What situation is this?

Coarse cognition and behavior over a space of games[*]

Robert Gibbons[a], Marco LiCalzi[b], and Massimo Warglien[c]

September 2017

Abstract. We study strategic interaction between agents who distill the complex world into simpler situations. We assume agents share the same cognitive frame. In equilibrium, agents can be better or worse off than under full information: the frame creates a fog of cooperation or a fog of conflict. In repeated interaction, the frame is as important as agents' patience: for a fixed discount factor, when all agents coordinate on what they perceive as the best equilibrium, there remain significant performance differences across dyads with different frames. Finally, we analyze tensions between incremental versus radical changes in the cognitive frame.

Keywords: categorization, frame, mental model, small world, culture, leadership.

JEL Classification Numbers: C79, D01, D23, L14, M14.

Correspondence to:

Robert Gibbons	MIT Sloan School and Economics Department 100 Main Street, E62-519 Cambridge, MA 02142-1347, U.S.A.
E-mail:	rgibbons@mit.edu

[*] We thank Andreas Blume, Vincent Crawford, Emir Kamenica, Kate Kellogg, Margaret Meyer, Wanda Orlikowski, Alessandro Pavan, Andrea Prat, Phil Reny, Joel Sobel, Marco Tolotti, Cat Turco, and seminar audiences at Autonoma de Barcelona, Carnegie Mellon, Columbia, ESSET 2017, MIT, MPI Leipzig, London Business School, Padua, Pompeu Fabra, Siena, Stanford GSB, Vanderbilt, and Washington U. for helpful comments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732942 and from MIT Sloan's Program on Innovation in Markets and Organizations.

[a] Massachusetts Institute of Technology, rgibbons@mit.edu

[b] Università Ca' Foscari Venezia, licalzi@unive.it

[c] Università Ca' Foscari Venezia, warglien@unive.it

1 Introduction

An agent must apprehend her world before she can make decisions. Her perception generates a representation of the environment—"[a] 'small-scale model' of external reality" used to formulate and evaluate her options (Craik, 1943: 61).

Since Savage (1954), economics has recognized that distilling the "grand world" into a "small world" precedes rational decision-making. It is frequently assumed that the small world is a parsimonious but accurate model, exogenously given. Alternatively, the small world can result from a deliberate choice made by an agent (e.g., by allocating attention among different variables) or by a third party (e.g., by designing how much information is provided). However, Savage warned that the deliberate choice of an appropriate small world is a difficult task—"a matter of judgment and experience about which it is impossible to enunciate complete and sharply defined general principles" (1954: 16).

Cognitive science goes further than economics on this question: it is widely agreed that agents distill the environment into partial representations and that agents' mental models depend on cognitive mechanisms that usually escape conscious control. For example, Allport (1954: 20) stated that "the human mind must think with the aid of categories. Once formed, categories are the basis for normal prejudgment. We cannot possibly avoid this process." In short, in the cognitive approach, the small world cannot be attributed to a deliberate choice and need not be accurate.

In this paper, we develop and analyze a model of how agents' partial representations affect their strategic behavior and hence their performance. Motivated by the cognitive approach, our main goal is to explore the role of "small-scale" mental models in strategic interactions. Our model can also be interpreted in terms of information design and thus nests a version of the economic approach as well.

We study two agents who engage in strategic interaction using a shared mental model. The environment is an uncountable space $\mathcal{G}$ of games. Each agent distills $\mathcal{G}$ into a finite number of categories called *situations*. The collection of situations is a partition of $\mathcal{G}$, which we call the agent's *frame*.[1] The frame-based cognition of $\mathcal{G}$ is

[1] We borrow this term from Bacharach (2003: 63), who defines it as "the set of *concepts* an agent uses in thinking about the world" and assumes that a frame induces a partition. Less formally, Goffman (1974: 21) uses this term to connote "schemata of interpretation" that allow agents to "locate, perceive, identify, and label" events in the world around them.

coarse, because the frame reduces an uncountable set to a finite partition.

Under our economic approach, the frame is exogenously given or follows from known choices: when a game from $\mathcal{G}$ is realized, an agent learns only which situation has occurred and updates her belief accordingly. We call this case *generative* because the agent's frame is generated by an information structure known to the agent. The generative case is consistent with the "metaphorical" interpretation of information design (Bergemann and Morris, 2017: 4).

Under our cognitive approach, in contrast, agents have no access to the cognitive process that produces their mental representation: they are unaware that they are framing, and they cannot imagine that others perceive the world differently. More precisely, they know only (i) the set of situations that could be realized and (ii) which one has been; they are not aware of the underlying games in $\mathcal{G}$. We call this case *interpretive* because the frame summarizes how agents interpret their environment.[2]

We assume that both parties share the same frame. In the generative case, this occurs when the information structure is symmetric; in the interpretive case, the shared frame may be culturally determined, as we discuss below. We study how the frame affects strategic behavior, under minimal requirements of rationality: conditional on what their frame lets them perceive, agents have correct beliefs and play their dominant strategy. Our analysis thus applies to both our generative and interpretive cases.

We provide three main results. First, in a one-shot interaction, the coarse representation induced by a frame can either decrease or increase the parties' payoffs, compared to having full information about the environment. We say that the shared frame may induce either a *fog of conflict* or a *fog of cooperation.* In the generative case, this implies that limiting access to information can make both agents better (or worse) off. In the interpretive case, it suggests that performance differences may be due to differences in cognitive frames across dyads.

Second, we consider a repeated interaction. In each period an independent draw from $\mathcal{G}$ selects (a) the game the parties actually face and (b) conditional on their shared frame, which situation they perceive. Assuming an infinite horizon and subgame perfection, standard arguments allow the parties to increase their payoffs above the

[2] Our terminology is similar to Hong and Page (2009: 2175), who suggest that "generated signals [...] are passively received by the agents, [whereas to] create an interpreted signal, an agent filters reality into a set of categories."

static level if they are sufficiently patient. We focus on the opposite comparative static: fix the parties' discount rate δ and analyze how their frame affects payoffs in a repeated interaction. We find that, for fixed δ, there are frames under which the parties' highest equilibrium payoffs are greater (or lower) than in a repeated interaction under full information. In this respect, the configuration of the situations induced by a frame, called the *frame's footprint*, is as important as discounting in facilitating cooperation.

Third, moving beyond comparative statics, we investigate the dynamics of changes in the shared frame within a dyad. We distinguish between *incremental change*, when only the footprint of the situations perceived by the parties varies, but not the dominant action in a given situation, versus *radical change*, where both the footprint and the action vary. When the cost of reshaping the footprint is increasing in the size of the change, the optimal change in frame may be incremental and lead to lower gross payoffs than a radical (but expensive) change would. When changes in the frame are costless but of uncertain effectiveness, we illustrate how behavioral inertia after a radical change may have perverse effects on short-term performance.

Framing this paper

We take two inspirations from economics and one from cognitive science. First, at the individual level, we follow Simon (1986), Kreps (1990a), and Rubinstein (1991) by separating cognition from behavior. Simon (1986: S211) cautions that "we must distinguish between the real world and the actor's perception of it and reasoning about it." Kreps (1990a: 155) explicitly separates cognition and (rational) behavior: "the individual builds a model of his choice problem, [. . .] which is typically a simplification or a misspecification (or both) of the 'true situation'. The individual finds his 'optimal' choice of action within the framework of this model and acts accordingly." Finally, Rubinstein (1991) calls for game-theoretic models to account for what agents perceive.

Second, at the group level, we follow Denzau and North (1994), Aoki (2001), and Ostrom (2005) by emphasizing the shared mental models that can be held by individuals with common backgrounds or experiences. Denzau and North (1994: 5) argue that the experiences of past generations are distilled into "a culturally provided set of categories and priors." Aoki (2001: 235) discusses how shifts in equilibria are associated with changes in the parties' "common cognitive representations," and Ostrom (2005:

106) emphasizes that "cultural beliefs systems affects the mental models that individuals utilize." More recently, Hoff and Stiglitz call for economic analyses to consider "socially constructed cognitive frames" (2010: 141) and "cultural mental models [such as] concepts, categories, social identities, [and] narratives" (2016: 26).

Finally, from cognitive science, we follow a wide consensus that subjective representations mediate between perception and choice: "we have mental structures, especially schematic representations of complex social phenomena, which shape the way we attend to, interpret, remember, and respond emotionally to the information we encounter and possess"' (DiMaggio, 1997: 273). We focus on categories as a basic form of such cognitive simplifications: "categorization is the mental operation by which the brain classifies objects and events [and] this operation is the basis for the construction of our knowledge of the world" (Cohen and Lefebvre, 2005: 2).

Related literature

Categorization is not new to the economics literature. For example, Mullainathan et al. (2008) and Fryer and Jackson (2009) use categorization to model coarse cognition; Wernerfelt (2004), Crémer et al. (2007), and Blume and Board (2014) use it to model coarse language. Another strand of research studies categorization of the elements of a given game; e.g., in Jehiel (2005) each player partitions the opponents' moves into similarity classes. We focus our review on contributions where the categorization concerns different games, because this case is closest to ours.

Heller and Winter (2016) assume that each agent initially and simultaneously decides a categorization over a finite set $\mathcal{G}$ of two-player games, committing to play the same strategy for all the games in the same category. Categorizations may be part of a subgame-perfect equilibrium and the strategic value of choosing a coarser partition can be positive, akin to our fog of cooperation.

Categorizations might emerge from an evolutionary process. Mengel (2012a) studies the evolutionary fitness of different cultures (viewed as different partitions of the game space) under the assumption of persistent noise in the transmission process across generations. Our model shows that categorizations can be advantageous even when errors play no role.

