < \underline{\rho}$ or $\delta' - \delta$ is sufficiently small, the value of the game with discount factor $\delta'$ (with miscoordination) is strictly larger than the value of the game with discount factor $\delta$ (with coordination failure).*

Theorem 3.5 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 surprising, perhaps, is the fact 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 when 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., $\rho > \overline{\rho}$). In this case, the direct effects are complex: Increasing the discount factor from $\delta$ to $\delta'$ increases $u_{12}$ but reduces $u_{21}$ and $u_{22}$ (while leaving $u_{11}$ unchanged). Theorem 3.5 shows that the positive effects dominate provided that the increase $\delta' - \delta$ is not too large.

One important implication of Theorem 3.5 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).[13] These predictions are in line with empirical evidence. For example, the US government's practice to buy vaccines in

[13] 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, $\delta$ should be replaced by $\delta^T$ 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

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 3.5 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 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., $p_{11} > 0$).[14] Again, these insights are robust: The online appendix shows that similar results obtain for other commonly-used models of collusion.

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.

The proof strategy we use for Theorem 3.5 is similar to, but somewhat distinct from that for Theorem 3.4. As before, we consider the difference $\Delta^{\mathcal{T}}$ between the value under miscoordination and the value under coordination failure (cf. Eq. (3)). However, because the direct payoff effects are complex – with a change in the discount factor affecting different payoff parameters at different rates – it is difficult to derive necessary and sufficient conditions under which miscoordination is more costly than coordination failure for this case. However, we show that it is possible to derive sufficient conditions under which coordination failure is more costly. That is, we derive a lower bound on $\Delta^{\mathcal{T}}$ (Lemma C.2 in the appendix). This delivers the testable predictions in

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

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

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.

[14] 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 3.5 suggests the striking conjecture that experimenting to identify appropriate collusive actions may carry little, if any, cost, relative to the baseline without collusion, implying that tacit collusion can be an important threat even in these more complex settings.

Theorem 3.5. In fact, because the lower bound in Lemma C.2 applies to any coordination game, it can be used for other applications as well.

# 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).[15] At any time $\tilde{t}$, the social context is modeled by the introspective type space $\mathcal{T}_{\tilde{t}} = (F, \tau^0_{\tilde{t}})$, where $\mathcal{T}_{\tilde{t}}$ satisfies Assumption 1–4. So, the share of matches involving types $t_1, t_2 \in T$ is $f(t_1, t_2)$. The key assumption is that players' 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 $\tau^0_0$ and the equilibrium threshold $\tau_{\tilde{t}-1}$ at time $\tilde{t} - 1$, with some noise $\varepsilon > 0$:

$$\min\{\tau^0_0, \tau_{\tilde{t}-1}\} - \varepsilon < \tau^0_{\tilde{t}} < \max\{\tau^0_0, \tau_{\tilde{t}-1}\} + \varepsilon. \tag{4}$$

The assumption that the original impulses (i.e., $\tau^0_0$) 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. The noise $\varepsilon$ captures the idea that there can be slight shocks to the social context. 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 Eq. (4) (for given $\varepsilon$).

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

**Proposition 4.1 (Introspective Equilibrium as a Steady State).** *If the noise is sufficiently small, then (generically) the introspective equilibrium remains largely unchanged over time: For*

[15] 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.

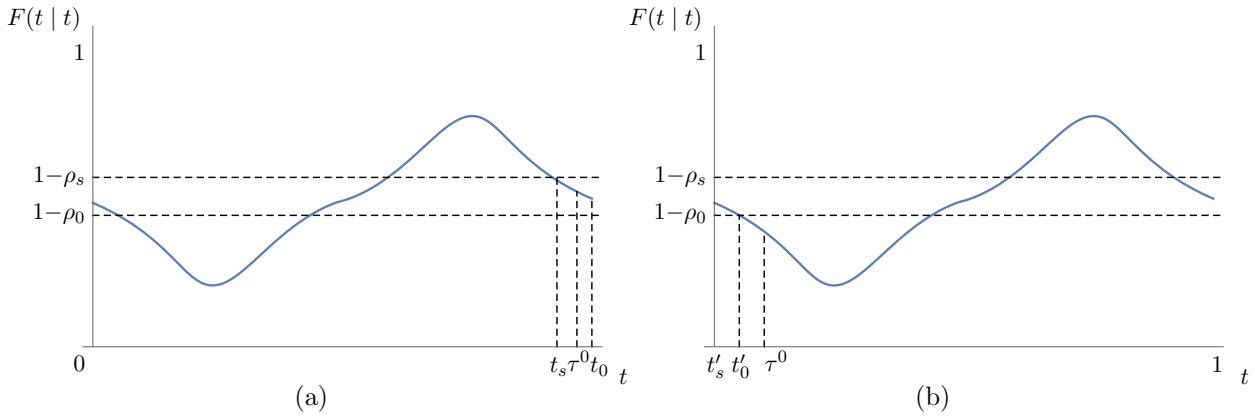


Figure 5: Introspective equilibria for the "animal spirits" type space with and without investment subsidy (with respective dominance parameters $\rho_s$ and $\rho_0$): (a) when players have an impulse to invest only if they expect public sentiment to favor investment; (b) when players have an impulse to invest unless they expect public sentiment to be strongly against investing.

*every $\chi > 0$, there is an $\varepsilon > 0$ such that for every dynamic $\{\tau^0_{\tilde{t}}\}_{\tilde{t}}$ that satisfies Eq. (4) and every pair of periods $\tilde{t}, \tilde{t}'$, the equilibrium thresholds $\tau_{\tilde{t}}, \tau_{\tilde{t}'}$ are within $\chi$ of each other, i.e., $|\tau_{\tilde{t}} - \tau_{\tilde{t}'}| < \chi$.*

Proposition 4.1 implies that when players' impulses may be driven by past population play, then they will continue to behave similarly. 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 $(\underline{t}, \bar{t})$: for any $\tau^0 \in (\underline{t}, \bar{t})$, the introspective process converges to the introspective equilibrium $\tau$. Because the basin of attraction contains both the level-0 threshold and the equilibrium threshold, any introspective process that begins between these two thresholds converges to the same equilibrium threshold $\tau$. Proposition 4.1 shows that this insight extends to the case where there can be shocks to the social context.

We apply Proposition 4.1 to our investment example (Section 3.3.1). Proposition 4.1 suggests that there can be *path dependence*. 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 in Figure 5. Figure 5 considers two extreme cases: one where players have an impulse to invest only if they expect public sentiment to strongly favor investment (i.e., $\tau^0$ close to 1; panel (a)), and one in which players have an impulse to invest unless they expect public sentiment to be strongly against investing (i.e., $\tau^0$ close to 0; panel (b)). Contrasting panels (a) and (b), we see remarkably different outcomes even for the same payoff environment. When players have an impulse to invest only if they expect public sentiment to favor investment (panel (a)), there is no or little investment in equilibrium even with subsidies in place (i.e., $t_0 = 1$, $t_s$ close to 1). Under the

alternative assumption that players have an impulse to invest unless they expect public sentiment to be against investing (panel (b)), there are high levels of investment in equilibrium even in the absence of subsidies (i.e., $t'_s = 0$, $t'_0$ close to 0). 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 4.1). 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, Theorem 3.4 shows that, depending on the environment, low but nonzero levels of investment can be costlier than no investment at all. 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.*

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.

