What Situation Is This?

Shared Frames and Collective Performance*

Robert Gibbons
Massachusetts Institute of Technology
rgibbons@mit.edu

Marco LiCalzi
Università Ca' Foscari Venezia
licalzi@unive.it

Massimo Warglien
Università Ca' Foscari Venezia
warglien@unive.it

June 2020

Abstract. We study agents who distill the complex world around them using cognitive frames. The frame reduces a continuum of interactive decisions into a finite number of categories, which we call "situations." We assume that agents from the same organizational culture share the same frame and analyze how the frame affects their collective performance. In one-shot and repeated interactions, the frame causes agents to be either better or worse off than if they could perceive the environment in full detail. In repeated interactions, the frame from the organizational culture is as important as agents' patience in determining the outcome: for a fixed discount factor, when all agents choose what they perceive as their best play, there remain significant performance differences induced by different frames. We distinguish between incremental versus radical changes in frames by their effects on agents' behavior. We develop a model of category formation to illustrate some challenges of modifying the agents' frame to improve their collective performance. When parties rely on different frames, we differentiate between incremental versus radical discord, and we sketch one path to resolving discord.

Keywords: categorization, cognitive frame, culture, leadership, performance.

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

*We thank Andreas Blume, Vincent Crawford, Emir Kamenica, Katherine Kellogg, Anton Kolotilin, Margaret Meyer, Wanda Orlikowski, Alessandro Pavan, Andrea Prat, Phil Reny, Joel Sobel, Marco Tolotti, Catherine 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, from MIT Sloan's Program on Innovation in Markets and Organizations, and from COPE, Danish Council for Independent Research.

1 Introduction

This paper studies how shared cognition can create a link between organizational culture and organizational performance. We analyze how different cultures carrying different cognitive representations support stable differences in performance, and we explore challenges faced by a leader who seeks to change an organization's frame in order to improve performance.

In modeling organizational culture as shared cognition, we follow a long tradition. Pettigrew describes organizational culture as the "system of terms, forms, categories and images that interprets a people's own situation to themselves" (1979: 574). Similarly, Schein argues that organizational culture creates "mindsets and frames of reference ... [that are] invisible and to a considerable degree unconscious" (1985; fourth edition, 2010: 14). And a more specific tradition supports our particular modeling approach, viewing shared categorization as a building block of culture. For example, Patterson asserts that "The basis of all cultural knowledge is our capacity to categorize." (2014: 8) and Denzau and North discuss "a culturally provided set of categories" (1994: 5).

Our model assumes that an organization's members share a *cognitive frame* that distills the complex world around them into a finite number of categories we call *situations*. Consistent with much literature in cognitive science, agents in our model are unaware that they distill the world through the categories of such frames: they know only (i) the set of situations that could arise and, when an underlying state of the world is realized, (ii) which of these situations has arisen. In this sense, agents in our model ask themselves "What situation is this?" (Goffman, 1974; March and Olsen, 1983) and reach answers determined by their frame.

We close the model by assuming not only that organizational culture determines how agents see the world, but also that agents act rationally given what they perceive. The agents' shared frame thus determines a mapping from situations to their optimal actions. We obtain the following results.

First, in a one-shot interaction, the coarse representation induced by the categories of a frame results in a unique equilibrium that can either decrease or increase the parties' payoffs, compared to having full information about the environment. We say that an organization's shared frame may induce either a fog of conflict or a fog of cooperation. This initial result is consistent with the argument that differences in organizational performance may stem from differences in cognitive frames.

Second, in a repeated interaction, standard arguments from repeated games allow the parties to increase their payoffs above the one-shot level if they are sufficiently patient. We focus on the opposite comparative static: fix the parties' discount factor and analyze how their

frame affects their *highest equilibrium payoffs* in a repeated interaction. Holding discounting constant, there are again frames under which the parties' payoffs are higher (or lower) than under full information. Importantly, this is not a standard result about multiple equilibria, as follows.

Kreps (1990) proposed long ago that different equilibria in a repeated game might correspond to different corporate cultures (shared understandings of "how we do things around here") associated with different performance levels across plants and firms. While highly suggestive, there is a concern with modeling performance differences as resulting from different equilibria in a given game: low performers know that better equilibria exist, and yet the model gives these parties no way to try to reach a better equilibrium and 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.

Whether in the unique equilibrium of our static analysis or the best equilibrium in our repeated analysis, a unifying feature is that a difference in frames can cause parties to (a) ascribe a state of the world to different situations, and even (b) see different actions as optimal in situations that they describe equivalently. As a result, if we imagine low performers visiting a high-performing organization, the low performers may see their hosts achieving higher performance in ways that the low performers cannot understand how to imitate. This inimitability is necessary if an organizational culture is to create competitive advantage (Barney, 1986).

Building on these classic contributions by Kreps and by Barney, our third set of results concerns the consequences and mechanisms of attempting to change cognitive frames. Regarding the consequences, we distinguish between *incremental change*, when the boundaries of situations change but the parties' optimal actions in given situations do not, versus *radical change*, where both the boundaries and the optimal actions vary. We show that, if revising the perceived boundaries of situations is more rapid than adjusting perceived optimal actions, radical change can induce either worse-before-better or better-before-worse performance paths (Repenning and Sterman, 2002).

Regarding the mechanisms, we follow an established literature in psychology (Medin and Schaffer, 1978; Nosofsky, 1986) and develop a model of category formation based on the exemplars stored in agents' memories. We then explore how a range of actions by a leader may change the organization's frame, and we uncover some trade-offs associated with attempting such changes.

Finally, we offer an initial sketch of parties with different frames interacting with each other. We distinguish between *incremental discord*, when the parties apply the same rules

of behavior to what they perceive, versus *radical discord*, where even their rules of behavior are different. We explore how, after discord arises, the parties may enter into a dialogue and coordinate their actions using sincere communication.

In summary, we see our model as exploring an integrated account of aspects of organizational culture, performance and leadership that have heretofore been considered separately (and typically not in formal models). We defer a detailed review of relevant literature to Section 7 after the development of our model, thereby facilitating the comparison with its precursors and alternatives.

2 The model

Interactions that involve either pure common interest or pure conflict are archetypes of social and economic life, and a rich repertoire of cultural and linguistic resources is available to represent these poles. Yet, many interactions mix collaborative and conflictual motives. In contexts of organizational interest—including team production, labor relations, strategic alliances, interactions along the supply chain, and many more—agents must interpret the combination of collaborative and conflictual motives before deciding how to behave.

The polar cases are cognitively simple: agents' interests are perfectly correlated, either positively or negatively, making it easy for agents to process such games. But interactions that mix collaborative and conflictual motives are more difficult to apprehend. As Schelling (1960) remarked, "[t]he difficulty is finding a sufficiently rich name for the mixed game in which there is both conflict and mutual dependence. [For example, ...] in the common-interest game we can refer to [the players] as "partners" and in the pure-conflict game as "opponents" or "adversaries"; but the mixed relation that is involved in wars, strikes, negotiations, and so forth, requires a more ambivalent term" (1960: 89). More recently, this middle ground has begun to be populated with terms such as "co-opetition" (Brandenburger and Nalebuff, 1996) and "frenemy."

In this section, we develop a simple model of how frames shape the interpretation of Schelling's mixed-motive interactions. Our model combines three features. First, we consider a space of games rather than a single game; second, agents have a coarse representation of their environment and act according to it; and third, they perceive the coexistence of motives as a blend of the common-interest and pure-conflict archetypes.

2.1 A space of games

Consider two archetypical games: a common-interest game (CI) on the left of Figure 1 and a zero-sum game (ZS) on the right, with r in $[0, 1]$. These games represent basic situations

	H	L
H	r, r	$0, 0$
L	$0, 0$	$-r, -r$

CI

	H	L
H	$0, 0$	$-(1-r), (1-r)$
L	$(1-r), -(1-r)$	$0, 0$

ZS

Figure 1: A common-interest game (left) and a zero-sum game (right).

of collaboration and conflict: for any r in $(0, 1)$, they have dominant strategies; that is, there are unequivocal motivations to cooperate (H) or compete (L), respectively.

These two archetypes provide building blocks for more complex interactions. For example, as we will see below, a prisoner's dilemma is a mixed-motive game that combines the two archetypes. More generally, we analyze a blended interaction involving two aspects: (1) the *reward* r from cooperation in the common-interest game and $1 - r$ from defection in the zero-sum game; and (2) the *prominence* p in $[0, 1]$ for the common-interest component and $1 - p$ for the zero-sum component.[1]

Figure 2 shows a typical game $G(r, p)$ generated from this blending of the archetypes in Figure 1. The nature (and attractiveness) of a blended interaction depends on both

	H	L
H	pr, pr	$-(1-p)(1-r), (1-p)(1-r)$
L	$(1-p)(1-r), -(1-p)(1-r)$	$-pr, -pr$

Figure 2: The payoff matrix for a generic game $G(r, p)$.

the prominence p of the common-interest archetype and the payoffs that each archetype contributes to the blend. For example, suppose $p = 1/2$ (e.g., the two archetypes are equally likely). Then, for $r > 1/2$ the blended game in Figure 2 is a common-interest game, but for $r < 1/2$ the blended game is a prisoners' dilemma. More generally, for $r + p > 1$, the blended game has common interests (and H is the dominant strategy); for $r + p < 1$ it is a prisoners' dilemma (where L is the dominant strategy). In other words, r and p jointly determine whether the cooperative or competitive motive prevails.