Samuelson (2001) studies finite automata that bundle bargaining games to save cognitive resources for more demanding games. Mengel (2012b) considers two players with arbitrary small reasoning costs who, under reinforcement learning, end up bundling games into categories that are then played identically. Bednar and Page (2007) use agent-based simulations to demonstrate that different rules of behavior emerge when different pairs of games are bundled in the same category.

Moving from theory to experiments, psychology has produced a vast literature on categorization in individual decision-making (Cohen and Lefebvre, 2005), but categorization has attracted far less attention within games. Halevy et al. (2012) show that individuals map the outcome interdependence from a variety of conflicts to only four archetypal situations. Grimm and Mengel (2012) provide evidence of learning spillovers across six games and match it to a model of coarse partitions of the space of games where agents best reply to the "average game" in each category; see Bednar et al. (2012) and Huck et al. (2011) for related experimental results. Transfer effects across games played in sequence have been attributed to perceiving their similarity; see Knez and Camerer (2000).

2 The model

We study an environment involving symmetric interactions where rational agents will either cooperate (as in common-interest games) or compete (as in zero-sum games). When these clear-cut interactions are conflated, interesting tensions can emerge.

For example, the prisoners' dilemma (PD) can be associated to a lottery between a common-interest (CI) game and a zero-sum (ZS) game; see Kalai and Kalai (2013). Consider the interaction between two parties facing a 50–50 lottery over the CI and ZS games below.

$$\begin{array}{c|c|c|} & H & L \\ \hline H & 10,10 & 6,6 \\ \hline L & 6,6 & 2,2 \\ \hline \end{array} \qquad\qquad \begin{array}{c|c|c|} & H & L \\ \hline H & 0,0 & -6,6 \\ \hline L & 6,-6 & 0,0 \\ \hline \end{array}$$

CI   ZS

This interaction is best-reply equivalent (Morris and Ui, 2004) to the PD game

	H	L
H	$5, 5$	$0, 6$
L	$6, 0$	$1, 1$

PD

based on the expected payoffs from the lottery. The cooperation motive in the CI game encourages agents to play (H, H), while the competitive motive in the ZS game suggests they play (L, L). Given the payoffs in the CI and ZS games above, when the lottery puts probability $p = 1/2$ on the CI game, the cooperative motive is weaker than the competitive one. The cooperative motive would instead prevail for $p > 3/5$.

We generalize this example by considering CI games with payoff parameter $a > 0$ and ZS games with payoff parameter $z > 0$. In the game $G(a, z; p)$ shown in Figure **??**, the two parties face a CI game with probability p and a ZS game with probability $1-p$.

	H	L
H	a, a	$0, 0$
L	$0, 0$	$-a, -a$

CI

	H	L
H	$0, 0$	$-z, z$
L	$z, -z$	$0, 0$

ZS

Figure 1: Nature draws CI or ZS, with probabilities p and $1 - p$.

Assuming that the parties move before uncertainty is resolved, the game $G(a, z; p)$ with imperfect information is best-reply equivalent to the game below.

	H	L
H	pa, pa	$-(1-p)z, (1-p)z$
L	$(1-p)z, -(1-p)z$	$-pa, -pa$

Defining

$$\pi = \frac{a}{a+z},$$

the dominant strategy for each party is to play H when $\pi > 1-p$ and L when $\pi < 1-p$. From a strategic viewpoint, it is therefore possible to reduce the number of dimensions from three to two: there is a projection from the $(a, z; p)$-space to the (π, p)-space that preserves best replies, while losing information about payoffs. Under this projection,

(π, p) is the (representative) element for a class of games that are best reply-equivalent. Therefore, with some abuse of language, we refer to (π, p) as a *game*.

We assume that p is uniformly distributed on $[0,1]$; moreover, a and z have independent exponential distributions with parameter $\lambda = 1$, so $\pi = a/(a+z)$ has a uniform distribution on $[0,1]$. The bidimensional *space of games* $\mathcal{G} = [0,1]^2$ is thus uniformly distributed and is depicted in Figure **??**. Any game with $\pi > 1-p$ is a CI game where H is the dominant strategy, and any game with $\pi < 1-p$ is a PD game where L is the dominant strategy.[3]

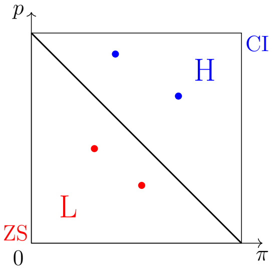

Figure 2: The space $\mathcal{G}$ of games.

As a benchmark, in this paragraph we consider the case where each party perceives any game drawn from the space $\mathcal{G}$ as distinct and plays the appropriate dominant strategy in whatever game is drawn. The expected payoff is $1/3$; see Proposition **??** in the Appendix, where we have collected theorems and proofs.

We henceforth assume that each dimension of the space $\mathcal{G}$—the payoff ratio π in $[0,1]$ and the probability p in $[0,1]$—is too rich to allow either party to perceive all its elements as distinct. Instead, each agent apprehends each dimension by means of a categorization that bundles uncountable points into a finite number of intervals. For simplicity, we work with binary categorizations, respectively defined by the thresholds $\hat{\pi}$ and $\hat{p}$. Thus, an agent categorizes π as High (h) if $\pi > \hat{\pi}$ and Low (ℓ) if $\pi < \hat{\pi}$; similarly, p is High (h) if $p > \hat{p}$ and Low (ℓ) if $p < \hat{p}$.[4]

An agent with binary categorizations for π and p perceives four cells, as depicted in Figure **??**. We call each of the four cells a *situation*. A cell bundles together many games, all of which are perceived by a party as instances of the same situation. That

[3] One may reformulate the space of games as a single game with payoff uncertainty.

[4] Which categorization applies at $\pi = \hat{\pi}$ or at $p = \hat{p}$ is immaterial, because this event has zero probability.

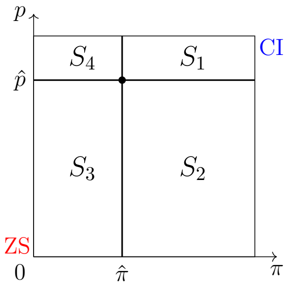

Figure 3: A categorization of $\mathcal{G}$ into four situations.

is, when an agent faces a game from $\mathcal{G}$ and wonders *"what kind of situation am I in?"*, only four answers come to her mind. For example, the northeastern cell S_1 corresponds to the situation where both π and p are perceived as h. If she views the dimension π as the (relative) *salience* of the cooperation payoff and the dimension p as the (likelihood of the) *opportunity* for cooperation, the situation S_1 involves high salience and high opportunity. The other three situations have similar interpretations.

The *frame* of an agent is the collection of the situations that she perceives, identified by the threshold pair $(\hat{\pi}, \hat{p})$. In the generative case, the frame is a direct consequence of the information structure, which is known by the agents. In the interpretive case, she simply perceives a game (π, p) as a situation, described by each of the dimensions π and p taking value h or ℓ: only the model-builder, not the agent, knows that the agent (a) categorizes games and (b) does so via the threshold pair $(\hat{\pi}, \hat{p})$.

Throughout this paper, we assume that interacting parties share the same frame $(\hat{\pi}, \hat{p})$: in each of the four situations associated with the frame, they perceive a single 2×2 symmetric game with payoffs equal to the expected payoffs from all the games ascribed to that situation.[5] In sum, agents' strategic understanding of the space of games $\mathcal{G}$ is coarsened into the four situations S_1, S_2, S_3, S_4 in Figure **??**. Using Lemmata **??** and **??**, the expected payoffs to the first party (rescaled by a factor of 2) are shown in Figure **??** for each of the four situations perceived under the frame $(\hat{\pi}, \hat{p})$.

After playing a perceived situation, the parties receive the payoffs associated with the actual game drawn: either CI with probability p or ZS with probability $1 - p$ from

[5] Conditional on the frame, the agents have correct beliefs about the distribution of payoffs in each situation. In the generative case, this occurs because they can compute this distribution. In the interpretive case, we assume that the cognitive process selecting their frame also provides them with correct beliefs about payoffs.

	H	L
H	$\hat{\pi}^2(1-\hat{p}^2)$	$-\hat{\pi}(2-\hat{\pi})(1-\hat{p})^2$
L	$\hat{\pi}(2-\hat{\pi})(1-\hat{p})^2$	$-\hat{\pi}^2(1-\hat{p}^2)$

S_4

	H	L
H	$(1-\hat{\pi}^2)(1-\hat{p}^2)$	$-(1-\hat{\pi})^2(1-\hat{p})^2$
L	$(1-\hat{\pi})^2(1-\hat{p})^2$	$-(1-\hat{\pi}^2)(1-\hat{p}^2)$

S_1

	H	L
H	$\hat{\pi}^2\hat{p}^2$	$-\hat{\pi}(2-\hat{\pi})\hat{p}(2-\hat{p})$
L	$\hat{\pi}(2-\hat{\pi})\hat{p}(2-\hat{p})$	$-\hat{\pi}^2\hat{p}^2$

S_3

	H	L
H	$(1-\hat{\pi}^2)\hat{p}^2$	$-(1-\hat{\pi})^2\hat{p}(2-\hat{p})$
L	$(1-\hat{\pi})^2\hat{p}(2-\hat{p})$	$-(1-\hat{\pi}^2)\hat{p}^2$

S_2

Figure 4: Perceived payoffs in the four situations under the frame $(\hat{\pi}, \hat{p})$.

Figure **??**. In the interpretive case, they ascribe the difference between the expected payoff and the realized payoff to noise.

We intend the interpretive case of this model as one way to capture differences— sometimes attributed to "culture"— in how agents perceive not only what situation they are in but also what the likely consequences of alternative actions are. By assuming that two agents share a frame, we imagine them coming from the same culture.