We have given two examples of introspective type spaces that induce non-monotone rank beliefs (Sections 3.2–3.3). 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 salience, 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, animal spirits), aggregate uncertainty is not sufficient to induce non-monotone rank beliefs. To see this, note that the type spaces considered in the global games literature also feature aggregate uncertainty but do not induce non-monotone rank beliefs (Morris et al., 2016). 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, predictions that hold across introspective type spaces for given payoff parameters. That is, if an analyst has only limited information on players' impulses or beliefs, can he make any predictions about behavior or the value for a given game form $\boldsymbol{u}$?

A first observation is that introspective equilibrium rules out any behavior that is inconsistent with correlated equilibrium. We show this for any finite game form $\hat{\boldsymbol{u}} = \langle N, \{S_j\}_{j\in N}, \{u_j\}_{j\in N}\rangle$, where $N$ is a finite set of players, and for each player $j \in N$, $S_j$ is a finite set set of actions and $u_j : S_j \times S_{-j} \to \mathbb{R}$ is a payoff function. We also allow any introspective type space $\widehat{\mathcal{T}} = \langle (T_j)_{j\in N}, (\mathcal{I}_j)_{j\in N}, F\rangle$, where for each player $j$, $T_j$ is the set of types, taken to be a closed subset of the real line, $\mathcal{I}_j$ is a function that maps each type $t_j \in T_j$ into an impulse $\mathcal{I}_j(t_j) \in S_j$, taken to be Borel measurable, and $F$ is a common prior on $\prod_j T_j$. In particular, $\widehat{\mathcal{T}}$ does not need to satisfy any of the assumptions in Section 2. We then have:[16]

**Proposition 5.1.** *Fix a game $\widehat{\mathcal{G}} = (\hat{\boldsymbol{u}}, \widehat{\mathcal{T}})$ where $\hat{\boldsymbol{u}}, \widehat{\mathcal{T}}$ are as defined above. Any introspective equilibrium of $\widehat{\mathcal{G}}$ corresponds to a correlated equilibrium of the underlying game form $\hat{\boldsymbol{u}}$.*

Thus, behavior in introspective equilibrium is consistent with correlated equilibrium. Without imposing further restrictions, introspective equilibrium does not impose any further restrictions on behavior beyond behavior being consistent with correlated equilibrium.[17] This follows from a version of the revelation principle (Myerson, 1994): Given any correlated equilibrium of a game form $\hat{\boldsymbol{u}}$, simply choose the introspective type space $\widehat{\mathcal{T}}$ such that the action distribution induced by the level-0 strategy $\sigma^0$ coincides with that of the correlated equilibrium. Correlated equilibrium already imposes some, albeit limited, restrictions on behavior: For example,

[16] Kets and Sandroni (2021) prove a similar result.

[17] Restrictions on beliefs are also critical for other concepts to have cutting power: see, e.g., Brandenburger and Dekel (1987) on a posteriori equilibrium (a refinement of correlated equilibrium), Battigalli and Siniscalchi (2003) on Bayesian-Nash equilibrium, and Bergin and Lipman (1996) on evolutionary dynamics.

in any symmetric coordination game, players' behavior is positively correlated in the sense that $\mu_{11}\,\mu_{22} \geq \mu_{12}\,\mu_{21}$, where $\mu_{nm}$ is the probability of action profile $(s^n, s^m)$ in equilibrium (Calvó-Armengol, 2006).

By focusing on coordination games and imposing some arguably minimal assumptions on the type space (Section 2), we can say more: Any introspective equilibrium is symmetric (by the proof of Proposition 2.1) and strict (for generic $\boldsymbol{u}$). While simple, this observation is essentially sufficient to deliver the predictions for $\boldsymbol{u}_w$ with $w = 1$ (Proposition 3.2). For example, Proposition 3.2 rules out the mixed Nash equilibrium for $\boldsymbol{u}_w$ with $w = 1$, as well as arguably unreasonable predictions, such as the correlated equilibrium in which players randomize with equal probability over $(s^1, s^1)$, $(s^2, s^2)$, and $(s^1, s^2)$ (since this correlated equilibrium is neither symmetric nor strict).[18] Beyond this, introspective equilibrium also imposes further restrictions: While the set of (symmetric, strict) correlated equilibria is convex, the set of introspective equilibria need not be.[19] We leave a full characterization of all action distributions consistent with introspective equilibrium for future work.

## 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. To see this, fix a game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$, where $\mathcal{T} = (F, \tau^0)$ satisfies Assumptions 1–4. Suppose there is a change in payoffs, i.e., the game form changes to $\tilde{\boldsymbol{u}}$. A natural assumption is that players are more likely to choose an action when it becomes more attractive in terms of payoffs, i.e., the level-0 threshold increases with the dominance parameter. 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 or payoffs is not too large, it does not affect the introspective equilibrium:

[18] In fact, because the set of symmetric correlated equilibria has Lebesgue measure 0 in the set of all correlated equilibria (Calvó-Armengol, 2006), the set of introspective equilibria is small (in the standard measure-theoretic sense) relative to the set of all correlated equilibria.

[19] To be precise, fix a game form $\tilde{\boldsymbol{u}}$ and suppose $\mu, \mu'$ are correlated equilibria of $\tilde{\boldsymbol{u}}$. Also suppose that there are introspective type spaces $\widetilde{\mathcal{T}}$ and $\widetilde{\mathcal{T}}'$ such that the introspective equilibria of $\widetilde{\mathcal{G}} = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}})$ and $\widetilde{\mathcal{G}}' = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}}')$ induce behavior consistent with $\mu$ and $\mu'$, respectively. Then, while $\mu'' := \lambda\,\mu + (1 - \lambda)\,\mu'$ is a correlated equilibrium of $\tilde{\boldsymbol{u}}$ for any $\lambda \in (0, 1)$, there need not be an introspective type space $\widetilde{\mathcal{T}}''$ such that the introspective equilibrium of $\widetilde{\mathcal{G}}'' = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}}'')$ induces behavior consistent with $\mu''$.

**Corollary 5.2.** *Let $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$ and $\widetilde{\mathcal{G}} = (\tilde{\boldsymbol{u}}, \widetilde{\mathcal{T}})$ be as defined above. Then (generically), if $|\rho(\boldsymbol{u}) - \rho(\tilde{\boldsymbol{u}})|$ and $|\tau^0 - \tilde{\tau}^0|$ are not too large, the introspective equilibrium of $\widetilde{\mathcal{G}}$ coincides with the introspective equilibrium of the game $\widetilde{\mathcal{G}}' = (\tilde{\boldsymbol{u}}, \mathcal{T})$ that has the original type space $\mathcal{T}$.*

The proof is a straightforward adaptation of the proof of Proposition 4.1, so we refer to this result as a corollary. Corollary 5.2 shows that, even when the level-0 threshold changes with payoffs, as long as this change is not too large, the comparative statics results will not be affected by ignoring this effect. The intuition is similar to before: As long as the level-0 thresholds $\tau^0$ and $\tilde{\tau}^0$ lie in the same "basin of attraction," then our prediction will be independent of which of the two thresholds we use. So, at least in this sense, our results are robust to relaxing the assumption that impulses are independent of payoffs.