We assume that r and p are independently and uniformly distributed on $[0, 1]$ so the agents face a bidimensional *space of games* $\mathcal{G} = [0, 1]^2$, depicted in Figure 3. This space of games captures a wide range of interactions. As just one example, in relations between suppliers and assemblers of complex components, a component may blend standard and dedicated elements,

[1] *Prominence* is the weight attributed by the agent to the common-interest archetype when perceiving a blended interaction. More concretely, our analysis can be seen as studying games where agents believe that (after they move) Nature chooses the common-interest game with probability p and the zero-sum game with probability $1 - p$.

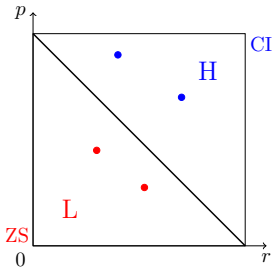


Figure 3: The space $\mathcal{G}$ of games, with dominant strategy H or L.

where the former are associated with a competitive interaction but the latter with common interests (p), and the value added by dedicated elements (r) also matters in interpreting the blended interaction and deciding how to play.

We use this space of games to model interactions between agents who have limited ability to discriminate among the blended games in $\mathcal{G}$. Strictly as a benchmark for comparing our model's results, we conclude this subsection by briefly considering the case where each party can discriminate any game g from $\mathcal{G}$ and then play the appropriate dominant strategy for g. Then each party's expected payoff is 1/6; see Proposition A.1 in the Online Appendix, where we have collected propositions and proofs.

2.2 Coarse perception

We now impose our assumption that the parties have limited ability to discriminate among games in $\mathcal{G}$. In particular, we henceforth assume that each dimension of the space $\mathcal{G}$—the reward r in $[0, 1]$ and the prominence p in $[0, 1]$—is too rich to allow either party to perceive all its elements as distinct. Instead, each agent apprehends each dimension using a finite partition. For simplicity, we work with binary categorizations, respectively defined by the thresholds $\hat{r}$ and $\hat{p}$. Thus, an agent categorizes r as High (h) if $r > \hat{r}$ and Low (ℓ) if $r < \hat{r}$; similarly, p is High (h) if $p > \hat{p}$ and Low (ℓ) if $p < \hat{p}$.

An agent with binary categorizations for r and p perceives four cells, as depicted in Figure 4. 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 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 r and p are perceived as h. That is, S_1 involves both high reward and high prominence, close to the common-interest archetype with $r = p = 1$. The other three situations have analogous interpretations.

The *frame* of an agent is the collection of the situations that she perceives, identified by

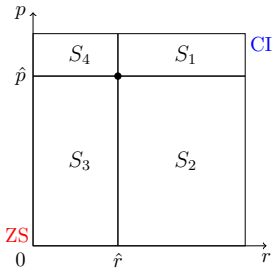


Figure 4: A categorization of the game space $\mathcal{G}$ into four situations.

the threshold pair $(\hat{r}, \hat{p})$. We assume that an agent is unaware that she is framing. The model-builder, not the agent, knows that the agent (a) categorizes games and (b) does so via the threshold pair $(\hat{r}, \hat{p})$.

Until Section 6, we assume that two interacting parties share the same frame $(\hat{r}, \hat{p})$, with $0 < \hat{p}, \hat{r} < 1$. This is how we model (admittedly, quite starkly) the idea that these two parties have been shaped by the same organizational culture. A more realistic assumption might be that individuals' frames are more highly correlated within organizations than between organizations, but not necessarily perfectly correlated for individuals within a given organization. We offer an initial sketch of this alternative case in Section 6. Until then we assume that two parties from a given organization share the same frame: in each of the four situations associated with the frame, the two parties perceive a single 2×2 symmetric game with payoffs equal to the expected payoffs from all the games ascribed to that situation. In this sense, 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 4.

We label the northeast and southwest situations S_1 and S_3 *consonant*, because their descriptors r and p are both high or both low: the reward r and the prominence p of cooperation are aligned. As we will see, each consonant situation has an unambiguous interpretation, with clear implications for agents' actions. In contrast, we say that the two situations S_2 and S_4 are *dissonant* because their descriptors are misaligned: one is high and the other is low. As we will see, dissonant situations have ambivalent interpretations, whose resolution may diverge under (even slightly) different frames, producing different implications for agents' actions.

3 One-shot interaction

This section considers a one-shot interaction between two parties under a shared frame $(\hat{r}, \hat{p})$. We assume that their behavior is rational conditional on their frame: once they have interpreted a given situation, they are rational players within their interpreted world.

The expected payoffs to the first party (rescaled by a factor of 4) for each of the four situations perceived under the frame $(\hat{r}, \hat{p})$ are shown in Figure 5.

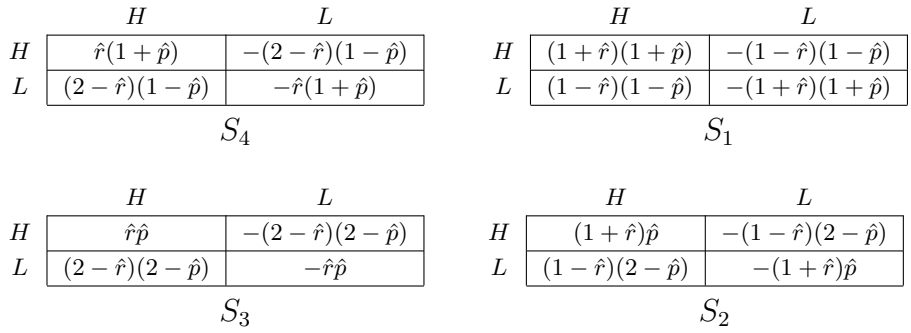


Figure 5: Perceived payoffs for the Row player in the four situations under the frame $(\hat{r}, \hat{p})$.

Conditional on the frame, the agents have correct beliefs about the distribution of payoffs in each situation; see the notion of *interpreted signal* in Hong and Page (2009). After playing a perceived situation, the parties receive the payoffs associated with the actual game $G(r, p)$ that was drawn and ascribe the difference between expected payoff and realized payoff to noise.

Rational behavior in the two consonant situations is unequivocal. Figure 5 shows that situation S_1 is always perceived as a CI game under any frame $(\hat{r}, \hat{p})$, so H (cooperate) is the dominant strategy. Similarly, situation S_3 is always perceived as a PD game, so L (defect) is the dominant strategy. Regardless of the frame, the rational behavior is to play H in S_1 and L in S_3. Intuitively, the two consonant situations are adjacent to the common-interest and zero-sum archetypes, respectively: their interpretation (and the resulting behavior) matches their close proximity to an archetype.

Assuming that the frame satisfies $\hat{r} + \hat{p} \neq 1$, there is also a unique dominant strategy for the dissonant situations S_2 and S_4. This is characterized in the next proposition, which is a corollary of Proposition A.4 in the Online Appendix.

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

Unlike consonant situations, the dominant strategy in dissonant situations depends on the frame. Intuitively, the descriptors are misaligned because a dissonant situation exhibits Schelling's mixed motives: the agent's frame resolves the ambivalent interpretation in favor of one strategy or the other.

Combining the dominant strategies over the four situations, we find two rules of behavior, shown in Figure 6. The first rule, depicted on the left, is optimal if $\hat{r} + \hat{p} > 1$: play H in

any situation except S_3, and then play L; we call this rule *Cooperation by Default* because it prescribes playing H unless both r and p are low. The second, shown on the right, is optimal if $\hat{r} + \hat{p} < 1$: play H only in S_1 and otherwise play L; we call this rule *Defection by Default* because it prescribes playing L unless both r and p are high.

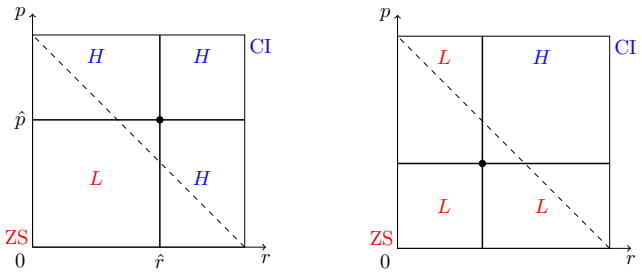


Figure 6: Cooperation by Default (left) prescribes H unless both p and r are low; Defection by Default (right) prescribes L unless both p and r are high.

These rules of behavior are useful in our exposition below. Note that "Default" refers to the situations, not to their probability of occurring. That is, Cooperation by Default prescribes H unless a special situation (S_3) holds, and Defection by Default prescribes L unless a special situation (S_1) holds. The overall probabilities with which H and L are played depend on further details, beyond these rules of behavior.

Since the frame is shared and payoffs are symmetric, the parties will play the same strategy in a given situation. If $\hat{r} + \hat{p} > 1$, they will cooperate by default, playing (H, H) in all situations except (L, L) in S_3; if $\hat{r} + \hat{p} < 1$, they will defect by default, playing (L, L) in all situations except (H, H) in S_1. In sum, different frames can induce different strategy profiles when parties encounter dissonant situations.

Having computed optimal strategies, we next analyze how the parties' expected payoffs depend on the thresholds $(\hat{r}, \hat{p})$ of their shared frame. First, payoffs change continuously in $(\hat{r}, \hat{p})$ if the variation in thresholds does not change the parties' rule of behavior; second, if the rule of behavior switches, then there is a discontinuous change in payoffs.