3 One-shot interaction

This section considers the one-shot interaction between two parties under a shared frame $(\hat{\pi}, \hat{p})$, with $\hat{\pi} + \hat{p} \neq 1$. We label the northeast and southwest situations S_1 and S_3 *congruous*, because their descriptors π and p are both high or both low: the salience of cooperation π and the opportunity for cooperation p are aligned. In contrast, we say that the two situations S_2 and S_4 are *incongruous* because their descriptors are misaligned: one is high and the other is low.

Rational behavior in the two congruous situations is unequivocal. Figure **??** shows that the situation S_1 is always perceived as a CI game under any frame $(\hat{\pi}, \hat{p})$, so H is the dominant strategy. Similarly, the situation S_3 is always perceived as a PD game, so L is the dominant strategy. In short, regardless of the frame, rational behavior for a party facing a congruous situation is to play H in S_1 and L in S_3.

Assuming that the frame satisfies $\hat{\pi} + \hat{p} \neq 1$, we find that there is also a unique dominant strategy for the incongruous situations S_2 and S_4. This is characterized in the next proposition, which is an immediate corollary of Proposition **??** in the Appendix.

Proposition 1. *The unique dominant strategy for both S_2 and S_4 is H if $\hat{\pi} + \hat{p} > 1$, and it is L if $\hat{\pi} + \hat{p} < 1$.*

Unlike the congruous situations, the dominant strategy for the incongruous situations depends on the frame.

Combining the dominant strategies over the four situations, we find two rational rules of behavior, shown in Figure **??**. The first rule, depicted on the left, is optimal if $\hat{\pi} + \hat{p} > 1$: play H in any situation except S_3, and then play L; we call it the *OR rule*, because it prescribes playing H when either π or p is perceived as high. The second rule, shown on the right, is optimal if $\hat{\pi} + \hat{p} < 1$: play H only in S_1 and otherwise play L; we call it the *AND rule*, because it prescribes playing H only when both π and p are perceived as high.

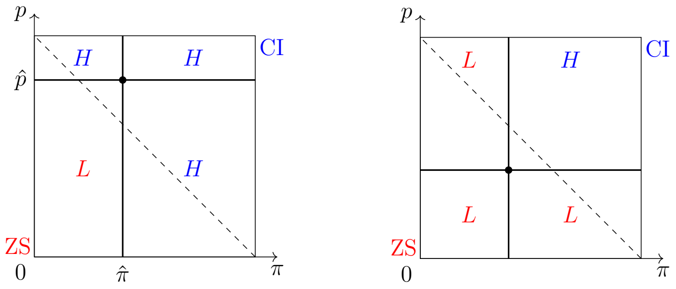

Figure 5: The OR and AND rules of behavior.

Since the frame is shared and payoffs are symmetric, the parties will play the same strategy in a given situation. If $\hat{\pi}+\hat{p} > 1$, they will play (H, H) in all situations except S_3 and (L, L) in S_3 (OR rule); if $\hat{\pi} + \hat{p} < 1$, they will play (H, H) only in S_1 and (L, L) otherwise (AND rule). Therefore, different frames can induce different strategy profiles when parties encounter incongruous situations.

Having computed optimal strategies, we next analyze how the parties' expected payoffs depend on the thresholds $(\hat{\pi}, \hat{p})$ of their shared frame. First, payoffs change continuously in $(\hat{\pi}, \hat{p})$ unless the thresholds induce a switch in the parties' rule of behavior; second, if the rule of behavior switches, then there is a discontinuous change in payoffs.

Proposition **??** gives the expected payoff to each party as a function of $\hat{\pi}$ and $\hat{p}$. As an illustrative example, suppose $\hat{\pi} = \hat{p} = x$ so that a change in x makes the thresholds shift in lockstep. The parties play the OR rule for $x > 1/2$ and the AND rule for

$x < 1/2$. Figure **??** shows the payoff to each party as a function of the common value x for the two thresholds. Within either the OR or the AND region, payoffs are continuously decreasing in x. On the other hand, moving x leftward across $1/2$ implies an abrupt drop in payoffs, as the parties switch to the less cooperative AND rule of behavior. Nonetheless, depending on x, the AND rule may outperform the OR rule.

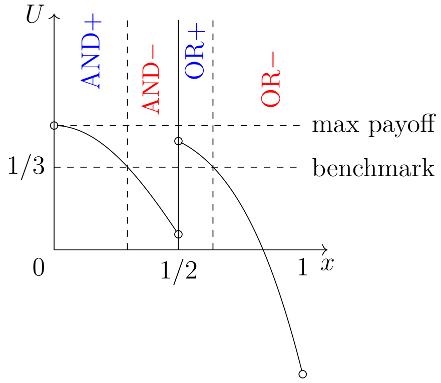

Figure 6: Payoffs as a function of x when $\hat{\pi} = \hat{p} = x$.

Framing games as situations can either help or hurt the parties' payoffs, relative to the benchmark case where each game is perceived as distinct. This is apparent in Figure **??**, where the benchmark payoff of $1/3$ cuts across the payoff curve. Intuitively, one may think of the frame as creating a fog that confounds different games into a single situation, forcing a party to deal with all such games in one way. Depending on the frame, the result is either a *fog of conflict* (marked $-$), making agents play less cooperatively than they would under full information, or a *fog of cooperation* (marked $+$), making them play more cooperatively. Note that either fog can occur under either rule of behavior, so frames evidently do more than determine rules of behavior.

We can identify which frames generate which kind of fog. Proposition **??** states formally a simple characterization, but the main message is conveyed through a picture. Each frame is associated with a threshold pair $(\hat{\pi}, \hat{p})$ in $[0,1]^2$ so the unit square in Figure **??** stands for the space of the frames. We emphasize that Figure **??** *does not* portray the space $\mathcal{G}$ of the games (π, p), but rather the set of threshold pairs $(\hat{\pi}, \hat{p})$ that define a frame.[6]

The OR rule prevails when the parties' shared frame is a threshold pair $(\hat{\pi}, \hat{p})$ above

[6] One of us enjoys the mnemonic "put your hat on" when keeping track of p versus $\hat{p}$.

the diagonal, while the AND rule prevails when it is a pair below. The curve above

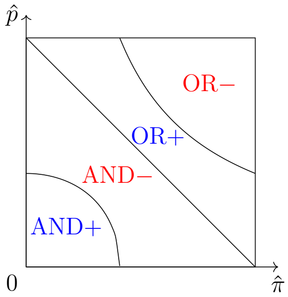

Figure 7: Fog of cooperation (+) and fog of conflict (−).

the diagonal separates the OR region into the shared frames $(\hat{\pi}, \hat{p})$ generating a fog of conflict (marked −) versus those generating a fog of cooperation (marked +). Similarly, the curve below the diagonal separates the AND region into fog of conflict (marked −) versus fog of cooperation (marked +). Consistent with the special case depicted in Figure **??**, moving from northeast to southwest within a rule in Figure **??** improves payoffs continuously; on the other hand, crossing the boundary coming from the OR into the AND rule causes a discontinuous drop in payoffs. Switches in the rule of behavior motivate part of our discussion about changing frames in Section **??**.

As one way to summarize the interpretive case of this static model, we find it useful to imagine two parties who share a low-performing frame arranging a site visit to observe two other parties who share a high-performing frame. All parties perceive situations in terms of their own categorizations; the low-performing parties observe the actions chosen by the high-performing parties.

As a dramatic example, consider the discontinuity at $x = 1/2$ in Figure **??**, and suppose that the low- and high-performing frames have $x = .45$ and $x = .55$, respectively. For all the games (π, p) from the set $(.45, .55) \times (.45, .55)$, the low-performing parties perceive the situation S_1 and hence a CI game, expecting the high-performing parties to choose (H, H), while the high-performing parties perceive the situation S_3 and hence a PD game, leading them to choose (L, L). Thus, on this set of games, the high-performing team does worse.

The more important difference cuts the other way: in the incongruous situations the high-performing parties perceive a CI game and so choose (H, H), whereas the low-performing parties perceive a PD (even for games when both teams agree that an incongruous cell has been realized) and so choose (L, L). In sum, the low-performing parties will be mystified by the site visit: when they see CI, they observe their hosts playing (L, L) with small probability; and when they see PD, they observe their hosts playing (H, H) with larger probability.

We see the interpretive case of this stylized model as a small step towards understanding widespread evidence of differences in cooperation during evolution (Boyd and Richerson, 2009) and among cultures (Henrich et al., 2005), communities (Ostrom, 1990), firms (Leibenstein, 1982), and teams (Cole, 1991). Moving from the field to the laboratory, experiments show that cultural frames differ in how they perceive situations as "cooperative" or "competitive" (Keller and Lowenstein, 2011) and how different frames affect cooperation levels (Pruitt, 1970; Liberman et al., 2004). Many explanations of such differences in cooperation emphasize differences in preferences; we provide a complementary explanation based on differences in shared cognition which, in turn, might arise from cultural differences.

4 Repeated interaction

Having constructed a model where shared frames shape behavior in static situations, we next consider the case of infinitely repeated interactions. Under any frame, the congruous situation S_3 is perceived as a PD. Furthermore, if $\hat{\pi} + \hat{p} < 1$, then the incongruous situations S_2 and S_4 are also perceived as PDs. In a repeated interaction, familiar logic might allow the parties to cooperate in some or all of these PDs, even if they would defect in a one-shot interaction.