## 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 $\tilde{t}$, players choose a best response to the opponent's strategy in period $\tilde{t} - 1$, much like how in our model, for each level $k$, types choose a best response at level $k$ to the opponent's level-$(k-1)$ strategy. 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 use their mental state to form beliefs about other players. More precisely, introspective players form beliefs by conditioning on their (private) type. To see this, fix a game $(\mathcal{G}, \mathcal{T})$. Then, 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 $k-1$, and this posterior belief varies with $t$ (by Assumption (MON-B), as $\mu_{-j}^{k-1}(s^2 \mid t) = F(\tau^{k-1} \mid t)$). By contrast, under the best response dynamic, players do not condition their beliefs on any private information (for example, because they do not have any private information, as in the standard Cournot model, or because types are independent).

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. By contrast, the introspective process always converges and allows for miscoordination (i.e., miscoordination can be attracting; Lemma 2.2 and Figure 1). In fact, even if we consider stronger notions of stability, miscoordination remains stable: The online appendix shows that, under mild conditions, introspective equilibrium is (generically) asymptotically stable, i.e., it is both attracting and Lyapunov stable. Thus, miscoordination is asymptotically stable whenever the payoff structure of the game provides little guidance (i.e., $\rho$ intermediate) and Assumption 5 (NMRB) holds. 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).

Miscoordination can be stable in richer adaptive models, but the predictions of these models are otherwise fundamentally different from 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). In these models, mixed Nash equilibrium can be asymptotically stable or not, depending on the class of perturbations being considered (Ellison and Fudenberg, 2000; Echenique and Edlin, 2004; Sandholm, 2007). However, even if miscoordination is asymptotically stable, these payoff-based extensions of standard dynamics still deliver fundamentally different predictions: Because mixed Nash equilibrium predicts that players are less likely to choose an action when its payoffs improve, these models do not always deliver intuitive comparative statics on behavior, in contrast with introspective equilibrium (Proposition 2.1).

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 unfolds over time. One way to model this is to view each type as an agent, and the distribution $f(\cdot)$ as a local interaction network that governs the interactions between agents (Mailath et al., 1997; Morris, 1997; Kets, 2011). Our results suggest that, if a network satisfies an analogue of (NMRB), then an action may not spread to the entire population, i.e., $\tau \in (0,1)$ (cf. Morris, 2000). Which natural properties of a network imply (NMRB) 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 economic and non-economic factors affect the value, through their impact on 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 $\boldsymbol{u}_w$ and $\tilde{\boldsymbol{u}}_x$ 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 many learning-based refinements (Kandori et al., 1993; Young, 1993).[20] On the other hand, mixed Nash equilibrium predicts miscoordination but cannot account for coordination failure; moreover, it has unattractive comparative statics on behavior. 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.

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), precedent (Crawford and Haller, 1990), or symmetry (Alós-Ferrer and Kuzmics, 2013) 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 and show that our results hold 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.[21] However, these models do not deliver testable comparative

[20] 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). 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, 2022).

[21] This approach has been applied widely in economics, see, e.g., Kreps (1990) on corporate culture, Greif (1994) on economic history, Myerson (2004) on the foundations of political institutions, Ray (2004) on development,

statics. 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. While the comparative statics on equilibrium behavior are well-understood, this paper is among the first to deliver intuitive 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 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). The trade-off between direct and indirect effects is especially important when considering the relative costs of miscoordination (i.e., failing to coordinate on a pure Nash equilibrium) and coordination failure (i.e., coordinating on a Pareto-dominated Nash equilibrium). As we show, the direct and indirect effects can be subtle and interact with each other in intricate ways. We show that it is nevertheless possible to obtain testable comparative statics on the value. An important question for future research is to identify conditions under which policies that have the desired effect on behavior also have a positive impact on welfare, that is, when monotone comparative statics on behavior imply monotone comparative statics on the value.

# Acknowledgments

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.

and Cass and Shell (1983) on sunspots.

# Appendix A The value under miscoordination

This appendix derives Eq. (1) for the value for a given (known) type space and characterizes how the value changes with payoffs in the regime with miscoordination when the rank belief function is differentiable. We show that the change in value can be decomposed into a direct and an indirect effect in this case.

Consider a game $\mathcal{G} = (\boldsymbol{u}, \mathcal{T})$, where $\mathcal{T}$ satisfies Assumptions 1–5. Recall that, by the proof of Proposition 2.1, 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 $\boldsymbol{u}$ (given $\mathcal{T}$). We focus here on the case where there is miscoordination, i.e., $\rho \in (\underline{\rho}, \overline{\rho})$ and $\tau \in (0,1)$. Note that the value of a game can in general 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, a type $t$ chooses $s^1$ if $t > \tau$ and $s^2$ 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 $F(t) := F(t, 1)$ is the marginal distribution function of a player's type. Hence, we can rewrite the value of the game as

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

which is Eq. (1) in Proposition 3.1.

Note that, by symmetry,

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

where $f(t) := F'(t)$ is the probability density of a player's type. Hence, using that $F(\tau \mid \tau) = 1 - \rho$, the partial derivative of the value with respect to $\tau$ 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

$$\nabla_{\boldsymbol{u}} V(\boldsymbol{u}; \mathcal{T}) = (u_{12} - u_{21})\, f(\tau)\, \frac{\partial \tau}{\partial \rho}\, \nabla_{\boldsymbol{u}} \rho + \boldsymbol{p}(\tau), \tag{5}$$

where $\boldsymbol{p}(\tau) = (\, p_{11}(\tau), p_{12}(\tau), p_{21}(\tau), p_{22}(\tau)\,)$. Equation (5) shows that, under miscoordination, the change in value in response to one of the payoff parameters $u_{nm}$ can be separated into a direct and an indirect effect: the first term is the indirect effect, and the second term is the direct effect. Using the equilibrium condition $F(\tau \mid \tau) = 1 - \rho$ (under miscoordination), we have

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

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

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

where the first term is again the indirect effect and the second term is the direct effect. The direct effect is always positive; the sign of the indirect effect depends on how $\rho$ changes with payoffs (e.g., $\rho$ increases with $u_{22}$ but decreases with $u_{11}$) and the relative magnitude of $u_{12}$ and $u_{21}$. If the indirect effect is negative, the net effect on the value of a change in one of the payoff parameters $u_{nm}$ depends on the relative magnitude of the direct and indirect effects. The situation becomes more complex when we consider changes in economic primitives that affect multiple payoff parameters. This is the case in Section 3.3, where the introduction of a subsidy affects both $u_{11}$ and $u_{12}$, and a change in discount factor affects $u_{12}$, $u_{21}$, and $u_{22}$, sometimes at different rates. In that case, one needs to consider the total derivative. However, for any given type space, the value and all underlying drivers (e.g., $F(t \mid t)$) can be calculated numerically (as we have done to generate Figures 1–5; see Appendix B for details).

# Appendix B Example type spaces

This appendix shows that the social salience type space (Section 3.2) and the "animal spirits" type space (Section 3.3.1) satisfy our conditions.

## B.1 Social salience

This section defines a type space that naturally fits the experiments discussed in Section 3.2 and shows that it satisfies Assumptions 1–4. We also derive conditions under which the type space satisfies Assumption 5 (NMRB) and the assumption that no action is socially salient.