Proposition A.5 gives the expected payoff to each party as a function of $\hat{r}$ and $\hat{p}$. As an example, suppose $\hat{r} = \hat{p} = x$ so that a change in x makes both thresholds shift in lockstep. The parties Cooperate by Default (denoted CbD) for $x > 1/2$ and Defect by Default (DbD) for $x < 1/2$. Figure 7 shows the payoff to each party as a function of x. Within each default-rule region, payoffs continuously decrease in x. On the other hand, moving x rightward across $1/2$ implies an abrupt increase in payoffs, as the parties switch from Defection by Default to Cooperation by Default. Nonetheless, depending on x, the former rule may outperform the latter.

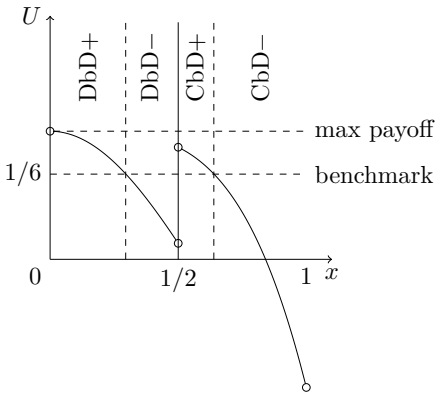


Figure 7: Payoffs as a function of x when the frame is $(\hat{r}, \hat{p}) = (x, x)$. (The labels DbD and CbD identify the dominant strategy; the modifier $+/-$ denotes payoffs higher or lower than the benchmark.)

Figure 7 also shows that framing games as situations can either help or hurt the parties' payoffs, relative to the benchmark case where each game is perceived as distinct: the benchmark payoff of 1/6 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 $-$), under which agents achieve lower expected payoffs than they would under full information, or a *fog of cooperation* (marked $+$), under which expected payoffs are higher. Note that either fog can occur under either rule of behavior, so frames evidently do more than determine rules of behavior.

Beyond the special case of $\hat{r} = \hat{p} = x$ shown in Figure 7, we can identify which frames generate which kind of fog. See the Online Appendix, where Proposition A.6 states a formal characterization followed by a visual summary. The main message is that a tiny change in the threshold(s) that causes a switch in the rule of behavior yields an abrupt change in payoffs. This discontinuity motivates part of our discussion about changing frames in Section 5.

As one way to summarize this static model, imagine two parties who share a low-performing frame visiting two other parties who share a high-performing frame. All parties perceive situations in terms of their own frames, and the low performers observe the actions chosen by the high performers. For now, we simply consider what the low performers will see and what they might then infer; we defer discussions of attempts to change frames and of parties with different frames until Sections 5 and 6, respectively.

Consider the discontinuity at $x = 1/2$ in Figure 7, and suppose that the low- and high-performing frames have $x_\ell < 1/2 < x_h$, with x_ℓ and x_h close but on different sides of 1/2. The high-perfoming frame supports Cooperation by Default, as in the left panel of Figure 6,

whereas the low-performing frame supports Defection by Default, as in the right panel of Figure 6. There are then some games (r, p) in $\mathcal{G}$ that the low and high performers see as different situations, but the more important difference is that the low performers see dissonant situations as PD games and hence choose (L, L), whereas the high performers see dissonant situations as CI games and so choose (H, H). The low performers thus will be mystified by the visit: when they see CI, they observe their hosts playing (L, L) occasionally; and when they see PD, they observe their hosts playing (H, H) frequently.

If the low and high performers discuss what they saw or why the high performers acted as they did, the low performers will occasionally discover that the high performers saw different situations, and they will frequently discover that the high performers perceived the same situation but considered different actions to be optimal. Neither discovery would necessarily make them aware that anyone perceives the world coarsely (not to mention differently so). In short, the difference in cognitive frames may be an inimitable source of competitive advantage (Barney, 1986).

To summarize this section, we see our static model as a small but novel contribution towards understanding widespread evidence of differences in cooperation.[2] It is common to interpret such differences in cooperation as arising from differences in preferences; our model provides a complementary explanation based on differences in cognition—specifically, differences in interpretation.[3] While we do not expect our simple model to capture this wide range of empirical evidence, we believe that cognition (and especially interpretation) can offer a promising explanatory approach.

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 consonant situation S_3 is perceived as a PD. Furthermore, if $\hat{r} + \hat{p} < 1$ then the dissonant 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 $(\hat{r}, \hat{p})$. As in the static model,

[2] Some of the field evidence points to differences in cooperation during evolution (Boyd and Richerson, 2009) and among cultures (Henrich et al., 2005), communities (Ostrom, 1990), firms (Leibenstein, 1982), organizations (Schein, 1985), and teams (Cole, 1991).

[3] Experiments show that cultural frames cause individuals to perceive situations as "cooperative" or "competitive" (Keller and Lowenstein, 2011) and how inducing different frames affects cooperation levels (Pruitt, 1970; Liberman et al., 2004; Ellingsen et al., 2012).

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 $(1-\hat{r})(1-\hat{p})$, situation S_2 with probability $(1-\hat{r})\hat{p}$, situation S_3 with probability $\hat{r}\hat{p}$, and situation S_4 with probability $\hat{r}(1-\hat{p})$. 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 consider subgame-perfect equilibria where an unexpected defection (i.e., playing L when H was expected in a PD situation) triggers Nash reversion thereafter (i.e., defection in all future PD situations and cooperation in all future CI situations).

There are two cases of interest. The first is *Full Cooperation*, when agents play (H, H) across all situations. The second case is *Improved Cooperation*, when the static model leads to Defection by Default but the repeated interaction can support Cooperation by Default; that is, in the repeated game players switch from defection to cooperation when facing dissonant situations, but not when facing situation S_3.

Recall that the rule of behavior in the static model is Cooperation by Default if $\hat{r}+\hat{p} > 1$ and Defection by Default if $\hat{r}+\hat{p} < 1$. As above, we reduce the number of parameters by assuming $\hat{r} = \hat{p} = x$; then the two rules obtain for $x > 1/2$ and $x < 1/2$. When $x > 1/2$, the static model leads to Cooperation by Default, and we study when the repeated interaction may support Full Cooperation. When $x < 1/2$, the static model leads to Defection by Default, and we study when the repeated interaction may support either Full Cooperation or Improved Cooperation.

Consider $x > 1/2$: the only situation perceived as a PD is S_3. The frame generated by x has three effects. First, there is a probability x^2 that the PD situation occurs in the future: the greater is x, the larger the PD situation looms. Second, when the PD situation does occur and the other party is expected to cooperate, there is a temptation to play defection (L) instead of cooperation (H) that is decreasing in x. Finally, the threat of a long-term payoff loss from Nash reversion after defection is increasing 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. Nevertheless, the familiar intuition that a sufficiently high δ supports Full Cooperation survives: if

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

then there is a Nash-reversion equilibrium where the parties play H in situation S_3; see Proposition A.7. This is illustrated in Figure 8 for $x > 1/2$: given δ and x, either the parties can sustain Full Cooperation (FC) across all situations, or they Cooperate by Default (CbD) which, given $x > 1/2$, is the rule from the static game. In particular, for $\delta \geq 16/17$, sustaining

Full Cooperation is possible for any value $x > 1/2$.

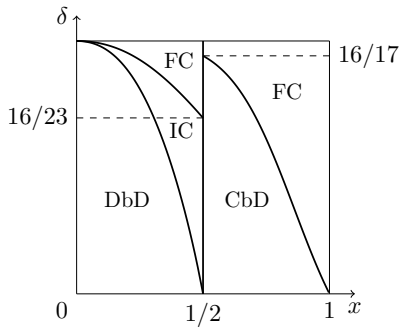


Figure 8: Best feasible cooperation in the repeated interaction. (The labels DbD, CdB, IC, FC identify the strategy profiles yielding the highest payoffs under Nash reversion for a frame x and a discount factor δ.)

Consider now $x < 1/2$, when the static model yields Defection by Default. Proposition A.8 demonstrates a richer result. Let

$$\overline{\delta}_1(x) = \frac{2-2x}{2-2x+2x^2-x^4} \quad \text{and} \quad \overline{\delta}_2(x) = \frac{1-2x}{1-2x+2x^2-2x^4}.$$

In a repeated interaction, Nash reversion may be used to support Full Cooperation (FC) across all situations if $\delta \geq \overline{\delta}_1(x)$ and to support Improved Cooperation (IC) if $\delta \geq \overline{\delta}_2(x)$. Clearly, Full Cooperation is harder to achieve because $\overline{\delta}_1(x) \geq \overline{\delta}_2(x)$ for all $x < 1/2$. If neither inequality holds, the parties are stuck with Defection by Default (DbD) as in the static model.

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 fix a discount factor $16/23 < \delta < 16/17$. Then Figure 8 shows that, as x progresses from 0 to 1, the best outcome that parties can sustain in a repeated interaction changes from Defection by Default to Improved Cooperation (i.e., Cooperation by Default) to Full Cooperation, then back to to Cooperation by Default and to Full Cooperation again—all for the same discount factor.

For comparison, consider the benchmark case where the parties can distinguish all the games in $\mathcal{G}$. Proposition A.9 shows that in the benchmark case Nash reversion supports Full Cooperation if $\delta \geq 12/13$. Because this value of δ is between 16/23 and 16/17, in repeated interaction there are values of x where coarse perception creates a fog of cooperation (Full

Cooperation is feasible under framing but not without) as well as values of x where it creates a fog of conflict (Full Cooperation is not feasible under framing but is without).