We analyze such opportunities for long-term cooperation using a multi-period model, where in each period the stage game is randomly drawn from the space $\mathcal{G}$ of games and perceived as one of four situations under the shared frame $(\hat{\pi}, \hat{p})$. As in the static model, given their frame, the parties have correct beliefs: before a game is drawn in a given period, the parties expect to face situation S_1 with probability $(1 - \hat{\pi})(1 - \hat{p})$, situation S_2 with probability $(1 - \hat{\pi})\hat{p}$, situation S_3 with probability $\hat{\pi}\hat{p}$, and situation S_4 with probability $\hat{\pi}(1 - \hat{p})$. We assume that the parties have the same discount factor $\delta < 1$, and we rescale their discounted payoffs by a factor $(1 - \delta)$ to make them comparable to the one-shot payoffs.

We analyze subgame-perfect equilibria in trigger strategies where defection (i.e., playing L) in a PD situation is met by defection in all future PD situations, discarding the possibility that punishment calls for a party to play (the dominated) action L in a CI situation. We are especially interested in the case of full cooperation, when agents play (H, H) everywhere.

As above, we briefly consider the benchmark case where the parties can distinguish all the games in $\mathcal{G}$. Then the parties can support full cooperation if $\delta \geq 24/25$; see Proposition **??**.

Returning to the case of a shared frame, recall that the one-shot model calls for the OR rule of behavior if $\hat{\pi} + \hat{p} > 1$, but the AND rule if $\hat{\pi} + \hat{p} < 1$. As above, we reduce the number of parameters by assuming $\hat{\pi} = \hat{p} = x$; then the two rules obtain for $x > 1/2$ and $x < 1/2$.

Consider $x > 1/2$: the only situation perceived as a PD is S_3. The frame's footprint, here given by x, has three effects. First, when x increases, there is a probability x^2 that the PD situation occurs in the future. Second, when the PD situation does occur, the temptation of a short-term payoff improvement from defecting (L) rather than cooperating (H) initially goes up (until $x = 2/3$) and then shrinks to zero. Finally, in the trigger-strategy equilibrium, the threat of a long-term payoff loss from reversion to the static equilibrium after defecting is decreasing in x. These effects of shared cognition on the *perceived* frequency of PD situations and on the *perceived* relative strength of temptation versus punishment all influence the viability of long-term cooperation.

In spite of these three effects of the frame's footprint, the familiar intuition that a sufficiently high δ supports cooperation (i.e, both parties playing H) in the S_3 situation survives: if

$$\delta \geq \frac{2 - 2x}{2 - 2x + x^4}$$

then there is a trigger-strategy equilibrium where the parties play H in the S_3 situation; see Proposition **??**. This critical region for $x > 1/2$ is illustrated in Figure **??**: given δ and x, either the parties can sustain cooperation in **ALL** situations, or they cooperate only in those situations perceived to be CI (which, given $x > 1/2$, is the OR rule from the static game). In particular, for $\delta \geq 16/17$, sustaining ALL is possible for any value $x > 1/2$.

Consider now $x < 1/2$. Proposition **??** demonstrates a similar result, shown in Figure **??** for $x < 1/2$: given δ and x, either the parties can sustain cooperation in

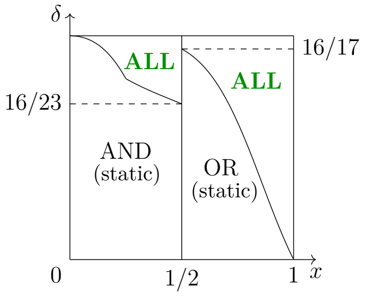

Figure 8: Optimal cooperation in the repeated interaction.

ALL situations, or they cooperate only in the situation perceived to be CI (which, given $x < 1/2$, is the AND rule from the static game).

Our analysis reiterates the familiar theme that repetition and patience may allow the parties to achieve higher payoffs than in the static model. The novel point here is that performance differences may arise from differences in shared frames, even when all parties share the same discount factor and are playing the best repeated–interaction equilibrium they can, given how they perceive the space of games.

This novel point is illustrated most vividly if we assume a discount factor $16/23 < \delta < 16/17$. Then Figure **??** shows that, as x progresses from 0 to 1, the best outcome that parties can sustain in a repeated interaction changes from AND to ALL to OR to ALL again. For instance, taking $\delta = .8$, Figure **??** shows that either expected payoffs are at their maximum (normalized) value of 1/2 (such as for $x_1 < x < 1/2$ or for $x_2 < x < 1$) or they are at their static level (such as for $0 < x < x_1$ or for $1/2 < x < x_2$), where the critical values $x_1 \approx 0.255$ and $x_2 \approx 0.648$ depend on the chosen value $\delta = .8$.

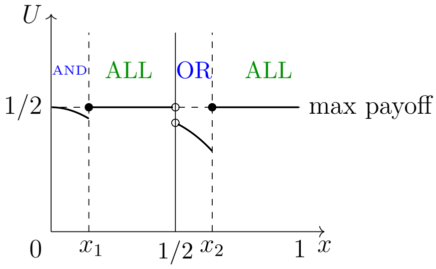

Figure 9: Payoffs in the repeated interaction for $\delta = .8$.

In the generative case, this analysis of repeated interaction shows that, for a given

δ, there may be quite different information structures that are equally conducive to full cooperation: $x_1 < x < 1/2$ or $x_2 < x < 1$. Furthermore, the full information benchmark says that full cooperation is feasible only if $\delta \geq 24/25$. But, taking for instance $\delta = .8$ in Figure **??**, it is possible to have full cooperation using coarser frames, regardless of which rational rule of behavior (AND/ OR) these frames support in one-shot interactions. This latter result shows that adding fog in agents' information structures can help them sustain full cooperation.[7]

For the interpretive case, we again imagine low-performing parties conducting a site visit to observe high-performing parties. Importantly, all parties share a common discount factor, such as $\delta = .8$ in Figure **??**. Parties able to achieve only the static equilibrium (whether under the AND rule for $0 < x < x_1$ or under the OR rule for $1/2 < x < x_2$) understand that the interaction is repeated but calculate that δ is not high enough to allow the ALL equilibrium, whereas parties with different shared frames ($x_1 < x < 1/2$ or $x_2 < x < 1$) share the same δ but calculate that a trigger-strategy equilibrium will support cooperation in ALL situations.

The low-performing parties will again be mystified by the site visit, observing cooperation on (H, H) in many situations that all four parties perceive as PDs, and perhaps hearing explanations from the high-performing parties such as "Sure, we see a PD, but the shadow of the future makes cooperation credible today."

We saw the interpretive case of the static model as a small step towards understanding a broad set of findings concerning evolution, cultures, communities, firms, teams, and more. In contrast, we see the interpretive case of this repeated model as a larger step towards a smaller set of findings.

A growing literature documents persistent performance differences among seemingly similar plants and firms; e.g., Syverson (2011). Bloom et al. (2016) find that measures of management practices correlate with these performance differences, and Gibbons and Henderson (2012) argue that many competitively significant management practices rely on relational contracts that seem difficult to replicate.

Moving to theory, Kreps (1990b) suggested long ago that different equilibria in a repeated game might correspond to different corporate cultures (shared understandings

[7] This point may be reinforced. As noted above, our benchmark with agents distinguishing games up to (π, p) allows full cooperation for $\delta \geq 24/25$. If we let them have a finer partition and distinguish games up to (a, z, p), then Proposition **??** shows that full cooperation is impossible for any δ.

of “how we do things here”) associated with different performance levels across plants and firms. But modeling performance differences as resulting from different equilibria in a given game implies that low performers know that better equilibria exist; and yet the model gives these parties no way to try to reach a better equilibrium and hence offers no rationale for why moving to a better equilibrium might be difficult. Our model formalizes one such difficulty: low performers are playing the best equilibrium they can perceive; reaching a better equilibrium would require changing the parties' frame.

As one example along these lines, consider the innovations introduced by Toyota in supply-chain management, which redefine how some key interactions are perceived. The difficulties of reshaping these shared representations contribute to the difficulties in replicating the results of the Toyota system even after having imported its contractual framework and management tools (Helper and Henderson, 2014). In short, Toyota's system is geared more towards explaining when a supplier is expected to cooperate than towards excluding him from future interactions (Womack et al., 1990). Ostrom (1990) makes similar observations about education versus exclusion in communities that successfully manage common-pool resources.

5 Shifts in the frame

The parties' shared frame could change in many ways; for instance, it could be refined by increasing the number of cells. In this section we consider shifts in the thresholds, which change the size of the cells but not their number. The change in the frame may be driven by an outside shock or a purposeful third agent. For simplicity, we assume that the new frame is simultaneously shared by both parties.

Consider the simple example where the current threshold pair is $(\hat{\pi}, \hat{p})$, with $\hat{p} = 1/2$ and the initial threshold $\hat{\pi} > 1/2$ changing to a new threshold $\hat{\pi}' < \hat{\pi}$. We distinguish two cases: (a) $\hat{\pi}' > 1/2$, versus (b) $\hat{\pi}' < 1/2$.

The case $\hat{\pi}' > 1/2$ is shown in Figure **??**. When the parties' frame changes, they recategorize some games as different situations. Because $\hat{\pi}' < \hat{\pi}$, the probabilities that the agents perceive S_1 and S_4 increase, and the probabilities that they perceive S_2 and S_3 decrease. On the other hand, because $\hat{\pi}' + \hat{p} > 1$, their rational rule of behavior does not change: it remains the OR rule, and S_2 and S_4 remain CI games. In short, the

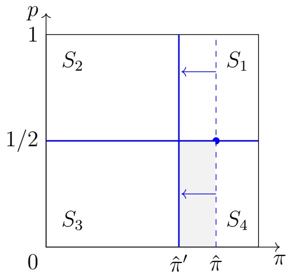

Figure 10: Lowering the threshold from $\hat{\pi}$ to $\hat{\pi}' > 1/2$.

parties' rational behavior changes only for the games shaded in Figure ??: the parties recategorize these games from S_3 (PD) to S_4 (CI) and thus switch behavior from L to H.