We assume that impulses are sensitive to social cues: Either action may be "socially salient," where we say that an action is socially salient if players are likely to have an impulse to choose that action. To be precise, we suppose action $s^1$ is socially salient (denoted $\theta = s^1$) with probability $p \in$

$(0,1)$, while $s^2$ is socially salient (denoted $\theta = s^2$) with probability $1-p$. Conditional on action $s$ being socially salient, each player $j$ has an impulse to choose $s$ with probability $q_j$, where $q_j \geq \frac{1}{2}$. 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. The parameter $q_j$ is drawn from a continuous density $g(\cdot)$ with full support on $[\frac{1}{2},1)$ independently across players.[22]

We define an introspective type space as follows: We identify the type of a player who has 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 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 B.1.** *The introspective type space $\mathcal{T}^*$ satisfies the conditions (SYM), (MON-I), (MON-B), and (REG).*

**Proof.** A first observation is that for given $q_j$,

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

Clearly, (SYM) is satisfied. We next show that (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 (MON-B) holds. Consider a player with type $t$ and corresponding parameter $q$. By inverting the relations above between a player's type and the parameter $q$, we

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

find that the parameter $q = q(t)$ that corresponds to type $t$ is

$$q(t) = \begin{cases} r(t) & t \in [0, p); \\ \hat{r}(t) & t \in [p, 1]; \end{cases} \tag{6}$$

where

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

We use this to calculate the density of 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: A 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, for types $t < p$ the density $f(t)$ is $g(q(t))\,|q'(t)|$ times the conditional probability that $I_j = s^2$ given that $q_j = q(t)$, and for types $t > p$ the situation is analogous (here, $q'(t)$ 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 $t \in [0, 1]$,

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

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

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

for all $t \in [0, 1]$. Similarly, the joint probability density[23] of types $t$ and $u$ for the two players is given by

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

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

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

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 $\tau \in [p, 1]$,

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

[23] 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.

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

$$F(t \mid t) = \begin{cases} \displaystyle\int_{q(t)}^{1} (q - t\,(2q-1))\, g(q)\, \mathrm{d}q & t \in [0,p); \\ \displaystyle 1 - \int_{q(t)}^{1} ((1-q) + t\,(2q-1))\, g(q)\, \mathrm{d}q & t \in [p,1]. \end{cases} \tag{7}$$

Since $\lim_{t \downarrow 0} q(t) = \lim_{t \uparrow 1} q(t) = 1$, it follows from this 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 now turn to Assumption 5 (NMRB). Recall that (NMRB) says that there is a $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$, or there is a $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both). In particular, (NMRB) holds if the rank belief function $F(t \mid t)$ is differentiable and has a negative derivative at $t = \tau^0$. We next characterize the conditions under which the introspective type space $\mathcal{T}^*$ has these properties:

**Proposition B.2.** *The rank belief function $F(t \mid t)$ for the introspective type space $\mathcal{T}^* = \mathcal{T}^*(p,g)$ is continuously differentiable on $[0,1]$. Moreover, the derivative at $t = \tau^0$ is negative if and only if $g(\frac{1}{2}) < 8p\,(1-p)\,\mathbb{E}(2q-1)$. Therefore, if this inequality holds, then $\mathcal{T}^*$ satisfies (NMRB).*

**Proof.** 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 from Eq. (7) that for all $t \neq p$,

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

Since $g(\cdot)$ and $q(t)$ are continuous, the derivative is continuous on $[0,p)$ and $(p,1]$. Moreover, 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, $F(t \mid t)$ has a negative derivative at $t = \tau^0$ precisely when $g(\frac{1}{2}) < 8p\,(1-p)\,\mathbb{E}(2q-1)$. □

Finally, a sufficient condition under which no action is strongly salient in $\mathcal{T}^*$ is that the rank belief function has a negative derivative at $t = \tau^0$ and $p$ is equal to $\frac{1}{2}$:

**Proposition B.3.** *Suppose the rank belief function for the introspective type space $\mathcal{T}^* = \mathcal{T}^*(p, g)$ has a negative derivative at $t = \tau^0$. Then no action is strongly salient if $p = \frac{1}{2}$.*

**Proof.** Recall that $\tau^0 = \frac{1}{2}$ for $p = \frac{1}{2}$. By Proposition B.2, when the derivative of the rank belief function is negative at $t = \tau^0$, the interval $(\underline{\rho}, \overline{\rho})$ is nonempty and contains $1 - F(\tau^0 \mid \tau^0)$. Finally, by Eq. (7), $F(\frac{1}{2} \mid \frac{1}{2}) = \frac{1}{2}$. □

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 if $F(t \mid t)$ has a negative derivative at $t = \tau^0$.

The introspective type space $\mathcal{T}^*$ was used to generate Figures 1–3, with the following specifications: $p$ is equal to $\frac{1}{2}$ and $g(\cdot)$ is the truncated normal distribution with mean 7/8 and variance 1/64. By Propositions B.1–B.3, this introspective type space satisfies Assumptions 1–5, 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 Morris and Yildiz's assumptions, the introspective type space in Section 3.3.1 satisfies Assumptions 1–5.

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

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

where $\varepsilon_j$ and $\eta$ 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 $h : \tilde{t} \mapsto t$ 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 joint density stays bounded on $T \times T$. Given an appropriate differentiable transformation $h$, we take the set of 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 $t_1 = h(\tilde{t}_1)$ and $t_2 = h(\tilde{t}_2)$, respectively, then $f(t_1, t_2) = f_{shocks}(\tilde{t}_1, \tilde{t}_2)\, h'(\tilde{t}_1)^{-1}\, h'(\tilde{t}_2)^{-1}$. Because the transformation is continuous, $f$ is a continuous density; by construction, it has full support on the interior of $T \times T$. Now 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 $\mathcal{I}(t) = s^1$ if $t \geq \tau^0$.

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) for appropriate choices of the transformation $h$; moreover, if $\tau^0$ 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 $\eta$ has a Student's-t distribution with parameter $n = 4$ while the idiosyncratic shocks have a standard normal distribution.[24] For the transformation, it will be convenient to define a mapping from $[0,1]$ to $[-\infty, \infty]$ and then use its inverse to calculate the rank belief function. 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 $\alpha$ controls the shape of the transformation; we use $\alpha = 1/4$. To avoid numerical problems, we use a slight modification of this transformation to generate our figures: We replace the first map by $t \mapsto (1 - (\alpha R)^{-1})\,(2t - 1)$ so that the extreme types 0 and 1 get mapped to values close to $\pm R$ and $R$ thus acts as a cutoff on extreme types; we use $R = 200$.

# Appendix C Proofs

## C.1 Proof of Proposition 2.1

We start by proving existence and uniqueness. Say that a strategy $\sigma_j$ is a *switching strategy* with threshold $t^* \in T$ if types $t \in T$ with $t < t^*$ choose $s^2$ (i.e., $\sigma_j(t) = s^2$), and types $t \in T$ with $t > t^*$ choose $s^1$ (i.e., $\sigma_j(t) = s^1$). (Type $t^*$ may choose either action.) At level 0, types follow their impulse. By Assumptions (MON-I) and (SYM), the level-0 strategy $\sigma_j^0$ for each player $j$ is

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

a switching strategy with (common) threshold $\tau^0$. Suppose that, at level 1, type $\tau^0$ 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 $\tau^1$ be the largest type not larger than $\tau^0$ 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 $\tau^1$: By (MON-B) and (SYM), action $s^1$ is a strict best response for 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 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 $\tau$.