In short, even with a shared discount factor, differences in frames can cause parties to disagree about the best equilibrium feasible in the repeated game. For example, in Figure 8, consider a discount factor between 16/23 and 16/17 and two values of x below 1/2—one value of x such that those parties see Improved Cooperation as the best feasible equilibrium in the repeated game and another (larger) value of x such that those parties see Full Cooperation as feasible. In this example, all parties believe that there is an equilibrium in the repeated game that outperforms spot play, but the former think that (H, H) cannot be sustained in situation S_3, while the latter think that it can. The low-performers might diagnose this disagreement about equilibrium strategies as disagreement about the probabilities of (or the payoffs in) the situation S_3. Thus, as in our static model, disagreement about equilibrium in the repeated interaction might *not* cause the parties to imagine that they perceive the world coarsely (not to mention differently so).

We viewed our static model as a small step towards understanding a broad set of findings concerning widespread evidence of differences in cooperation. In contrast, we see our repeated model as a larger step towards a much more specific goal: moving beyond the intriguing suggestion by Kreps (1990) that different equilibria in a repeated game might correspond to different corporate cultures. There is a concern with modeling performance differences where low performers know that better equilibria exist, and yet the model gives these parties no way to try to reach a better equilibrium and 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, to which we now turn.

5 Changing frames

Our basic model assumes that, within a given organization or group, frames are (a) fixed and (b) shared. This section relaxes the first assumption and explores some difficulties in changing frames to improve performance. Section 6 relaxes the second and sketches how parties with different frames might resolve their discord.

Within this section, we first consider the consequences of frame change. In particular, in Section 5.1 we distinguish between *incremental* versus *radical* change in the frame: the former modifies only the boundaries of situations, while the latter changes also the optimal actions in some situation. We then build on this distinction to illustrate some risks of attempting change, including "worse before better" dynamics (Repenning and Sterman, 2002).

Section 5.2 turns to mechanisms behind frame change. We offer a basic model for the formation and evolution of a frame, inspired by the psychological literature on categorization, and we then analyze its implications for a leader who seeks to manage her followers' frame.

In most of this section's discussion of frame change we focus on the long run, after the parties have accomplished not only (i) appraisal (i.e., new thresholds are in place) but also (ii) evaluation (i.e., beliefs about payoffs in new situations are in place). Both appraisal and evaluation could take time, and both deserve their own analyses, but we cannot conduct those here.

5.1 Consequences of frame change

We begin by distinguishing between two kinds of frame change: incremental versus radical. Imagine that the current thresholds for the frame are $\hat{p} = \hat{r} = x$. Suppose that the initial common threshold $x > 1/2$ shifts down to $x' < x$. We distinguish two cases: (a) $x' > 1/2$, versus (b) $x' < 1/2$.

The case $x' > 1/2$ is shown on the left of Figure 9. When the parties' frame changes, they recategorize some games as different situations. Because $x' < x$, the probabilities that the

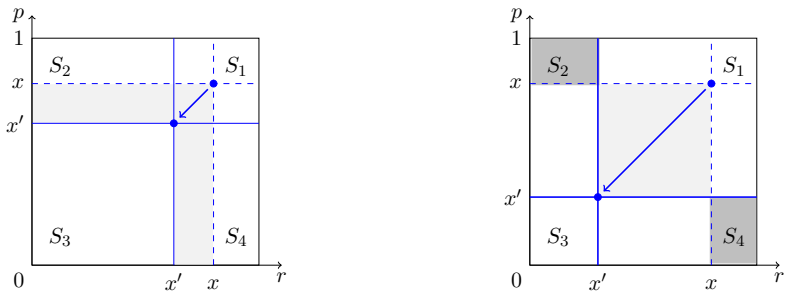


Figure 9: The changes in perceived situations after lowering thresholds from $\hat{p} = \hat{r} = x > 1/2$ to $x' > 1/2$ (left) or to $x' < 1/2$ (right).

parties perceive S_1, S_2 and S_4 increase, and the probability that they perceive S_3 decreases. On the other hand, because $x' > 1/2$, the rational rule of behavior does not change: it remains Cooperation by Default, and S_1, S_2 and S_4 are still played cooperatively. In short, the parties' behavior changes only because the parties recategorize some games from S_3 to S_1, S_2, or S_4 and thus switch behavior from L to H; these games correspond to the light gray area on the left of Figure 9.

The case $x' < 1/2$ is shown on the right of Figure 9. As above, the parties recategorize some games from S_3 to S_1 and switch their behaviors in these games from L to H; see the light gray area on the right of Figure 9. Notably, there is now a second source of change in behavior: the parties used to play Cooperation by Default but, when $x' < 1/2$, they switch

to Defection by Default and change actions (from H to L) in the dissonant situations S_2 and S_4. Hence, games ascribed to dissonant situations *both before and after the change in frame* are now played differently—see the dark gray areas on the right of Figure 9.

Clearly, any shift in the frame's boundaries changes what states of the world parties ascribe to particular situations (i.e., the size of the frame's cells). More importantly, after they learn the payoffs associated with these new situations, the frame change may also affect which actions they choose in specific situations (their rule of behavior). We call a change in frame *incremental* when the agents' rule of behavior does not change, and we call it *radical* when it does.

We now imagine a *leader* who seeks to change a group's frame to improve its performance. The group members (i.e., the "parties" in Sections 3 and 4, who share a given frame) are referred to as the *followers*. We assume that the followers' frame changes in the same way at the same time for both followers. We also assume that the leader knows the whole model, including that the followers perceive the space of games as situations generated by a frame $(\hat{r}, \hat{p})$.

Returning to the example from Section 5.1 where $\hat{r} = \hat{p} = x$, suppose that the leader may lower or raise the threshold x from its initial value. Figure 10 shows the expected payoff U to each follower, analogous to Figure 7 above; see Proposition A.5 for details.

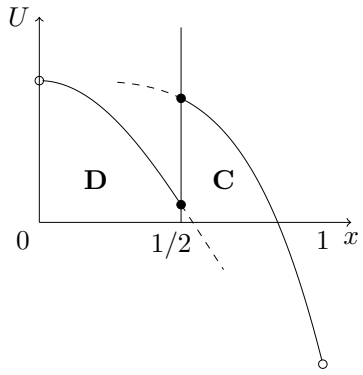


Figure 10: Follower's payoff as a function of x when the frame is $(\hat{r}, \hat{p}) = (x, x)$.

Suppose that the leader lacks full control of the followers' new frame.[4] For example, assume that the leader controls the direction of frame change—either L (shift x to the Left) or R (shift x to the right)—but not its magnitude.

Suppose that the initial threshold is $x > 1/2$, so the followers initially use Cooperation by Default. The action R is dominated by staying put. The action L, on the other hand,

[4] March (1981: 563) reminds us that "Organizations are continually changing, [...] but change within organizations cannot be arbitrarily controlled."

is risky: a small shift to a new threshold $x' > 1/2$ increases payoffs, but a larger shift to a new threshold just below 1/2 discontinuously decreases payoffs. Thus, under Cooperation by Default, attempts at *incremental change* (i.e., a mild reduction of the threshold) are worthwhile, unless the risk of a radical change—with followers switching to Defection by Default—is too high.

Alternatively, suppose that the initial threshold is $x < 1/2$, so the followers initially use Defection by Default. The action L now increases payoffs through incremental change, whereas the action R may lead to a radical change: if the threshold crosses the 1/2 barrier then payoffs increase substantially, but if the threshold moves right without crossing the barrier then payoffs are worse than before.

A second risk of frame change concerns the speed of change. As noted above, suppose the followers go through two steps after the leader's intervention: (i) appraisal (whether and how much thresholds change), and (ii) evaluation (how long it takes before followers update their beliefs about payoffs in a reconfigured situation). Here we suppose that (i) occurs instantly but there is delay in (ii). That is, the followers exhibit inertia (because of a delay in updating beliefs about payoffs) and hence stick to their previous rule of behavior for a while.

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 example). But suppose that $x < 1/2$ and the leader achieves a radical change to $x' > 1/2$. The followers were using Defection by Default, so after a radical change crossing 1/2 from the left the payoff stays on the lower dashed curve until the followers complete the evaluation step (b), after which behavior changes and the payoff jumps up to the higher solid curve—at the new threshold $x' > 1/2$. In short, cognitive inertia in the evaluation of a radical change in frame may cause a transient decline in performance before producing its positive effects: the "worse-before-better" dynamic emphasized by Repenning and Sterman (2002).

Conversely, suppose $x > 1/2$ and the leader intends incremental change to the left. If the frame changes too far, becoming a radical change to $x' < 1/2$ (and close to 1/2), the 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.

In much of the business strategy and organization literature, radical change is conceived as "long jumps" in some space of organizational features (Levinthal 1997; Roberts 2004). In this view, the costs and risks of change stem from the need to reach distant points by small steps. Our notion of radical change is different, because it refers to a switch in behavior rules that engenders a discontinuity in performance. In our model, when an organization is close to the point of discontinuity, even a small step may cause radical change; when this occurs, the asynchronous update of frames and behavior rules can generate an implementation dip.