The case $\hat{\pi}' < 1/2$ is shown in Figure ??. The parties used to play the OR rule

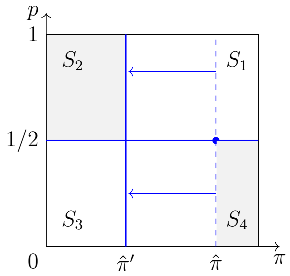

Figure 11: Lowering the threshold from $\hat{\pi} > 1/2$ to $\hat{\pi}' < 1/2$.

but, because $\hat{\pi}' + \hat{p} < 1$, their rational behavior now switches to the AND rule and they change strategies (from H to L) in the incongruous situations S_2 and S_4 shaded in Figure ??. Note that behavior changes only in games ascribed to the same incongruous situations both before and after the change in frame; this change in behavior occurs because the change in expected payoffs for the incongruous situations causes the parties to now perceive these situations as PD instead of CI.

Even in this simple example we see that a change of the frame may affect not only how the agents perceive situations (the frame's footprint) but also how they rationally behave in a given situation (their rule of behavior). Within the context of this example, we call a change in frame *incremental* if the frame's footprint changes but the rational rule of behavior (OR or AND) does not, and we call it *radical* if the rational rule of behavior also changes.

We henceforth assume that frame change is driven by a purposeful third person (called *leader*) who seeks to influence the parties (called *followers*). As an initial model, we assume that the leader knows everything the modeler knows; in particular, the leader is aware that the followers perceive the space of games as situations generated by a frame $(\hat{\pi}, \hat{p})$.

The novel—and, in our interpretive case, realistic—aspect of our analysis is that the leader lacks full control of the parties' new frame: the final outcome of an attempt to change thresholds is uncertain. This uncertainty may concern: (a) appraisal (e.g., whether and how much thresholds change), (b) evaluation (e.g., how long it takes before followers revise their beliefs about payoffs in a reconfigured situation), or (c) behavior (e.g., how long followers stick with the old rule of behavior after revising their beliefs about payoffs).

Concerning (a) *appraisal*, the leader takes unmodelled actions that may shift the followers' threshold $\hat{\pi}$, thus reframing their understanding of the four situations. In our interpretive case, we imagine the leader taking actions such as using certain language, or extolling certain behaviors, or telling certain stories—all as efforts to change the followers' frame. The direction of this change may be clear, but its magnitude will often be uncertain. Thus we assume that the leader controls the direction of change (aiming for a higher or a smaller threshold $\hat{\pi}'$), but not the exact shift. As above, the new frame is simultaneously shared by both followers.

Concerning (b) *evaluation*, the followers may need time to revise their beliefs about the payoffs associated with the situations defined by the new frame. We consider only two extreme cases: either the followers revise payoffs immediately, or with some delay.

Concerning (c) *behavior*, we make the standard assumption that followers immediately adjust behavior to a perceived change in payoffs. All this yields two scenarios: (1) the followers react immediately to a change in frame, revising payoffs at once and changing their rule of behavior (if needed); or (2) the followers exhibit inertia (because of a delay in revising beliefs about payoffs) and hence stick to their previous rule of behavior for a while.

Returning to our simple example, the leader may lower or raise the $\hat{\pi}$ threshold from its initial value. We now allow $\hat{\pi}$ to be greater or less than $1/2$. We code the leader's two options as L (Lower the threshold by moving it to the Left) and R (Raise the threshold by moving it to the Right). The leader is benevolent and maximizes the sum of the payoffs obtained by the followers in their static interaction under a new frame. Proposition **??** gives the expected payoff U to each party. Doubling U, we obtain the leader's payoff U_L represented in Figure **??**.

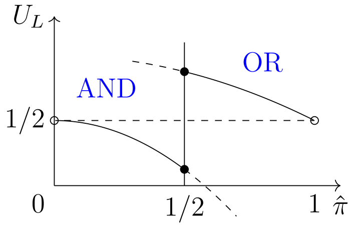

Figure 12: Leader's payoff as a function of $\hat{\pi}$ for $\hat{p} = 1/2$.

Consider scenario (1), where the followers' appraisal and evaluation are updated immediately after the frame changes. Suppose that the initial threshold is $\hat{\pi} > 1/2$, so the followers use the OR rule. What are the leader's options? Choosing R is dominated by staying put. Choosing L, on the other hand, is risky: moving to a new threshold $\hat{\pi}' < \hat{\pi}$ increases payoffs only if $\hat{\pi}'$ does not cross over 1/2. In short, under the OR rule, the leader might want to pursue a strategy of *incremental change* (i.e., a mild reduction of the threshold), unless the risk of a radical change—with followers switching to the AND rule—is too high.

Suppose instead that the initial threshold is $\hat{\pi} < 1/2$, so the followers use the AND rule. Choosing L is a safe option that increases payoffs through incremental change. The alternative is to choose R and attempt a radical change: if the leader can get the followers' threshold to cross the 1/2 barrier, this yields a substantial improvement in payoffs, but an attempt for a radical change carries the risk that the new threshold moves right without crossing the barrier, making payoffs worse than before.

Now consider scenario (2), where only the followers' appraisal is updated immediately, but they then need time to update their evaluation of payoffs and change their rule of behavior. This adds new trade-offs to the choice between incremental versus radical change.

Under incremental change, the original rule of behavior is still optimal, so there are no delayed effects on behavior (even if evaluation is slow, as postulated in this scenario). But suppose that $\hat{\pi} < 1/2$ and the leader attempts a radical change to $\hat{\pi}' > 1/2$. The followers are initially using the AND rule, so after a radical change crossing 1/2 from the left the payoff stays on the lower dashed curve until the followers adjust behavior, and afterwards jumps up to the higher solid curve—both at the new threshold $\hat{\pi}' > 1/2$. In short, things get worse before they get better: inertia in the adaptation of behavior to a radical change in the frame may cause a transient decline in performance before producing its positive effects—an "implementation dip" (Fullan, 2001).

Conversely, suppose $\hat{\pi} > 1/2$ and the leader attempts an incremental change to the left that goes too far, becoming a radical change to $\hat{\pi}' < 1/2$. Such radical change would enjoy an initial success before its ultimate failure: transient payoffs would be on the upper dashed curve, but long-run payoffs would be on the lower solid curve.[8]

The view that leaders take action to change followers' mental models is central to the leadership literature (Argyris, 1982; Schein, 1985; Senge, 1990). Our model is also consistent with theory and evidence from the organizations literature that radical change is a risky activity: its development and timing are often opaque (Hannan et al., 2003), leading to a high rate of failure. And there is broad consensus that radical change is easier in greenfield sites, where it can be seeded from scratch, than in brownfield sites, where it is inhibited by pre-existing frames (Brynjolfsson et al., 1997).

Because the leader lacks full control over the followers' new frame, the leader's risk attitude could influence his propensity to attempt incremental versus radical change. In addition, the leader's overconfidence (Weinstein, 1980; Camerer and Lovallo, 1999) about the transience of a performance decline or about her ability to minimize this decline may bias her towards attempting radical change. On the other hand, if pressures for current performance push an organization towards abandoning radical change prematurely (Repenning and Sterman, 2002), the leader's overconfidence might beneficially promote persistence in the change effort.

6 Conclusions

This paper analyzes how coarse cognition can influence strategic interactions. We constructed a simple framework consistent with both economic and cognitive perspec-

[8] In Figure **??** $\hat{p}$ is fixed at 1/2 and the OR rule of behavior dominates the AND rule of behavior, but Figure **??** in Section **??** shows that this is general: therefore, an initial success before ultimate failure may also occur when radical change is actively sought, not just when it arises mistakenly.

tives, uniting them through the assumption that the parties share a cognitive frame. The shared frame can be attributed either to the information the parties receive (the generative approach favored by economic theory) or to the parties' perception of the environment (the interpretive viewpoint preferred by the cognitive sciences).

We provide three main results. The first is that changes in the cognitive frame may induce a fog of cooperation or a fog of conflict. The second is that, in a repeated interaction, the frame's footprint is as important as agents' patience in achieving full cooperation. And the third is that incremental or radical changes in frame may have starkly different (short- and long-run) consequences for performance.

In future work, there may be interesting applications of our generative case to information design; we intend to explore our interpretive case in various ways. For example, when two parties hold frames that differ only slightly, they may proceed in harmony for some time before an unexpected outcome occurs: we are interested in how the parties' efforts to diagnose such a misalignment in frames may lead to a repair or a rupture of their relationship. We also plan to study some of the leader's actions for changing a frame; we hope such work will contribute to new perspectives on language and stories in organizations. More generally, we would like to deepen our understanding of how shared and persistent cognitive representations—typical of a culture—affect economic interactions.

References

[1] G.W. Allport (1954), *The Nature of Prejudice.* Cambridge (MA): Addison-Wesley.

[2] M. Aoki (2001), *Toward a Comparative Institutional Analysis*, Cambridge (MA): The MIT Press.

[3] C. Argyris (1982), *Reasoning, Learning, and Action*, San Francisco: Jossey-Bass.

[4] M. Bacharach (2003), "Framing and cognition in economics", in: N. Dimitri, M. Basili and I. Gilboa (eds.), *Cognitive Processes and Economic Behaviour*, London: Routledge, pp. 63–74.

[5] J. Bednar and S. Page (2005), "Can game(s) theory explain culture?", *Rationality and Society* **19**, 65–97.