For $k > 1$, suppose, inductively, that for each player the level-$(k-1)$ strategy is a switching strategy with threshold $\tau^{k-1}$, and that, furthermore, $F(t \mid t) < 1 - \rho$ for all $t \in [\tau^{k-1}, \tau^0]$. Define $\tau^k$ to be the largest type not larger than $\tau^{k-1}$ 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 $\tau^k$, 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 $\tau$ not larger than $\tau^0$ 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 $\tau^0$, the equilibrium threshold $\tau = \lim_{k \to \infty} \tau^k$ is the smallest $\tau$ not smaller than $\tau^0$ such that $F(\tau \mid \tau) \leq 1 - \rho$ if such a type exists, and $\tau = 1$ otherwise. Finally, if type $\tau^0$ 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 $\tau$ is just $\tau^0$. So, in introspective equilibrium, each player follows a switching strategy with threshold $\tau$. As a consequence, introspective equilibrium is monotone in type. The equilibrium is essentially unique: It pins down the behavior for all types $t \neq \tau$, and this set has probability 1.

We next show that introspective equilibrium is monotone in payoffs. 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, considering how the equilibrium threshold varies with the dominance parameter allows us to assess how improving the payoffs to an action changes the probability that players choose that action in introspective equilibrium. Fix an introspective type space and dominance parameters $\rho, \tilde{\rho}$ such that $\tilde{\rho} > \rho$. Denote the games with dominance parameters $\rho$ and $\tilde{\rho}$ by $\mathcal{G}$ and $\widetilde{\mathcal{G}}$, respectively, and let $\tau$ 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 argument showing existence, 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 existence proof, either $\tau = 0 \leq \tilde{\tau}$ or otherwise

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

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 existence proof shows that either $\tilde{\tau} = 1 \geq \tau$ or

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

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

## C.2 Proof of Lemma 2.2

We prove the necessity of (NMRB). Suppose that (NMRB) does not hold. Then, 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 $t > \tau^0$. But then, by the proof of Proposition 2.1, the equilibrium threshold is $\tau = 0$ for $\rho < \rho^0$ and $\tau = 1$ for $\rho > \rho^0$. Thus, we have $\tau \in (0,1)$ only if $\rho = \rho^0$. The sufficiency of (NMRB) follows from the proof of Proposition 3.1 below. □

## C.3 Proof of Proposition 3.1

Fix a type space $\mathcal{T} = (F, \tau^0)$ that satisfies Assumptions 1–4. Note that by (REG), the rank belief function $F(t \mid t)$ is well-defined (and continuous) for all $t \in (0,1)$, and can be extended to a continuous function on $[0,1]$ by defining $F(0 \mid 0)$ and $F(1 \mid 1)$ as the limits of $F(t \mid t)$ as $t$ tends to 0 and 1, respectively. We then define $\underline{\rho}, \overline{\rho}$ by

$$1 - \underline{\rho} = \max\{F(t \mid t) : t \in [0, \tau^0]\};$$
$$1 - \overline{\rho} = \min\{F(t \mid t) : t \in [\tau^0, 1]\};$$

and let

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

be the types "closest" to $\tau^0$ whose rank beliefs attain the relevant extrema; see Figure 2 for an illustration. We will first show in Lemma C.2 below that, for any type space, we have $\underline{\rho} > 0$ and $\overline{\rho} < 1$, and that we cannot have both $\underline{\tau} = 0$ and $\overline{\tau} = 1$, i.e., at least one of them must lie in the interior of $T$. This will be central to proving Proposition 3.1(a)–(b). Lemma C.2 relies on the following auxiliary result:

**Lemma C.1.** *We have* $\lim_{t \downarrow 0} F(t \mid t) \leq 1/2$ *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 $t \in (0, \delta)$,

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

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 C.2.** *It is the case that $0 < \underline{\rho} \leq \overline{\rho} < 1$, where the middle inequality is strict 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 of $\underline{\rho}$ and $\overline{\rho}$ that $\underline{\rho} \leq \overline{\rho}$ and that we have the strict inequality $\underline{\rho} < \overline{\rho}$ if and only if there is a $t < \tau^0$ such that $F(t \mid t) > F(\tau^0 \mid \tau^0)$ or there is a $t > \tau^0$ such that $F(t \mid t) < F(\tau^0 \mid \tau^0)$ (or both). But the latter statement 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 attains a maximum on $[0, \tau^0]$ and a minimum on $[\tau^0, 1]$) and $F(t \mid t) \in (0,1)$ for all $t \in (0,1)$. It thus remains to show that $\sup\{F(t \mid t) : t \in [0, \tau^0]\} < 1$ and $\inf\{F(t \mid t) : t \in [\tau^0, 1]\} > 0$. But this follows from Lemma C.1.

We next prove the second claim. By Lemma C.1 and the definition of $\underline{\tau}$ we have 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 Proposition 3.1. By the proof of Proposition 2.1, 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 must show that the value is generically not equal to the expected payoff in one of the Nash equilibria. By Lemma C.2, we have $\underline{\rho} < \overline{\rho}$ if and only if (NMRB) holds. Moreover, for $\rho \in (\underline{\rho}, \overline{\rho})$, the equilibrium threshold $\tau$ 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 $p_{11}^{\text{MNE}} + p_{22}^{\text{MNE}}$ 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 $\tau$ 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 2.1), we have

$$\begin{aligned} p_{11}(\tau) &= \int_{\tau}^{1} (1 - F(\tau \mid t))\, \mathrm{d}F(t) > \rho\,(1 - F(\tau)) = \rho\,(p_{11}(\tau) + p_{12}(\tau)); \\ p_{22}(\tau) &= \int_{0}^{\tau} F(\tau \mid t)\, \mathrm{d}F(t) > (1-\rho)\, F(\tau) = (1-\rho)\,(p_{22}(\tau) + p_{21}(\tau)); \end{aligned} \tag{8}$$

where the inequalities follow from (MON-B) and (REG). Rearranging the terms, it follows that

$$\begin{aligned} (1-\rho)(p_{11}(\tau) + p_{12}(\tau)) &> p_{12}(\tau); \\ \rho\,(p_{22}(\tau) + p_{21}(\tau)) &> p_{21}(\tau). \end{aligned} \tag{9}$$

Using that $p_{12}(\tau) = p_{21}(\tau)$ (by (SYM)), we thus have

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

Hence, $p_{12}(\tau) + p_{21}(\tau) < 2\rho\,(1-\rho)$, or, equivalently,

$$p_{11}(\tau) + p_{22}(\tau) > 1 - 2\rho\,(1-\rho). \tag{10}$$

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), we express the value in introspective equilibrium as

$$V = p_{11}(\tau)\, u_{11} + p_{12}(\tau)\, u_{12} + p_{21}(\tau)\, u_{21} + p_{22}(\tau)\, u_{22}$$

and note that there exists an $\alpha_\rho > 0$ (dependent on $\rho$) and a $\beta_\rho$ (which could be positive or negative) such that

$$\begin{aligned} p_{12}(\tau) &= p_{21}(\tau) = \rho\,(1-\rho)(1-\alpha_\rho); \\ p_{11}(\tau) &= \rho^2 + \rho\,(1-\rho)(\alpha_\rho - \beta_\rho); \\ p_{22}(\tau) &= (1-\rho)^2 + \rho\,(1-\rho)(\alpha_\rho + \beta_\rho). \end{aligned}$$

It follows that the difference between the expected payoff $V_{\text{MNE}}$ in mixed Nash equilibrium and the value $V$ in introspective equilibrium is

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

Hence, we have that $V_{\text{MNE}} = V$ 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 $\rho$ (and thus $\alpha_\rho$ and $\beta_\rho$), 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.