5.2 Managing frames

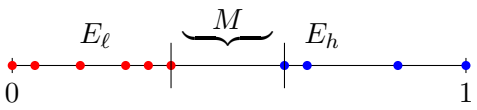
Schein defined the culture of a group as "a pattern of shared basic assumptions" (1985: 9) and argued that the essence of leadership is creating and managing this culture. We take the frame of organization members as such a shared basic assumption and now consider how a leader might attempt to change an organization's frame.

First, we enrich our basic model by providing a mechanism for the formation and evolution of a frame, inspired by the psychological literature on categorization. Then we analyze some trade-offs faced by a leader who seeks to change her followers' frame.

There is an established literature on categorization in the cognitive sciences, with a variety of formal models that describe or predict how human subjects organize their sensory experience into categories; see Pothos and Wills (2011). Two dominant approaches to categorization are prototype theory (Rosch, 1973; Osherson and Smith, 1981) and exemplar theory (Medin and Schaffer, 1978; Nosofsky, 1986). The first postulates that there is some central element (the prototype) for each cluster of similar objects; a novel stimulus is attributed to the category associated with the closest prototype. The second stipulates that each category is associated with some exemplars stored in memory rather than by an abstract summary representation; a novel stimulus is attributed to the category that maximizes the stimulus's overall similarity with the category's set of exemplars. The huge literature comparing these two (and other related) approaches has produced mixed evidence, depending on the fine details of the specific applications. We use a mixed approach that is simpler to present.

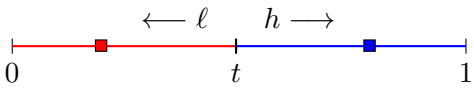
We impose two simplifications. First, the categorization is deterministic: a stimulus is uniquely assigned to a category. Second, all the exemplars lie on the main diagonal; that is, for each exemplar $e = (r, p)$, we have $r = p = x$ for some x in $(0, 1)$. Abusing notation, we write x to denote both the exemplar and its coordinates.

There are only two categories, ℓ (low) and h (high), with *exemplar sets* E_ℓ and E_h. Each exemplar set contains its archetype: the zero-sum game $(0, 0)$ is in E_ℓ and the common-interest game $(1, 1)$ is in E_h. The cardinalities of the exemplar sets $n_\ell \geq 1$ and $n_h \geq 1$ may be different. We assume $\max E_\ell < \min E_h$ so that there is a *middle ground* M separating E_ℓ from E_h. Intuitively, the middle ground is where the tug-of-war between the exemplars for ℓ and for h may be usually summarized into a threshold $\hat{x}$. A case with $n_\ell = 6$ and $n_h = 4$ is shown below.



For the two exemplar sets E_ℓ and E_h we compute the *average values* $\bar{e}_\ell$ and $\bar{e}_h$, depicted

below as squares.



We use $\bar{e}_\ell$ and $\bar{e}_h$ as prototypes for the two categories ℓ and h: a novel stimulus x is categorized as low (ℓ) if it is closer to $\bar{e}_\ell$ and as high (h) if it is closer to $\bar{e}_h$. This construction is equivalent to using the threshold $t = (1/2)\bar{e}_\ell + (1/2)\bar{e}_h$ and categorizing x as ℓ if $x < t$ and as h if $x > t$. This yields a partition for ℓ and h into adjacent intervals, as shown above.

Putting back the diagonal in the unit square, the frame with thresholds $\hat{r} = \hat{p} = t$ is shown in Figure 11.

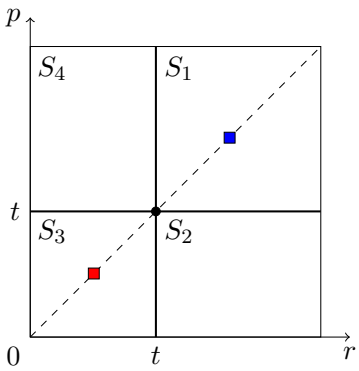


Figure 11: A categorization with four situations.

We assume that the leader knows the whole model, including how exemplars stored in (or removed from) the organization's memory affect the followers' frame. We imagine the leader attempting to change the followers' frame by taking actions such as using specific language, or extolling certain behaviors, or telling particular stories.

We illustrate four considerations for a leader, each through its own simple vignette. The goal is to present simple examples; each vignette might be subject to a deeper analysis. The leader seeks to maximize the sum of the followers' static payoffs: up to an irrelevant constant, the leader's payoff is $V(\hat{x}) = 1 - 2\hat{x}^4$ if $\hat{x} > 1/2$ and $1 - 4\hat{x}^2 + 2\hat{x}^4$ if $\hat{x} < 1/2$.

1. Changing exemplars. An obvious way to shift category boundaries is to add exemplars. For example, the acquisition of Lotus by IBM marked a key moment in Gertstner's culture change at IBM, shifting the boundaries of accepted sources of new technology to include acquisitions (Gerstner, 2002).

In addition to changing frames by adding exemplars, in our model frames can also change by deleting exemplars—"unlearning," as Hedberg (1981) calls it. Note that deletion is con-

strained by the existing exemplars in organizational memory. If the leader needs to fine-tune the frame to the threshold $\hat{t}$, none of the available deletions may get it right, whereas $\hat{t}$ may be feasible with the right addition(s) of exemplars.

2. The weight of memory. The size n_i of an exemplar set E_i dictates the degree of resistance to changing the frame: when n_i is large, addition or deletion of individual exemplars has a smaller effect. An organizational memory that stores a larger number of exemplars has a dampening effect on the addition (or deletion) of an exemplar. Consider an organization spread over two sites. Assume both sites have the same average values $\bar{e}_\ell, \bar{e}_h$ for the exemplar sets, and hence the same threshold t and the same frame. But suppose the first is a "brownfield" with a deep memory (higher values for n_i) and the second is a "greenfield" with a shallow memory (lower values for n_i). Then the frame at the first site is more resistant to change. If the leader attempts a frame shift for the whole organization by adding a new exemplar, each site will process it with respect to its own memory, leading to different new frames at the two sites.

3. Fictitious exemplars. Organizations learn not only from direct experience, but also from various forms of vicarious learning (Levitt and March, 1988), including stories (Selznick, 1957). When actual occurrences cannot serve as useful exemplars, a leader may attempt frame change by telling stories as fictitious exemplars. Compared to an actual exemplar, stories may have vagaries of interpretation that make their effects more difficult to predict (Boje, 1991). Whether a story is worth telling or not will depend on tradeoffs of the potential advantages of a new threshold versus the risk that the followers interpret the story differently from the leader's intention. The Appendix reports an example illustrating this point.

There are cases, especially when the organizational memory is deep, in which the leader may need to add an extreme exemplar to achieve desired change. One way is to introduce an *outlandish exemplar* describing a case that is very expressive but possibly infeasible in practice. Allegories, analogies, and parables are constructs that need not be factually possible, but may effectively promote a different viewpoint.

Our model could be extended to accommodate outlandish exemplars. Recall the assumption x in $[0, 1]$. Adding an actual exemplar from $[0, 1]$ to E_h can move the threshold up at most to $t' = (n_h \bar{e}_h + 1)/(n_h + 1)$. In contrast, suppose an outlandish exemplar for E_h is associated with $x = 1 + \alpha$, where $\alpha > 0$ is the fictitious excess over the highest feasible exemplar. If the leader succeeds in adding x to E_h, then the value of the new threshold is $t' + \alpha/(n_h + 1)$. Note that the amount by which the fictitious excess α changes the threshold is mediated by the depth n_h of the organizational memory.

The followers may interpret an outlandish exemplar with fictitious excess α as inspirational as intended by the leader, or they might reject it as preposterous. One could model these possibilities by assuming that the probability $\Pr(\alpha)$ that an outlandish exemplar $x = 1 + \alpha$ is assimilated into the organizational culture is a decreasing function of α. There would then be a tradeoff between the potential impact of an outlandish exemplar and the risk of its rejection.

4. Acting now or later. If the leader needs to wait for an actual occurrence to use it as an exemplar, she might face a real-option problem. Using the current (candidate) exemplar for incremental change may improve organizational performance but also increase organizational inertia, both by making organizational memory deeper and by shrinking the set of future exemplars that can achieve radical change. For example, in Figure 10, if the current threshold x is just below $1/2$, adding an exemplar that moves the threshold slightly left to $x' < x$ improves performance immediately but also increases the cardinality n_ℓ of one exemplar set and increases the distance that the threshold must travel to achieve radical change (to $x'' > 1/2$). Put more evocatively, because "inertia of organizational capabilities is the source of the value of real options" (Kogut and Kuklatilaka, 2001: 746), a leader may prefer to pass on an incremental improvement now and preserve the opportunity to trigger radical change later. Waiting to act may be an investment in strategic flexibility. In the Appendix, we provide such an example.

6 Discord

We have thus far explored the consequences of assuming that parties from the same organization share the same frame. The general case where agents have different frames is outside the scope of this paper, but this section takes a first step toward analyzing how parties with different frames might react to the discovery that they do not view the world the same way and attempt to resolve their discord.

As above, the parties are not aware that they are using frames to categorize situations. In the previous Sections the parties share the same frame and hence always take the same action. In this section, however, the parties have different frames so their actions might be miscoordinated.

We assume that, if the parties' actions are miscoordinated, they enter into a sincere dialogue. For example, they might ask each other "What situation did you see?" or "What action to you think is optimal in that situation?" We explore the idea that after miscoordination the parties agree that in future they will truthfully report to each other what situation

each perceives before either chooses an action.