[6] J. Bednar, Y. Chen, T.X. Liu and S. Page (2012), “Behavioral spillovers and cognitive load in multiple games: An experimental study”, *Games and Economic Behavior* **74**, 12–31.

[7] D. Bergemann and S. Morris (2017), “Information design: A unified perspective”, *Cowles Foundation Discussion Paper n. 2075*, February.

[8] N. Bloom, R. Sadun and J. van Reenen (2016), “Management as a technology?”, *CEPR Discussion Paper* **1433**, June.

[9] A. Blume and O. Board (2014), “Intentional vagueness”, *Erkenntnis* **79**, 855–899.

[10] R. Boyd and P.J. Richerson (2009), “Culture and the evolution of human cooperation”, *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**, 3281–3288.

[11] E. Brynjolfsson, A.A. Renshaw and M. Van Alstyne (1997), “The matrix of change”, *Sloan Management Review* **38**, 37–54.

[12] C. Camerer and D. Lovallo (1999), “Overconfidence and excess entry: An experimental approach”, *American Economic Review* **89**, 306–318.

[13] H. Cohen and C. Lefebvre (2005), eds., *Handbook of Categorization in Cognitive Science*, Amsterdam: Elsevier.

[14] R.E. Cole (1991), *Strategies for Learning: Small-group Activities in American, Japanese, and Swedish Industry*, Berkeley: University of California Press.

[15] K. Craik (1943), *The Nature of Explanation*, Cambridge (UK): Cambridge University Press.

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

[17] A. Denzau and D. North (1994), “Shared mental models: Ideologies and institutions”, *Kyklos* **47**, 3–31.

[18] P. DiMaggio (1997), “Culture and cognition”, *Annual Reviews of Sociology* **23**, 263–287.

[19] R. Fryer and M.O. Jackson (2008), “A categorical model of cognition and biased decision making”, *The B.E. Journal of Theoretical Economics* **8**, Article 6.

[20] M. Fullan (2001), *Leading in a Culture of Change*, San Francisco: Jossey-Bass.

[21] R. Gibbons and R. Henderson (2012), “Relational contracts and organizational capabilities”, *Organization Science* **23**, 1350–1364.

[22] E. Goffman (1974), *Frame Analysis: An Essay on the Organization of Experience*, Cambridge (MA): Harvard University Press.

[23] V. Grimm and F. Mengel (2012), “An experiment on learning in a multiple games environment”, *Journal of Economic Theory* **147**, 2220–2259.

[24] N. Halevy, E.Y. Chou and J.K. Murnighan (2012), “Mind games: The mental representation of conflict”, *Journal of Personality and Social Psychology* **102**, 132–148.

[25] M.T. Hannan, L. Polos and R.G. Carroll (2003), “The fog of change: Opacity and asperity in organizations”, *Administrative Science Quarterly* **48**, 399–432.

[26] Y. Heller and E. Winter (2016), “Rule rationality”, *International Economic Review* **57**, 997–1026.

[27] S. Helper and R. Henderson (2014), “Management practices, relational contracts, and the decline of General Motors”, *Journal of Economic Perspectives* **28**, 49–72.

[28] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N.S. Henrich, K. Hill, F. Gil-White, M. Gurven, F.W. Marlowe, J.Q. Patton, and D. Tracer (2005), “‘Economic man’ in cross-cultural perspective: Behavioral experiments in 15 small-scale societies”, *Behavioral and Brain Sciences* **28**, 795–815.

[29] K. Hoff and J.E. Stiglitz (2016), “Equilibrium fictions: A cognitive approach to societal rigidity”, *American Economic Review: Papers and Proceedings* **100**, 141–146.

[30] K. Hoff and J.E. Stiglitz (2016), “Striving for balance in economics: Towards a theory of the social determination of behavior”, *Journal of Economic Behavior and Organization* **126**, 25–57.

[31] L. Hong and S. Page (2009), “Interpreted and generated signals”, *Journal of Economic Theory* **144**, 2174–2196.

[32] S. Huck, P. Jehiel and T. Rutter (2011), “Feedback spillover and analogy-based expectations: A multi-game experiment”, *Games and Economic Behavior* **71**, 351–

365.

[33] P. Jehiel (2005), "Analogy-based expectation equilibrium", *Journal of Economic Theory* **123**, 81–104.

[34] A. Kalai and E. Kalai (2013), "Cooperation in strategic games revisited", *Quarterly Journal of Economics* **128**, 917–966.

[35] J. Keller and J. Loewenstein (2011), "The cultural category of cooperation: A cultural consensus model analysis for China and the United States", *Organization Science* **22**, 299–319.

[36] M. Knez and C. Camerer (2000), "Increasing cooperation in prisoner's dilemmas by establishing a precedent of efficiency in coordination games", *Organizational Behavior and Human Decision Processes* **82**, 194–216.

[37] D.M. Kreps (1990a), *Game Theory and Economic Modeling*, Oxford: Oxford University Press.

[38] D.M. Kreps (1990b), "Corporate culture and economic theory", in: J.E. Alt and K.A. Shepsle (eds.), *Perspectives on Positive Political Economy*, Cambridge (UK): Cambridge University Press, 90–143.

[39] H. Leibenstein (1982), "The prisoners' dilemma in the invisible hand: An analysis of intrafirm productivity", *American Economic Review* **72**, 92–97.

[40] V. Liberman, S.M. Samuels and L. Ross (2004), "The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves", *Personality and Social Psychology Bulletin* **30**, 1175–1185.

[41] F. Mengel (2012a), "On the evolution of coarse categories", *Journal of Theoretical Biology* **307**, 117–123.

[42] F. Mengel (2012b), "Learning across games", *Games and Economic Behavior* **74**, 601–619.

[43] S. Morris and T. Ui (2004), "Best response equivalence", *Games and Economic Behavior* **49**, 260–287.

[44] S. Mullainathan, J. Schwartstein and A. Shleifer (2008), "Coarse thinking and persuasion", *Quarterly Journal of Economics* **123**, 577–619.

[45] E. Ostrom (1990), *Governing the Commons: The Evolution of Institutions for*

Collective Action, New York: Cambridge University Press.

[46] E. Ostrom (2005), *Understanding Institutional Diversity*, Princeton: Princeton University Press.

[47] D.G. Pruitt (1970), “Motivational processes in the decomposed prisoner’s dilemma game”, *Journal of Personality and Social Psychology* **14**, 227–238.

[48] N.P. Repenning and J.D. Sterman (2002). “Capability traps and self-confirming attribution errors in the dynamics of process improvement”, *Administrative Science Quarterly* **47**, 265–295.

[49] A. Rubinstein (1991), “Comments on the interpretation of game theory”, *Econometrica* **59**, 909–924.

[50] L. Samuelson (2001), “Analogies, adaptation, and anomalies”, *Journal of Economic Theory* **97**, 320–366.

[51] L.J. Savage (1954), *The Foundations of Statistics*, New York: Wiley. Second edition published by Dover in 1972.

[52] E. Schein (1985), *Organizational Culture and Leadership.* San Francisco, CA: Jossey-Bass.

[53] P.M. Senge (1990), *The Fifth Discipline: The Art and Practice of the Learning Organization*, New York: Doubleday.

[54] H.A. Simon (1986), “Rationality in psychology and economics”, *Journal of Business* **59**, S209–S224.

[55] C. Syvverson (2011), “What determines productivity?”, *Journal of Economic Literature* **49**, 326–65.

[56] N.D. Weinstein (1980), “Unrealistic optimism about future life events”, *Journal of Personality and Social Psychology* **39**, 806–820.

[57] B. Wernerfelt (2004), “Organizational languages”, *Journal of Economics and Management Strategy* **13**, 461–472.

[58] J.P. Womack, D.T. Jones and D. Roos (1990), *The Machine that Changed the World*, New York: The Free Press.

For Online Publication

A Appendix

Throughout the appendix, we use majuscules to denote random variables and minuscules to denote their realizations. There are three independent r.v.'s.: A and Z have exponential distributions with $\lambda = 1$, and P has a uniform distribution on $(0, 1)$. It can be shown that the ratio $\Pi = \frac{A}{A+Z}$ has a uniform distribution on $(0, 1)$. Situations (Π, P) are uniformly distributed on $(0, 1)^2$.

Proposition A.1. *Under the benchmark, the expected payoff from playing a randomly drawn situation is* $1/3$*.*

Proof. Suppose that the parties use their dominant strategies. When $\Pi + P > 1$, the (expected) payoff to an agent is AP; when $\Pi + P < 1$, it is $-AP$. Write this payoff function as

$$V = -AP + 2AP\mathbb{1}_{\{\Pi \geq 1-P\}} = -AP + 2AP\mathbb{1}_{\left\{Z \leq \frac{AP}{1-P}\right\}}$$

Since $Z \sim \text{Exp}(1)$, $E\left(\mathbb{1}_{\left\{Z \leq \frac{AP}{1-P}\right\}} \mid A, P\right) = 1 - e^{-\frac{AP}{1-P}}$. Compute the conditional expectation

$$E(V|A, P) = -AP + 2AP\mathbb{1}_{\left\{Z \leq \frac{AP}{1-P}\right\}} = -AP + 2AP\left[1 - e^{-\frac{AP}{1-P}}\right] = AP - 2APe^{-\frac{AP}{1-P}}$$

Given that $E(Ae^{-kA}) = 1/(1+k)^2$, we have $E\left(Ae^{-\frac{AP}{1-P}} \mid P\right) = (1-P)^2$. Using independence, the expectation of $E(V|A, P)$ with respect to $A \sim \text{Exp}(1)$ is

$$E(V|P) = P - 2P(1-P)^2 = -P + 4P^2 - 2P^3$$

and, finally, computing its expectation with respect to $P \sim U[0, 1]$ we obtain

$$E(V) = -\frac{1}{2} + \frac{4}{3} - \frac{1}{2} = \frac{1}{3}$$

□

As discussed in the main text, a game $G(a, z; p)$ is best-reply equivalent to the game

	H	L
H	pa, pa	$-(1-p)z, (1-p)z$
L	$(1-p)z, -(1-p)z$	$-pa, -pa$

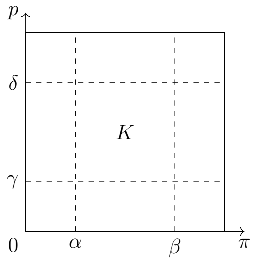

Figure 13: A cell in (the cartesian product of) a threshold categorization.