Finally, Eq. (1) for the value is derived in Appendix A. This completes the proof. □

**Remark C.1.** *The proof of Proposition 3.1 can be simplified if we strengthen (REG) to require that the density $f$ has full support on the whole of $T \times T$ (not just its interior): With this stronger assumption, it follows directly that $F(t \mid t)$ tends to 0 and 1 as $t$ approaches 0 and 1, respectively (which is obviously stronger than Lemma C.1). 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) with $g(1) = 0$ or the animal spirits type space (Section 3.3.1).*

## C.4 Proof of Proposition 3.2

Part (a) follows directly from Proposition 3.1 by noting that $\rho = 1/(w+1)$. To prove (b), 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

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

(see Appendix A). 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 from $V = p_{11}(\tau) + p_{22}(\tau) > 1 - 2\rho\,(1-\rho) = \frac{1}{2}$ (by Eq. (10)). □

## C.5 Proof of Proposition 3.3

Let $\mathcal{T}$ be an introspective type space with $\frac{1}{2} \in (\underline{\rho}, \overline{\rho})$. For $w > 1$ and $x \in [0, w-1)$, write $\rho(w,x)$ for the dominance parameter of $\tilde{\boldsymbol{u}}_x$. Then, there is a $\underline{w}$ such that for $w > \underline{w}$, $\rho(w,0) < \underline{\rho}$ and thus, by Proposition 3.1, $V((w,-c,0,1);\mathcal{T}) = w$. 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 a $\underline{w}'$ such that for $w > \underline{w}'$, $\lim_{x \uparrow w-1} \rho(w,x) \in (\underline{\rho}, \overline{\rho})$. Since the value under miscoordination is strictly smaller than $w$, the result then follows by choosing $w^* = \max\{\underline{w}, \underline{w}'\}$. □

## C.6 Proof of Theorem 3.4

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 $\rho^*$ 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 Eq. (9) in the proof of Proposition 3.1 (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,

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

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

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

Using Eq. (11) and that $\overline{p_{22}} = 1 - \overline{p_{12}} - \overline{p_{21}} - \overline{p_{11}}$ and $\overline{p_{21}} = \overline{p_{12}}$, we can rewrite this as follows:

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

Using the definition of $\rho^*$, 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 and $u_{22} > u_{12}$, $\Delta$ can be negative only if $\rho < \rho^*$. In that case, $\Delta < 0$ is equivalent to

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

where the equality follows from Eq. (11). □

## C.7 Proof of Theorem 3.5

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 C.3.** *Fix an introspective type space that satisfies Assumptions 1–5 and let $\rho, \rho'$ be such that $\underline{\rho} < \rho' < \overline{\rho} < \rho$, i.e., there is coordination failure in introspective equilibrium for the game with dominance parameter $\rho$ and there is miscoordination for the game with dominance parameter $\rho'$. Let $(u_{11}, u_{12}, u_{21}, u_{22})$ be a game form with dominance parameter $\rho > \overline{\rho}$, and $(u'_{11}, u'_{12}, u'_{21}, u'_{22})$ a game form with dominance parameter $\rho' \in (\underline{\rho}, \overline{\rho})$. 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 $\rho'$. Then, the difference $\Delta$ in value between the games with dominance parameters $\rho'$ (with miscoordination) and $\rho$ (with coordination failure) satisfies*

$$\Delta > (p'_{11} + p'_{12})(u'_{21} - u'_{22}) - (u_{22} - u'_{22}). \tag{13}$$

**Proof.** From

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

it follows that

$$\rho' (u'_{11} - u'_{22}) = u'_{22} - u'_{12} + \rho' (u'_{12} + u'_{21} - 2u'_{22}). \tag{14}$$

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 $\rho'$ (with miscoordination) and $\rho$ (with coordination failure) is

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

Using Eq. (14) we thus obtain

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

Combining this with $p'_{11} > \rho' (p'_{11} + p'_{12})$ (Eq. (8) in the proof of Proposition 3.1) gives (13). □

We also need the following lemma about the per-period profits:

**Lemma C.4.** *The various per-period profits in our model (i.e., the cheating, collusive, mutual cheating, Bertrand–Nash, and victim profit) satisfy $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$.*

Assuming for now that the lemma holds, we next derive the payoffs for the repeated game. Let $\delta \in (0,1)$ be the common discount factor. If firm $i$ chooses strategy $\sigma_i \in \{\sigma^*, \sigma^c\}$ and the

other firm chooses strategy $\sigma_{-i} \in \{\sigma^*, \sigma^c\}$, then 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 $(\sigma_i, \sigma_{-i})$. This yields the payoff matrix in the main text. By Lemma C.4 it satisfies $u_{22} > u_{12}$ and $u_{11} > u_{22}$. It also satisfies $u_{11} > u_{21}$ if $\delta > (\pi^c - \pi^*)/(\pi^c - \pi^N)$. So under this assumption on $\delta$, this is a coordination game, i.e., $\rho \in (0,1)$. Furthermore, from the payoff matrix we see that the dominance parameter $\rho = \rho(\delta)$ decreases to 0 as $\delta$ increases to 1.

We are now ready to prove Theorem 3.5, building on Lemmas C.3 and C.4. Let $\delta, \delta'$ be as in the statement of Theorem 3.5, and write $\rho = \rho(\delta)$ and $\rho' = \rho(\delta')$ for the corresponding dominance parameters. Then $\delta' > \delta$ and $\rho' < \overline{\rho} < \rho$. Now, if $\delta'$ is so large that $\rho' < \underline{\rho}$ (i.e., all players play the collusive strategy $\sigma^*$ in introspective equilibrium), then it follows immediately from $\pi^* > \pi^m > \pi^N$ that the game with discount parameter $\delta'$ has a strictly larger value than the game with discount parameter $\delta$. So it remains to consider the case $\rho' \in [\,\underline{\rho}, \overline{\rho}\,)$. Let $\varepsilon := \delta' - \delta$ be the difference between the two discount parameters. Then $\delta' < \overline{\delta} + \varepsilon$, where $\overline{\delta}$ denotes the discount parameter such that $\rho(\overline{\delta}) = \overline{\rho}$. Let $p_{nm}(\tau)$ be the probability with which the action profile $(s^n, s^m)$ is played in introspective equilibrium, and write $\overline{p_{nm}} := p_{nm}(\overline{\tau})$; observe that $p'_{11} + p'_{12} > \overline{p_{11}} + \overline{p_{12}}$ because $\rho' < \overline{\rho}$ (Appendix A). Hence, Eq. (13) from Lemma C.3 gives

$$\begin{aligned}\Delta &> (p'_{11} + p'_{12})(1-\delta')(\pi^c - \pi^m) - (\delta' - \delta)(\pi^m - \pi^N)\\ &> (\overline{p_{11}} + \overline{p_{12}})(1-\overline{\delta})(\pi^c - \pi^m) - \varepsilon\,(\overline{p_{11}} + \overline{p_{12}})(\pi^c - \pi^m) - \varepsilon\,(\pi^m - \pi^N).\end{aligned}$$

Since the first term on the right is positive, it follows that $\Delta > 0$ if the difference $\varepsilon$ between the two discount parameters is sufficiently small. This is what we had to prove, so to complete the proof of Theorem 3.5, it only remains to prove Lemma C.4.