Imagine that the current thresholds for the frames of the two parties are $\hat{p}_1 = \hat{r}_1 = x_1$ and $\hat{p}_2 = \hat{r}_2 = x_2$ with $x_1 < x_2$. We distinguish two cases: (a) $1/2 < x_1$ or $x_2 < 1/2$, versus (b) $x_1 < 1/2 < x_2$. In the first case, the agents use the same behavior rule; in the second case, they use different behavior rules.

The first case (for $x_1 > 1/2$) is shown in Figure 12, where the two panels show the different individual frames. Before miscoordination occurs, because their frames have thresholds above $1/2$, both parties use Cooperation by Default. However, because their individual frames are different, there are games that the parties play differently. We say that a discord is *incremental* when the parties are using the same behavior rule, but their rules are supported by different frames.

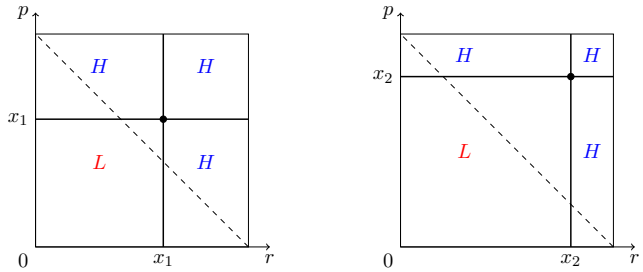


Figure 12: Two frames inducing the same behavior rule.

The left panel in Figure 13 depicts the refinement of the 4-cell individual frames into a 9-cell *joint categorization*, where each cell shows the strategy profile supported by the two individual frames—before the parties encounter miscoordination. Each cell in an individual frame is identified by two binary pieces of information: high/low r and high/low p. In comparison, each cell in the joint categorization is identified by four binary pieces of information: high/low r and high/low p for either agent.

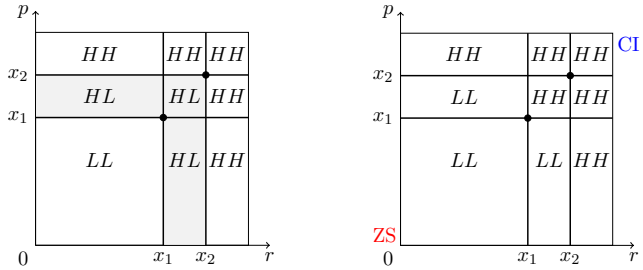


Figure 13: Incremental discord (left) and its resolution under truthful communication (right).

There are three cells where parties will miscoordinate (colored in light gray) and play

HL. Corresponding to the light gray area, the probability of miscoordination is $x_2^2 - x_1^2$: if two parties have slightly different frames (and x_1 is close to x_2), it may take a long time before they miscoordinate on HL.

Once miscoordination takes place and the parties debrief, they find out that both of them were using Cooperation by Default. Miscoordination has occurred because the first party saw high dimensions but the second party saw low dimensions. Being unaware of their cognitive mechanisms, they cannot elicit the source of their incremental discord. But if they exchange truthful information about their interpretations *before* the next play, they can pool their interpretations and use the joint categorization shown on the left panel of Figure 13.

We analyze the steady state after the parties learn their (expected) payoffs for each possible action in each cell of the joint categorization. At that point, we assume that agents are myopic optimisers, who decide how to play their current situation independently of future interactions, and that they resolve their discord by conditioning their individual strategies on the joint categorization. Using Proposition A.4 in the Appendix, it turns out that the dominant strategy for either agent is to play H *unless* at least 3 of the 4 signals are low, as shown on the right panel of Figure 13. We call this rule *Cooperation by Consensus*. In this terminology, an incremental discord over Cooperation by Default is resolved by moving to Cooperation by Consensus.

So far, we have analyzed the sub-case $1/2 < x_1 < x_2$. The other sub-case $x_1 < x_2 < 1/2$ is similar, with incremental discord between two parties now using Defection by Default. When the parties rely on the joint categorization, the dominant strategy for either agent is to play L *unless* at least at least 3 of the 4 signals are high. An incremental discord over Defection by Default is thus resolved by moving to Defection by Consensus.

A different case occurs for $x_1 < 1/2 < x_2$—i.e., when the frame of the first agent supports Defection by Default and the frame of the second supports Cooperation by Default. We say that a discord is *radical* when the parties are using distinct behavior rules, supported by different frames. One might expect only incremental discord when two parties have substantial shared experience in a single organization, and so miscoordination might be infrequent. In contrast, immediately after a merger or when a new boss or peer or subordinate is hired, there might be higher probability of radical discord, as we now analyze.

Figure 14 shows the three cells where parties miscoordinate under radical discord: they play HL in one of them (colored in light gray) and LH in two of them (colored in dark gray). The probability of miscoordination is $(x_2 - x_1)^2 + 2x_1(1 - x_2) \geq 1/4$: miscoordination is quite likely to occur. After it takes place and the parties debrief, they find out that they are using different behavior rules. If they resolve discord using the same process as above, they are led once again to Cooperation by Consensus (if $x_1 > 1 - x_2$, as in Figure 14) or Defection by

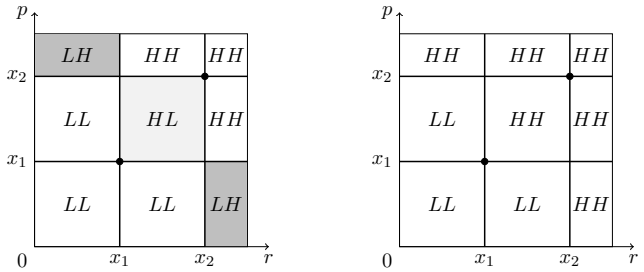


Figure 14: Radical discord (left) and its resolution under truthful communication (right).

Consensus (if $x_1 < 1 - x_2$). Intuitively, the type of consensus is driven by the agent who has a threshold closer to 1/2 and hence less extreme categories.

To summarize this initial analysis of parties with different frames, incremental discord between two agents who play Cooperation (respectively, Defection) by Default is always resolved by moving to Cooperation (Defection) by Consensus. Radical discord, instead, is resolved by moving to Cooperation or Defection by Consensus, depending on which agent has less extreme categories.

7 Literature review

This paper links categorization to organizational culture and performance. Taken singly, there are huge literatures on each of these three topics; any attempt to summarize them would be outside the scope of this paper. Instead, we proceed in three steps. First, we address the links between culture and performance. Then we discuss culture and categorization. Finally, we address issues of organizational change and the role of cognitive frames. We know no work linking all topics in a unified approach. We close our review with a short foray into the game-theoretic literature.

Organizational culture and performance. A stream of work in the management literature emphasizes the positive contribution of strong cultures to organizational performance; see Sørensen (2002) for a thoughtful review. Chatman and O'Reilly (2016) summarize recent evidence on how organizational culture affects performance and articulate the notion of culture strength and content. Our emphasis on culture as shared frames echoes the notion of strong culture and shows how its content may have diverging effects on organizational performance.

In a prescient paper, Barney (1986) analyzes the attributes an organizational culture must have to generate performance advantage—chiefly, inimitability, an issue we address in

Section 4 and 5. Weber and Camerer (2003) offer experimental evidence that cultural conventions are hard to reproduce and that cultural misalignment has performance implications.

Leibenstein (1982) conjectures—through informal use of Prisoners' Dilemma language—that under-performing enterprises might be stuck in Defect-Defect, whereas superior performers might have learned to play Cooperate-Cooperate. Kreps (1990) provides illustrative repeated-game models, highlighting gaps in the theory to be filled. Gibbons and Henderson (2013) connect Leibenstein and Kreps back to Barney by emphasizing that repeated-game models entail not just the familiar credibility problem (should you believe the promise being made?), but also an equally important clarity problem (is there a shared understanding of the promise being made?).

Culture and categorization. As early as 1952, anthropology had over 160 different definitions of culture (Kroeber and Kluckhohn, 1952), but a clear definition of the relationships between culture and cognition emerged only later (D'Andrade, 1995; Bender et al., 2010). It was not until the 1990s that cultural analysis broadly acknowledged cognitive science at its roots (DiMaggio, 1997; Zerubavel, 1991, 1997; Sperber, 1996). This increasing emphasis on the shared cognitive aspects of culture echoes Geertz's (1973:12) pithy "Culture is public because meaning is."

As discussed in the introduction, some management scholars have noted the connection between cognition and culture since the onset of studies on organizational culture. And there is work in economics and political science emphasizing that shared mental models can be held by individuals with common backgrounds or experiences (Denzau and North, 1994). Aoki (2001: 235) explores how shifts in equilibria are associated with changes in the parties' "common cognitive representations." More recently, Hoff and Stiglitz call for economic analyses to consider "cultural mental models [... such as] concepts, categories, social identities, [and] narratives" (2016: 26).

Our choice to focus on categories as basic cognitive entities is not arbitrary. Zerubavel (1991: 1, 3) argues that "the way we cut up the world clearly affects the way we organize our everyday life [... and] varies considerably from one society to another as well as across historical periods within the same society." Recently, Hannan and associates have argued that the analysis of categories can provide guiding principles for cultural analysis (Hannan et al., 2019).