A frame bundles several games as a single situation. We compute the expected payoffs for each strategy profile over all the games categorised in the same situation. The next two lemmas provide the building blocks for the computations. For generality, let $K = (\alpha, \beta) \times (\gamma, \delta)$ be the cell including the games $G(A, Z; P)$ with $\Pi = A/(A + Z) \in (\alpha, \beta)$ and $P \in (\gamma, \delta)$, where $0 \leq \alpha < \beta \leq 1$ and $0 \leq \gamma < \delta \leq 1$; see Figure **??**.

Lemma A.2. *The expected value of AP over all games $G(A, Z; P)$ in K is*

$$E_K \left(AP \right) = \frac{(\beta^2 - \alpha^2)(\delta^2 - \gamma^2)}{2}$$

Proof. Let

$$V = AP\mathbb{1}_{\{K\}} = AP\mathbb{1}_{\{\alpha \leq \Pi \leq \beta\}}\mathbb{1}_{\{\gamma \leq P \leq \delta\}}$$

Clearly, $\alpha \leq \Pi \leq \beta$ if and only if $\frac{\alpha}{1-\alpha}Z \leq A \leq \frac{\beta}{1-\beta}Z$; thus, we have

$$V = AP\mathbb{1}_{\left\{\frac{\alpha}{1-\alpha}Z \leq A \leq \frac{\beta}{1-\beta}Z\right\}}\mathbb{1}_{\{\gamma \leq P \leq \delta\}}$$

Using independence,

$$E(V) = E\left(A\mathbb{1}_{\left\{\frac{\alpha}{1-\alpha}Z \leq A \leq \frac{\beta}{1-\beta}Z\right\}}\right) \cdot E\left(P\mathbb{1}_{\{\gamma \leq P \leq \delta\}}\right)$$

Recall that $\int ae^{-a}\,\mathrm{d}a = -(1+a)e^{-a}$. For $k = \frac{\alpha}{1-\alpha} < \frac{\beta}{1-\beta} = m$, we have

$$\begin{aligned}
E\left(A\mathbb{1}_{\{kZ\le A\le mZ\}}\right) &= \int_0^{+\infty} e^{-z}\left(\int_{kz}^{mz} ae^{-a}\,\mathrm{d}a\right)\mathrm{d}z \\
&= \int_0^{+\infty} e^{-z}\left[(1+kz)e^{-kz} - (1+mz)e^{-mz}\right]\mathrm{d}z \\
&= \int_0^{+\infty} \left[(1+kz)e^{-(1+k)z} - (1+mz)e^{-(1+m)z}\right]\mathrm{d}z \\
&= \frac{1}{1+k} + \frac{k}{(1+k)^2} - \frac{1}{1+m} - \frac{m}{(1+m)^2} \\
&= (1-\alpha) + \alpha(1-\alpha) - (1-\beta) - \beta(1-\beta) = \beta^2 - \alpha^2,
\end{aligned}$$

where the last line follows from replacing $k = \frac{\alpha}{1-\alpha}$ and $m = \frac{\beta}{1-\beta}$. Since

$$E\left(P\mathbb{1}_{\{\gamma\le U\le\delta\}}\right) = \frac{\delta^2-\gamma^2}{2},$$

we multiply the two terms and get the result. □

Lemma A.3. *The expected value of $Z(1-P)$ over all games $G(A,Z;P)$ in K is*

$$E_K\left[Z(1-P)\right] = \frac{(\beta-\alpha)(2-\beta-\alpha)(\delta-\gamma)(2-\delta-\gamma)}{2}$$

Proof. Let

$$V = Z(1-P)\mathbb{1}_{\{K\}} = Z(1-P)\mathbb{1}_{\{\alpha\le\Pi\le\beta\}}\mathbb{1}_{\{\gamma\le P\le\delta\}}$$

Clearly, $\alpha \le \Pi \le \beta$ if and only if $\frac{1-\beta}{\beta}A \le Z \le \frac{1-\alpha}{\alpha}A$. Let $k = \frac{1-\beta}{\beta} < \frac{1-\alpha}{\alpha} = m$. By independence,

$$E(V) = E\left(Z\mathbb{1}_{\{kA\le Z\le mA\}}\right)\cdot E\left((1-P)\mathbb{1}_{\{\gamma\le U\le\delta\}}\right)$$

Since A and Z are identically distributed, the first term on the right-hand side yields

$$\begin{aligned}
E\left(Z\mathbb{1}_{\{kA\le Z\le mA\}}\right) &= E\left(A\mathbb{1}_{\{kZ\le A\le mZ\}}\right) \\
&= \frac{1}{1+k} + \frac{k}{(1+k)^2} - \frac{1}{1+m} - \frac{m}{(1+m)^2} \\
&= \beta + \beta(1-\beta) - \alpha - \alpha(1-\alpha) = (\beta-\alpha)(2-\beta-\alpha)
\end{aligned}$$

where the last line follows from replacing $k = \frac{1-\beta}{\beta}$ and $m = \frac{1-\alpha}{\alpha}$. Since

$$E\left((1-P)\mathbb{1}_{\{\gamma\le P\le\delta\}}\right) = \frac{(\delta-\gamma)(2-\delta-\gamma)}{2}$$

we multiply the two terms and get the result. □

We prove that each party has a dominant choice for almost every cell $K = (\alpha, \beta) \times (\gamma, \delta)$. We begin with two lemmata characterizing best replies.

Lemma A.4. *Suppose that i's opponent plays H over the cell $K = (\alpha, \beta) \times (\gamma, \delta)$. Then i's best reply over K is H if and only if*

$$\frac{\beta + \alpha}{2 - (\beta + \alpha)} \geq \frac{2 - (\delta + \gamma)}{\delta + \gamma} \tag{1}$$

Proof. Under the strategy profile (H, H), by Lemma **??** the expected payoff for i over K is

$$E_K \left(AP \right) = \frac{(\beta^2 - \alpha^2)(\delta^2 - \gamma^2)}{2}$$

Under the strategy profile LH, by Lemma **??** the expected payoff for i over K is

$$E_K \left[Z(1 - P) \right] = \frac{(\beta - \alpha)(2 - \beta - \alpha)(\delta - \gamma)(2 - \delta - \gamma)}{2}$$

Hence, H is preferred to L if and only if

$$\frac{(\beta^2 - \alpha^2)(\delta^2 - \gamma^2)}{2} \geq \frac{(\beta - \alpha)(2 - \beta - \alpha)(\delta - \gamma)(2 - \delta - \gamma)}{2}$$

Simplifying and rearranging, we obtain the inequality in (**??**). □

Lemma A.5. *Suppose that i's opponent plays L over the cell $K = (\alpha, \beta) \times (\gamma, \delta)$. Then i's best reply over K is H if and only if (**??**) holds.*

Proof. Under the strategy profile HL, by Lemma **??** the expected payoff for i over K is $E_K \left[-Z(1 - P) \right]$. Under the strategy profile LH, by Lemma **??** the expected payoff for i over K is $E_K \left(-AP \right)$. Hence, H is preferred to L if and only if $E_K \left(AP \right) \geq E_K \left[-Z(1 - P) \right]$, which leads again to the inequality in (**??**). □

Proposition A.6. *Given a cell $K = (\alpha, \beta) \times (\gamma, \delta)$, let $\overline{\pi} = (\beta + \alpha)/2$ and $\overline{p} = (\delta + \gamma)/2$. Then i has a (strictly) dominant strategy if $\overline{\pi} + \overline{p} \neq 1$ and is indifferent between H and L if equality holds. Moreover, the dominant strategy is H if $\overline{\pi} + \overline{p} > 1$, and it is L if $\overline{\pi} + \overline{p} < 1$.*

Proof. If we multiply and divide by 2 the expressions on either side of the inequality in (**??**), we find

$$\frac{\overline{\pi}}{1 - \overline{\pi}} \geq \frac{1 - \overline{p}}{\overline{p}}$$

that, upon rearrangement, is equivalent to $\overline{\pi} + \overline{p} \geq 1$. Then the result follows immediately from Lemma **??** and Lemma **??**. □

Proposition A.7. *Given the threshold pair $(\hat{\pi}, \hat{p})$, the expected payoff to each agent under the (unique) rational rule of behavior is*

$$\frac{1}{2} - \pi^2 p^2 \qquad \textit{when } \hat{\pi} + \hat{p} > 1$$

and

$$\frac{1}{2} - \pi^2 - p^2 + \pi^2 p^2 \qquad \textit{when } \hat{\pi} + \hat{p} < 1$$

Proof. Suppose $\hat{\pi} + \hat{p} > 1$. The dominant rule of behavior is OR, yielding a random payoff AP in situations S_1, S_2, S_4 and $-AP$ in S_3. Lemma **??** gives the expected payoffs over these four situations. Taking their sum yields

$$\frac{(1-\pi^2)(1-p^2)}{2} + \frac{\pi^2(1-p^2)}{2} - \frac{\pi^2 p^2}{2} + \frac{p^2(1-\pi^2)}{2} = \frac{1 - 2\pi^2 p^2}{2}$$

Suppose instead $\hat{\pi} + \hat{p} < 1$. The dominant rule of behavior is AND, yielding a random payoff AP in situation S_1, and $-AP$ in situations S_2, S_3, S_4. Proceeding similarly, we find

$$\frac{(1-\pi^2)(1-p^2)}{2} - \frac{\pi^2(1-p^2)}{2} - \frac{\pi^2 p^2}{2} - \frac{p^2(1-\pi^2)}{2} = \frac{1 - 2\pi^2 - 2p^2 + 2\pi^2 p^2}{2}$$

□

Proposition A.8. *When the OR rule prevails, there is fog of conflict if*

$$\hat{\pi}^2 \cdot \hat{p}^2 > 1/6 \tag{2}$$

and fog of cooperation if the opposite (strict) inequality holds.
When the AND rule prevails, there is fog of conflict if

$$\hat{\pi}^2 + \hat{p}^2 - \hat{\pi}^2 \cdot \hat{p}^2 > 1/6 \tag{3}$$

and fog of conflict if the opposite (strict) inequality holds.