**Proof of Lemma C.4.** Recall that $r = c/b$. Note that the two inverse demand functions

$$p_i = 1 - q_i - r\,q_{-i} \qquad (0 < r < 1) \tag{15}$$

$$\tilde{p}_i = a - b\,\tilde{q}_i - c\,\tilde{q}_{-i} \qquad (a > 0, b > c > 0) \tag{16}$$

are equivalent in the following sense: If we use the transformations $\tilde{p}_i = a\,p_i$ and $\tilde{q}_i = a\,q_i/b$, then we obtain a one-to-one correspondence between the solution of the system of equations defined by (15) and the solution of the system of equations defined by (16). Moreover, the corresponding per-period profits $\pi_i = q_i\,p_i$ and $\tilde{\pi}_i = \tilde{q}_i\,\tilde{p}_i$ differ only by a factor $a^2/b$. So, it suffices to consider (15). It will also be convenient to consider the rescaled profit $\psi_i := (1-r^2)\,\pi_i$ instead of $\pi_i$. Solving (15) for $q_i$ gives

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

Hence, as a function of a firm's price $p_i$, the firm's rescaled profit given that the competitor charges $p_{-i}$ is given by

$$\psi_i(p_i \mid p_{-i}) = (1 - r + r p_{-i} - p_i)\, p_i.$$

In particular, if both firms charge the same price $p$, this simplifies to a rescaled profit of

$$\psi(p) = (1 - r)(1 - p)\, p.$$

The Bertrand–Nash price $p^N$ is the value for $p$ that maximizes $\psi_i(p \mid p^N)$. Hence, it satisfies $p^N = \frac{1}{2}(1 - r + r\, p^N)$, which yields $p^N = (1 - r)/(2 - r)$. The collusive price $p^*$ maximizes $\psi(p)$, hence $p^* = \frac{1}{2}$. The cheating price $p^c$ depends on whether the constraint that the victim's share $q_{-i}$ (and hence also $\psi_{-i}$) is nonnegative is binding: If the constraint is not binding, then $p^c$ is the value for $p$ that maximizes $\psi_i(p \mid p^*)$, which yields $p^c = \frac{1}{2} - \frac{1}{4}r$. The victim's rescaled profit is then given by $\psi_{-i}(p^* \mid p^c) = \frac{1}{8}(2 - 2r - r^2)$, from which it is readily seen that the constraint $q_{-i} \geq 0$ is not binding if $r < \sqrt{3} - 1$, and binds if $r \geq \sqrt{3} - 1$. In the latter case, we must have $q_{-i} = 0$ when $p_i$ equals $p^c$ and $p_{-i}$ equals $p^* = \frac{1}{2}$. This yields $p^c = 1 - 1/(2r)$. In either case, the rescaled cheating, collusive, mutual cheating, Bertrand–Nash, and victim profit are given in terms of $p^N$, $p^*$ and $p^c$ by

$$\psi^c = \psi_i(p^c \mid p^*), \quad \psi^* = \psi_i(p^* \mid p^*) = \psi(p^*), \quad \psi^m = \psi(p^c), \quad \psi^N = \psi(p^N), \quad \psi^v = \psi_i(p^* \mid p^c).$$

Recall that $p^c$ is the value for $p$ that maximizes the parabola $\psi_i(p \mid p^*)$, while $p = p^*$ maximizes the parabola $\psi(p)$. Hence, to prove that $\psi^c > \psi^* > \psi^m > \psi^N$, it suffices to show that $p^* > p^c > p^N$. That the cheating price lies below $p^* = \frac{1}{2}$ is confirmed by the formulas $p^c = \frac{1}{2} - \frac{1}{4}r$ for $r < \sqrt{3} - 1$ and $p^c = 1 - 1/(2r)$ for $r \geq \sqrt{3} - 1$. Since $(2 - r)\, p^N = 1 - r$, we also have

$$\begin{aligned} (2 - r)(p^c - p^N) &= \tfrac{1}{4}r^2 > 0 && \text{for } r < \sqrt{3} - 1; \\ 2r\,(2 - r)(p^c - p^N) &= 3r - 2 > 0 && \text{for } r \geq \sqrt{3} - 1. \end{aligned}$$

This proves that $p^c > p^N$. So $\psi^c > \psi^* > \psi^m > \psi^N$, and since $\psi^N > \psi^v = 0$ for $r \geq \sqrt{3} - 1$, it only remains to show that $\psi^N > \psi^v$ when $r < \sqrt{3} - 1$ and $p^c = \frac{1}{2} - \frac{1}{4}r$. To this end, observe that

$$\psi^v = \psi_i(p^* \mid p^c) = \psi(p^*) - r\,(p^* - p^c)\, p^* = \tfrac{1}{4}(1 - r) - \tfrac{1}{8}r^2,$$

while $\psi^N = \psi(p^N) = (1 - r)^2/(2 - r)^2$. Since $(2 - r)^2 = 4\,(1 - r) + r^2$, it follows that

$$(2 - r)^2\,(\psi^N - \psi^v) = \tfrac{1}{4}r^2\,(1 - r) + \tfrac{1}{8}r^4 > 0.$$

Hence, $\psi^N > \psi^v$, and we conclude that $\pi^c > \pi^* > \pi^m > \pi^N > \pi^v$ for all $r \in (0, 1)$. □

## C.8 Proof of Proposition 4.1

Fix a game form $\boldsymbol{u}$ (with corresponding dominance parameter $\rho$) and fix the introspective type space $\mathcal{T}_0 = (F, \tau_0^0)$ at time 0. Recall that, by (REG), the rank belief function $F(t \mid t)$ is continuous in $t$. 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 $\tau_0$ 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 $\rho$ such that $F(\tau_0^0 \mid \tau_0^0) = 1-\rho$ or $1-\rho$ is a local extremum of $F(t \mid t)$, proving the claim for this case establishes the result for generic $\boldsymbol{u}$. Since $F(\tau_0 \mid \tau_0) = 1-\rho$ and, by assumption, $F(\tau_0^0 \mid \tau_0^0) \neq 1-\rho$, we have $\tau_0 \neq \tau_0^0$. We prove the result for the case $\tau_0^0 > \tau_0$; the proof for the case $\tau_0^0 < \tau_0$ is similar and thus omitted.