Culture, frames and organizational change. In the broad literature on organizational change, culture and frames are recurrent themes (Burke, 2017). However, in most cases they are associated to inertial forces, fostering stability and triggering defensive resistance

(Argyris, 1985). More recent attempts to reconsider how frames can play an active role in promoting change (Kaplan, 2008; Kellogg, 2011) have cast frames as cognitive/political resources for aligning new organizational coalitions supporting change.

Using a top-down approach, empirical case studies of macro-scale organizational change have focused on the role of corporate leadership as the key driver for the change of shared frames (Schein, 1985); see e.g. Goodstein and Burke (1991) and Kotter and Heskett (1992) on British Airways, Gerstner (2002) on IBM, Fiss and Zajac (2006) on the German corporate system. Across this literature, change is associated to agents who deliberately use discourse, stories, exemplar experiences and incentives to modify the frames that shape how organizational actors perceive the world.

It is a recurrent theme in the management literature that organizational changes do not lie on a continuum, but are better captured by a distinction between two different types of change: incremental vs radical (Greenwood and Hinings, 1996), convergence vs reorientation (Tushman and Romanelli, 1985), continuous vs discontinuous (Weick and Quinn, 1999). Radical change is usually defined as the simultaneous change of several key domains of organizational activity (Gersick, 1991; Romanelli and Tushman, 1994), possibly necessitated by complementarities (Milgrom and Roberts, 1995). The notion that radical change requires a cognitive discontinuity—a restructuring of frames—has been suggested in a number of contributions (Barr et al., 1992; Weick and Quinn, 1999; Gavetti, 2012; Werner and Cornelissen, 2014). Our approach explicitly connects cognition and behavior: radical change is defined by a discontinuous change of behavioral patterns (the mapping of situations to actions) associated to a shift in frame boundaries. We are not aware that this perspective has been developed in the literature on culture, frames and change.

Categorization in game theory. Categorization has appeared in the game-theoretic literature since Jehiel (2005), who considers single games where each player partitions the opponents' moves into categories. We focus on the case where the categorization spans many different games. Heller and Winter (2016) assume that agents simultaneously decide their own categorizations over games, committing to play the same strategy over the same category; in our model, instead, agents are unaware that they are framing. Samuelson (2001) and Mengel (2012) study alternative processes for the categorization of games that, differently from our Section 5.2, are not inspired by empirical evidence. Bednar and Page (2007) demonstrate how different rules of behavior may spontaneously emerge when different games are bundled in the same category.

8 Conclusions

This paper offers a new perspective on how organizational culture might be a strategic resource generating persistent performance differences across firms. We build a theoretical framework that combines organizational and cognitive approaches through the assumption that an organization's members perceive the environment through a shared frame.

We provide several results. First, changes in the cognitive frame may induce differences in collective performance. Second, in a repeated interaction, the frame is as important as agents' patience in achieving cooperation. Third, changes in frame may have starkly different consequences for performance, depending on how they affect the mapping of perceived situations to actions (incremental vs. radical change). We show that radical change may create worse-before-better dynamics. Fourth, we also consider the formation of categories, exploring how a leader may act to change the followers' frame, including the timing of change and the effects of organizational memory and direct vs indirect experience. Finally, we show how parties with different frames may resolve their discord using sincere communication.

In future work, we intend to develop our framework in various directions. One deals with modeling the effects of communication on frames, considering how messages affect the salience of the dimensions over which situations are categorized. We hope to explore the role of language in leadership and organizational culture change. Another concerns cognitive misalignment, when parties understand that frames are imperfectly shared. We are interested in how the parties' efforts may lead to a repair or a collapse of their cooperation.

More broadly, we hope that over time our approach will shed new light on established but puzzling phenomena, as well as suggest new questions. In particular, we see three interesting areas where our approach might contribute: relational contracts, culture as strategy, and leadership.

First, there is a rich and growing literature on relational contracts (see Malcomson (2013) for a survey), but all the models we know are in equilibrium from the beginning. As a result, learning can produce delight or disappointment, but never true surprise. We hope our future analysis of misaligned frames begins to capture such surprise. Another consequence of equilibrium models is that there is never any need to discuss strategies or intentions before the relationship begins. Since essentially no relationship in the real world begins without up-front discussions, it would seem useful for the theory to catch up with this fact, and we again hope that our analysis of misalignment will move in this direction.

Second, about culture as strategy, Barney's (1986) insight that culture must be inimitable if it is to create competitive advantage usually makes culture indescribable and/or taken for granted. In contrast, in our model parties who share a frame have no problem talking—to

themselves or to others—about their rule of behavior (i.e., their mapping from situations to actions), but parties with other frames will disagree at least about when different situations have been realized (as in Section 3) or whether proposed repeated-game strategies are equilibria (as in Section 4). The fact that parties are unaware of their framing prevents them from talking about their frames, even if they can communicate their strategies. We hope our future work on language and leadership provide an underpinning for Barney's inimitability.

Finally, the leader in Section 5 has a superior understanding of followers' framing. There are other models that imagine the leader knowing more than the followers do; e.g., in "leading by example" (Hermalin, 1998). We hope to explore the case of a leader who has some (necessarily) private information about an idea of her devising, such as her strategic intent, but lacks the language to fully share it with her followers.

References

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

[2] C. Argyris (1985), *Strategy, Change and Defensive Routines*, Boston: Pitman.

[3] J.B. Barney (1986), "Organizational culture: Can it be a source of sustained competitive advantage?", *Academy of Management Review* **11**, 656–665.

[4] P.S. Barr, J.L. Stimpert and A.S. Huff (1992), "Cognitive change, strategic action, and organizational renewal", *Strategic Management Journal* **13**, 15–36.

[5] J. Bednar and S. Page (2007), "Can game(s) theory explain culture?: The emergence of cultural behavior within multiple games", *Rationality and Society* **19**, 65–97.

[6] A. Bender, E. Hutchins and D. Medin (2010) "Anthropology in cognitive science", *Topics in Cognitive Science* **2**, 374–385.

[7] D.M. Boje (1991), "The storytelling organization: A study of story performance in an office-supply firm", *Administrative Science Quarterly* **36**, 106–126.

[8] 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.

[9] A. Brandenburger and B.J. Nalebuff (1996), *Co-opetition*, New York: Currency Doubleday.

[10] W.W. Burke (2017), *Organization Change: Theory and Practice*, fifth edition, Thousand Oaks (CA): Sage.

[11] J.A. Chatman and C.A. O'Reilly (2016), "Paradigm lost: Reinvigorating the study of organizational culture", *Research in Organizational Behavior* **36**, 199–224.

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

[13] R.G. D'Andrade (1995), *The Development of Cognitive Anthropology*, Cambridge (UK): Cambridge University Press.

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

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

[16] T. Ellingsen, M. Johannesson, J. Mollerstromm and S. Munkhammar (2012), "Social framing effects: Preferences or beliefs?". *Games and Economic Behavior* **76**, 117–130.

[17] P.C. Fiss and E.J.. Zajac (2006), "The symbolic management of strategic change: Sensegiving via framing and decoupling", *Academy of Management Journal* **49**, 1173–1193.

[18] G. Gavetti (2012), "Toward a behavioral theory of strategy", *Organization Science* **23**, 267–285.

[19] C. Geertz (1973), "Thick description: Toward an interpretive theory of culture", in: *The Interpretation of Cultures: Selected Essays*, New York: Basic Books, 3–30.

[20] C.J. Gersick (1991), "Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm", *Academy of Management Review* **16**, 10–36.

[21] L.V. Gerstner, Jr. (2002), *Who Says Elephants Can't Dance?: Leading a Great Enterprise through Dramatic Change*. New York: HarperCollins.

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

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

[24] L.D. Goodstein and W.W. Burke (1991), "Creating successful organization change", *Organizational Dynamics* **19**, 5–17.

[25] R. Greenwood and C.R. Hinings (1996), "Understanding radical organizational change: Bringing together the old and the new institutionalism", *Academy of Management Review* **21**, 1022–1054.

[26] M.T. Hannan, G. Le Mens, G. Hsu, B. Kovács, G. Negro, L. Pólos, E. Pontikes, and A.J. Sharkey (2019), *Concepts and Categories: Foundations for Sociological and Cultural Analysis*, New York: Columbia University Press.

[27] B. Hedberg (1981), "How organizations learn and unlearn", in: P.C. Nystrom and W.H. Starbuck, eds., *Handbook of Organizational Design*, Oxford: Oxford University Press, 1–27.

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

[29] 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.

[30] B.E. Hermalin (1998), "Toward an economic theory of leadership: Leading by example", *American Economic Review* **88**, 1188–1206.

[31] 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.

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

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

[34] S. Kaplan (2008), "Framing contests: Strategy making under uncertainty", *Organization Science* **19**, 729–752.

[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] K.C. Kellogg (2011), "Hot lights and cold steel: Cultural and political toolkits for practice change in surgery", *Organization Science* **22**, 482–502.

[37] B. Kogut and N. Kulatilaka (2001), "Capabilities as real options", *Organization Science* **12**, 744–758.

[38] J.P. Kotter and J.L. Heskett (1992), *Corporate Culture and Performance*, New York: Free Press.

[39] D.M. Kreps (1990), "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.

[40] A.L. Kroeber and C. Kluckhohn (1952), *Culture: A Critical Review of Concepts and Definitions*, Cambridge (MA): The Museum.

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

[42] B. Levitt and J.G. March (1988), "Organizational learning", *Annual Review of Sociology* **14**, 319–338.