Proof. By Proposition **??**, the expected payoff to a party under the benchmark is 1/3. Given a threshold pair $(\hat{\pi}, \hat{p})$, Proposition **??** characterizes the expected payoff to a party under the rational rule of behavior. There is fog of conflict (or cooperation) when this payoff is lower (or greater) than 1/3. If $\hat{\pi} + \hat{p} > 1$ and the OR rule applies, the expected payoff is

$\frac{1}{2} - \pi^2 p^2$. This is lower (or greater) than 1/3 when (**??**) (or its opposite) holds. The argument is similar if $\hat{\pi} + \hat{p} < 1$ and the AND rule applies, taking into account that the expected payoff is $\frac{1}{2} - \pi^2 - p^2 + \pi^2 p^2$. □

Proposition A.9. *Suppose that the parties perceive two games* $(\pi_1, p_1) \neq (\pi_2, p_2)$ *in* $\mathcal{G}$ *as distinct. Then cooperation on* (H, H) *for all games* (π, p) *can be supported by a subgame perfect equilibrium based on a Nash reversion trigger strategy if and only if*

$$\delta \geq \frac{24}{25} \tag{4}$$

Proof. Cooperation on (H, H) is a dominant strategy when the one-shot game maps to a pair (π, p) such that $\pi + p \geq 1$. This occurs with probability 1/2. Consider now the complementary event $D^- = \{(\pi, p) : \pi + p < 1\}$ and assume that the agents perceive (π, p) in D^-. Cooperation on (H, H) can be supported only if the long-term benefit from choosing H is never smaller than the short-term temptation to deviate and pick L. We compare the short-term temptation against the long-term benefit.

Conditional on the knowledge of (π, p), the short-term temptation ST is the difference in expected payoffs from choosing L or H when the other party plays H:

$$\text{ST} = E\left[(1-p)Z - pA \mid \pi, p\right]$$

Using $A/(A + Z) = \pi$ and $E(Z) = 1$, we obtain that the short-term temptation is

$$\text{ST} = E\left[(1-p)Z - p\frac{\pi Z}{1-\pi} \mid \pi\right] = \frac{1-\pi-p}{1-\pi}$$

The long-term benefit LB is the discounted sum of the (expected) incremental payoffs from sustaining cooperation on (H, H) against shifting to (L, L) on D^-. Since this occurs with probability 1/2, we have

$$\text{LB} = \frac{1}{2}\left(\frac{\delta}{1-\delta}\right) E\left[PA\mathbb{1}_{D^-}\right] = \frac{1}{2}\left(\frac{\delta}{1-\delta}\right) E\left[PA\mathbb{1}_{\{\Pi + P < 1\}}\right]$$

and, by the same technique used in the proof of Proposition **??**, we find

$$\text{LB} = \frac{1}{2}\left(\frac{\delta}{1-\delta}\right) E\left[P(1-P)^2\right] = \frac{1}{2}\left(\frac{\delta}{1-\delta}\right)\frac{1}{12}$$

Cooperation can be supported when the long-term benefit LB is not smaller than the short-term temptation ST; that is, if

$$\frac{1}{24}\left(\frac{\delta}{1-\delta}\right) \geq \frac{1-\pi-p}{1-\pi}$$

or, equivalently, if

$$\delta \geq \frac{24(1-\pi-p)}{1-\pi+24(1-\pi-p)}$$

This follows for any (π, p) in D^- if and only if (**??**) holds. □

Proposition A.10. *Suppose that the parties perceive two games G_1 and G_2 with $(a_1, z_1; p_1) \neq (a_2, z_2; p_2)$ as distinct. Then cooperation on (H, H) for all games cannot be a.s. supported by a subgame perfect equilibrium based on a Nash reversion trigger strategy.*

Proof. Given δ, for any game that maps to (π, p) with $\pi + p < 1$, the long-term benefit from playing H is $\delta/[6(1-\delta)]$ while the short-term temptation is $(1-p)z - pa$: cooperation fails whenever we draw a game such that $(1-p)z - pa > \delta/[6(1-\delta)]$. Since Z has unbounded support and $\delta < 1$, there is a strictly positive probability that cooperation fails for any combination of a, p, δ. □

Proposition A.11. *Suppose $x > 1/2$. Cooperation on (H, H) in all situations can be supported by a subgame perfect equilibrium based on a Nash reversion trigger strategy if and only if*

$$\delta \geq \frac{2-2x}{2-2x+x^4} \tag{5}$$

Proof. Cooperation on (H, H) is a dominant strategy for the one-shot games associated with situations S_1, S_2, S_4. So we only need to compare the short-term temptation in S_3 against the long-term benefit. The one-shot game perceived by the agents in S_3 is depicted on the left of Figure **??**. The short-term temptation in S_3 is the difference in payoffs from playing L instead of H when the other party plays H; that is,

$$\frac{x^2(2-x)^2}{2} - \frac{x^4}{2} = 2x^2(1-x)$$

The long-term benefit in S_3 is the discounted sum of the (expected) incremental payoffs from sustaining cooperation on (H, H). Since this situation occurs with probability x^2, we find

$$\frac{\delta x^2}{1-\delta}\left[\frac{x^4}{2} - \frac{-x^4}{2}\right] = \frac{\delta x^6}{1-\delta}$$

	H	L
H	$x^4/2$	$-x^2(2-x)^2/2$
L	$x^2(2-x)^2/2$	$-x^4/2$

S_3

	H	L
H	$x^2(1-x^2)/2$	$-x(2-x)(1-x)^2/2$
L	$x(2-x)(1-x)^2/2$	$-x^2(1-x^2)/2$

S_2

Figure 14: The one-shot games associated with S_3 (left) and S_2 (right) for $\hat{\pi} = \hat{p} = x$.

Cooperation can be supported when the long-term benefit is not smaller than the short-term temptation

$$\frac{\delta x^6}{1-\delta} \geq 2x^2(1-x)$$

which yields (**??**). □

Proposition A.12. *Suppose $x < 1/2$. Cooperation on (H, H) in all situations can be supported by a subgame perfect equilibrium based on a Nash reversion trigger strategy if and only if*

$$\delta \geq \begin{cases} \dfrac{(1-x)(1-2x)}{1-3x+4x^2-2x^3-2x^4+3x^5} & \textit{for } 0 < x \leq \dfrac{1}{4} \\[2ex] \dfrac{2-2x}{2-2x^2-2x^3+3x^4} & \textit{for } \dfrac{1}{4} < x < \dfrac{1}{2} \end{cases} \tag{6}$$

Proof. Cooperation on (H, H) is a dominant strategy only for the one-shot game associated with situation S_1, so we need to check the other three situations. By the simplifying assumption $\hat{\pi} = \hat{p} = x$, the payoffs for S_2 and S_4 are identical and we can restrict attention to S_3 and S_2, depicted in Figure **??**. The short-term temptation is $2x^2(1-x)$ in S_3 and $x(1-x)(1-2x)$ in S_2. Comparing the two, the short-term temptation is greater in S_3 for $x > 1/4$ and in S_2 for $x < 1/4$. Thus, we study the two cases separately for $x \gtrless 1/4$. The long-term benefit is the same across the three situations S_2, S_3, S_4, which occur with probability $x(1-x), x^2, x(1-x)$ respectively; thus, the long-term benefit is

$$\frac{\delta}{1-\delta}\left[2x(1-x)x^2(1-x^2) + x^2 \cdot x^4\right] = \frac{\delta}{1-\delta}\left[x^3\left(2-2x-2x^2+3x^3\right)\right]$$

When $x > 1/4$, the greater temptation is in S_3 and thus cooperation can be supported if

$$\frac{\delta}{1-\delta}\left[x^3\left(2-2x-2x^2+3x^3\right)\right] \geq 2x^2(1-x)$$

which can be rewritten as

$$\delta \geq \frac{2-2x}{2-2x^2-2x^3+3x^4} \tag{7}$$

When $x < 1/4$, the greater temptation is in S_2 and thus cooperation can be supported if

$$\frac{\delta}{1-\delta}\left[x^3\left(2-2x-2x^2+3x^3\right)\right] \geq x(1-x)(1-2x)$$

which yields

$$\delta \geq \frac{(1-x)(1-2x)}{1-3x+4x^2-2x^3-2x^4+3x^5} \tag{8}$$

Both right-hand sides for (**??**) and (**??**) are decreasing and take value $384/475 \approx 0.8084$ at $x = 1/4$, yielding (**??**). □