Fix $\chi > 0$. We claim that the following holds:

**Lemma C.5.** *There exist an equilibrium threshold $\hat{\tau}$ satisfying $\tau_0 - \chi < \hat{\tau} \leq \tau_0$ and an $\varepsilon > 0$, such that for every level-0 threshold $\tau^0$ in the interval $(\hat{\tau} - \varepsilon, \tau_0^0 + \varepsilon)$, the introspective process $\{\tau^k\}_k$ starting from $\tau^0$ converges to an equilibrium threshold $\tau$ that lies in the interval $[\hat{\tau}, \tau_0]$.*

Assume for the moment that this is true, and fix a dynamic $\{\tau_{\tilde{t}}^0\}_{\tilde{t}}$ that satisfies Eq. (4) for $\varepsilon$ as in the lemma. For $\tilde{t} \geq 0$, let $\tau_{\tilde{t}}$ be the introspective equilibrium for the game $\mathcal{G}_{\tilde{t}} = (\boldsymbol{u}, \mathcal{T}_{\tilde{t}})$, where $\mathcal{T}_{\tilde{t}} = (F, \tau_{\tilde{t}}^0)$. Then by Lemma C.5, using induction in $\tilde{t}$, we have that for every $\tilde{t} \geq 0$, the level-0 threshold $\tau_{\tilde{t}}^0$ lies in the interval $(\hat{\tau} - \varepsilon, \tau_0^0 + \varepsilon)$ and therefore $\tau_{\tilde{t}}$ lies in $[\hat{\tau}, \tau_0]$. Since $\tau_0 - \hat{\tau} < \chi$, this implies that $|\tau_{\tilde{t}} - \tau_{\tilde{t}'}| < \chi$ for every pair of periods $\tilde{t}, \tilde{t}'$. It therefore only remains to prove Lemma C.5.

**Proof of Lemma C.5.** Recall that $\tau_0$ is the supremum of all types $t$ in the interval $[0, \tau_0^0]$ such that $F(t \mid t) \geq 1-\rho$ or $t = 0$ (proof of Proposition 2.1). So by the continuity of $F(t \mid t)$, this means that there exists an $\bar{\varepsilon} > 0$ such that $F(t \mid t) < 1-\rho$ for all $t \in (\tau_0, \tau_0^0 + \bar{\varepsilon})$. But then the introspective process $\{\tau^k\}_k$ starting from any level-0 threshold $\tau^0$ in the interval $[\tau_0, \tau_0^0 + \bar{\varepsilon})$ converges to $\tau_0$. So if $\tau_0 = 0$, we can simply take $\hat{\tau} = \tau_0$ and $\varepsilon = \bar{\varepsilon}$, and the result follows. It remains to consider the case $\tau_0 > 0$. Since, by assumption, $F(\tau_0 \mid \tau_0) = 1-\rho$ is not a local extremum of the rank belief function and $F(t \mid t) < 1-\rho$ on the interval $(\tau_0, \tau_0^0 + \bar{\varepsilon})$, there must be a $\hat{t}$ in the interval $(\tau_0 - \chi, \tau_0)$ such that $F(\hat{t} \mid \hat{t}) > 1-\rho$. But then the introspective process $\{\tau^k\}_k$ starting from the level-0 threshold $\tau^0 = \hat{t}$ converges to an equilibrium threshold $\hat{\tau}$ that satisfies $\hat{t} < \hat{\tau} \leq \tau_0$. To be precise, $\hat{\tau}$ is the infimum over all types $t \geq \hat{t}$ for which $F(t \mid t) \leq 1-\rho$, so $F(t \mid t)$ must be strictly larger than $1-\rho$ on the interval $(\hat{t}, \hat{\tau})$. Now define $\underline{\varepsilon} := \hat{\tau} - \hat{t}$. Then the introspective process $\{\tau^k\}_k$ starting from any level-0 threshold $\tau^0$ in $(\hat{\tau} - \underline{\varepsilon}, \hat{\tau}]$ converges to $\hat{\tau}$. If we now take $\varepsilon := \min\{\underline{\varepsilon}, \bar{\varepsilon}\}$, then it follows that for every level-0 threshold $\tau^0$ in $(\hat{\tau} - \varepsilon, \tau_0^0 + \varepsilon)$, the introspective process starting from $\tau^0$ converges to an equilibrium threshold $\tau$ in the interval $[\hat{\tau}, \tau_0]$. □

## C.9 Proof of Proposition 5.1

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 1–5. Let $\hat{\boldsymbol{u}} = \langle N, \{S_j\}_{j\in N}, \{u_j\}_{j\in N}\rangle$ be a finite game form, where $N$ is the (finite) player set and for each player $j \in N$, $S_j$ is the (finite) set of actions and $u_j\colon S_j \times S_{-j} \to \mathbb{R}$ is the payoff function. Fix an introspective type space, that is, a set $T_j$ of 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 Borel $\sigma$-algebra $\mathscr{B}(T_j)$ on $T_j$. For each player $j \in N$, let $\Sigma_j$ be the set of (pure) *strategies*, i.e., measurable functions $\sigma_j\colon T_j \to S_j$. 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 $F$.

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

**Lemma C.6.** *Let $j \in N$. Then, for every $k$, $\sigma_j^k$ 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 $\psi_j$, 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 $\sigma_{-j}$. Given that the level-0 strategies are measurable for all players, it then follows that for each player $j \in N$, $\sigma_j^1$ is measurable. Iterating this argument gives that $\sigma_j^k$ 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 type $t_j \in T_j$ into its interim expected payoff, i.e.,

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

is measurable (e.g., Aliprantis and Border, 2006, Thm. 15.13). Let $\varphi_j(\cdot, \sigma_{-j})\colon T_j \twoheadrightarrow S_j$ be the best-response correspondence (given $\sigma_{-j}$), 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 $\psi_j$ that maps each nonempty subset $B_j \subset S_j$ into an element $s_j$ of $B_j$. Then, $\psi_j \circ \varphi_j(\cdot, \sigma_{-j})\colon T_j \to S_j$ is measurable. This proves the claim.

Hence, for every player $j \in N$, tie-breaking rule $\psi_j$, and $k > 0$, the level-$k$ strategy $\sigma_j^k$, defined by $\sigma_j^k(t_j) = \psi_j \circ \varphi_j(t_j, \sigma_{-j}^{k-1})$ 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 $j \in N$, $\sigma_j$ 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 each $t_j$ in $T_j$,

$$\int u_j(\sigma_j(t_j), \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) \geq \int u_j(s_j, \sigma_{-j}(t_{-j}))\,\mathrm{d}F(t_{-j} \mid t_j) \tag{17}$$

for $s_j \in S_j$. By Lemma C.6, the integrals in Eq. (17) 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. □

# References


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.

Alós-Ferrer, C. and C. Kuzmics (2013). Hidden symmetries and focal points. *Journal of Economic Theory 148*(1), 226–258.

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.

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

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. 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 64*(4), 943–956.

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.

Brandenburger, A. and E. Dekel (1987). Rationalizability and correlated equilibria. *Econometrica 55*, 1391–1402.

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 $(2 \times 2)$ 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.

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.

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.

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

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.

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.

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 62*(1), 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*, Chapter 3, pp. 56–114. 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.

Myerson, R. B. (1994). Communication, correlated equilibria and incentive compatibility. In R. Aumann and S. Hart (Eds.), *Handbook of Game Theory*, Volume 2, pp. 827–847. Elsevier.

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

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

Penta, A. and P. Zuazo-Garin (2022). 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, pp. 235–258. 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.

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*. Houghton Mifflin.

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.

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.

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. (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.