[43] D.A. Levinthal (1997), "Adaptation on rugged landscapes", *Management Science* **43**, 934–950.

[44] 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.

[45] J.M. Malcomson (2013), "Relational incentive contracts", in: R. Gibbons and J. Roberts (eds.), *The Handbook of Organizational Economics*, Princeton: Princeton University Press, 1014–1065.

[46] J.G. March (1981), "Footnotes to organizational change", *Administrative Science Quarterly* **26**, 563–577.

[47] J.G. March and J.P. Olsen (1983), "The new institutionalism: Organizational factors in political life", *American Political Science Review* **78**, 734–749.

[48] D.L. Medin and M.M. Schaffer (1978), "Context theory of classification learning", *Psychological Review* **85**, 207–238.

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

[50] P. Milgrom and J. Roberts (1995), "Complementarities and fit strategy, structure, and organizational change in manufacturing", *Journal of Accounting and Economics* **19**, 179–208.

[51] R.M. Nosofsky (1986), "Attention, similarity, and the identification-categorization relationship", *Journal of Experimental Psychology: General* **115**, 39–57.

[52] D.N. Osherson and E.E. Smith (1981), "On the adequacy of prototype theory as a theory of concepts", *Cognition* **9**, 35–58.

[53] E. Ostrom (1990), *Governing the Commons: The Evolution of Institutions for Collective Action*, New York: Cambridge University Press.

[54] O. Patterson (2014), "Making sense of culture", *Annual Review of Sociology* **40**, 1–30.

[55] A.M. Pettigrew (1979), "On studying organizational cultures", *Administrative Science Quarterly* **24**, 570–581.

[56] E.M. Pothos and A.J. Wills (2011), eds., *Formal Approaches in Categorization*, Cambridge, UK: Cambridge University Press.

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

[58] 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.

[59] J. Roberts (2007), *The Modern Firm: Organizational Design for Performance and Growth*. Oxford: Oxford University Press.

[60] E. Romanelli and M.L. Tushman (1994), "Organizational transformation as punctuated equilibrium: An empirical test", *Academy of Management Journal* **37**, 1141–1166.

[61] E.H. Rosch (1973), "Natural categories", *Cognitive Psychology* **4**, 328–350.

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

[63] E. Schein (1968), "Organizational socialization and the profession of management", *Industrial Management Review* **9**, 1–15.

[64] E. Schein (1985), *Organizational Culture and Leadership*, San Francisco: Jossey-Bass. (Fourth edition: 2010.)

[65] T.C. Schelling (1960), *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

[66] P. Selznick (1957), *Leadership in Administration: A Sociological Interpretation*, New York: Harper and Row.

[67] J.B. Sørensen (2002), "The strength of corporate culture and the reliability of firm performance", *Administrative Science Quarterly* **47**, 70–91.

[68] D. Sperber (1996), *Explaining Culture: A Naturalistic Approach*, Oxford: Blackwell.

[69] M.L. Tushman and E. Romanelli (1985), "Organizational evolution: A metamorphosis model of convergence and reorientation", *Research in Organizational Behavior* **7**, 171–222.

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

[71] K.E. Weick and R.E. Quinn (1999), "Organizational change and development", *Annual Review of Psychology* **50**, 361–386.

[72] M.D. Werner and J.P. Cornelissen, J. P. (2014), "Framing the change: Switching and blending frames and their role in instigating institutional change", *Organization Studies* **35**, 1449–1472.

[73] E. Zerubavel (1991), *The Fine Line*, New York (NY): Free Press.

[74] E. Zerubavel (1997), *Social Mindscapes: An Invitation to Cognitive Sociology*, Cambridge, MA: Harvard University Press.

A Online Appendix

Throughout this appendix, we use majuscules to denote random variables and minuscules to denote their realizations.

Proposition A.1. *Under the benchmark, the expected payoff to a party when both players use their dominant strategies is* $1/6$.

Proof. When both players use their dominant strategies, the payoff to an agent is PR if $R + P > 1$, and $-PR$ if $R + P < 1$. Write this payoff as

$$V = -PR + 2PR \cdot \mathbb{1}_{\{R \geq 1-P\}}$$

Given the independence of R and P, its expectation is

$$E(V) = E(-PR) + E\left(2PR \cdot \mathbb{1}_{\{R \geq 1-P\}}\right) = -\frac{1}{4} + \frac{5}{12} = \frac{1}{6}$$

□

As discussed in the main text, the payoff matrix for a game $G(r, p)$ is

	H	L
H	pr, pr	$-(1-p)(1-r), (1-p)(1-r)$
L	$(1-p)(1-r), -(1-p)(1-r)$	$-pr, -pr$

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. For generality, let $K = (\alpha, \beta) \times (\gamma, \delta)$ be the cell including the games $G(R, P)$ with $r \in (\alpha, \beta)$ and $P \in (\gamma, \delta)$, where $0 \leq \alpha < \beta \leq 1$ and $0 \leq \gamma < \delta \leq 1$; see Figure 15.

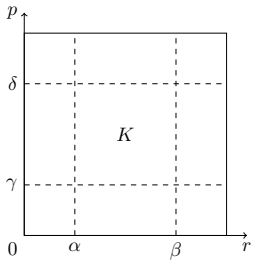


Figure 15: A cell in the game space.

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

Lemma A.2. *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*

$$\alpha + \beta + \gamma + \delta \geq 2 \tag{1}$$

Proof. The expected payoff for i over K under the strategy profile (H, H) is $E_K\Big(PR\Big)$. Because the two uniform random variables r and P are stochastically independent, we have

$$E_K\Big(PR\Big) = E_K\Big(P\Big) \cdot E_K\Big(R\Big) = \frac{(\alpha + \beta)(\gamma + \delta)}{4}$$

Analogously, under the strategy profile LH, the expected payoff for i over K is

$$E_K\Big((1 - P)(1 - R)\Big) = E_K\Big(1 - R\Big) \cdot E_K\Big(1 - P\Big) = \frac{(2 - \alpha - \beta)(2 - \gamma - \delta)}{4}$$

Hence, when the opponent plays H over K, H is preferred to L if and only if

$$\frac{(\alpha + \beta)(\gamma + \delta)}{4} \geq \frac{(2 - \alpha - \beta)(2 - \gamma - \delta)}{4}$$

Simplifying and rearranging, we obtain the inequality in (1). □

Lemma A.3. *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 (1) holds.*

Proof. Under the strategy profile HL, the expected payoff for i over K is $E_K\Big[-(1 - P)(1 - R)\Big]$. Under the strategy profile LL, the expected payoff for i over K is $E_K\Big(-PR\Big)$. Hence, H is preferred to L if and only if $E_K\Big(PR\Big) \geq E_K\Big[-(1 - P)(1 - R)\Big]$, which leads back to the inequality in (1). □

Proposition A.4. *Given a cell $K = (\alpha, \beta) \times (\gamma, \delta)$, let $\overline{r} = (\beta + \alpha)/2$ and $\overline{p} = (\delta + \gamma)/2$. Then i has a (strictly) dominant strategy if $\overline{r} + \overline{p} \neq 1$ and is indifferent between H and L if equality holds. Moreover, the dominant strategy is H if $\overline{r} + \overline{p} > 1$, and it is L if $\overline{r} + \overline{p} < 1$.*

Proof. If we divide by 2 the expressions on either side of the inequality in (1), the result follows immediately from Lemma A.2 and Lemma A.3. □

Proposition A.5. *Given a threshold pair $(\hat{r}, \hat{p})$, the expected payoff to each agent when they both use dominant strategies is*

$$\frac{1 - 2\hat{r}^2\hat{p}^2}{4} \qquad \textit{when } \hat{r} + \hat{p} > 1 \tag{2}$$

and

$$\frac{1-2\hat{r}^2-2\hat{p}^2+2\hat{r}^2\hat{p}^2}{4} \quad \textit{when } \hat{r}+\hat{p}<1 \tag{3}$$

Proof. Suppose $\hat{r}+\hat{p}>1$. The rational rule of behavior is Cooperation by Default, yielding a random payoff PR in situations S_1, S_2, S_4 and $-PR$ in S_3. Therefore,

$$E(V) = P(S_1)E_{S_1}(PR) + P(S_2)E_{S_2}(PR) - P(S_3)E_{S_3}(PR) + P(S_4)E_{S_4}(PR)$$

where $P(S_i)$ denotes the probability that the situation S_i occurs. The proof of Lemma A.2 provides general expressions for $E_K(Pr)$. Substituting these and dropping hats for simplicity, we find

$$\begin{aligned} E(V) &= (1-p)(1-r)E_{S_1}(PR) + p(1-r)E_{S_2}(PR) - prE_{S_3}(PR) + (1-p)rE_{S_4}(PR) \\ &= \frac{(1-r^2)(1-p^2)}{4} + \frac{p^2(1-r^2)}{4} - \frac{r^2p^2}{4} + \frac{r^2(1-p^2)}{4} = \frac{1-2r^2p^2}{4} \end{aligned}$$

Suppose instead $\hat{r}+\hat{p}<1$. The rational rule of behavior is Defection by Default, yielding a random payoff PR in situation S_1, and $-PR$ in situations S_2, S_3, S_4. Proceeding similarly, we find

$$\begin{aligned} E(V) &= P(S_1)E_{S_1}(PR) - P(S_2)E_{S_2}(PR) - P(S_3)E_{S_3}(PR) - P(S_4)E_{S_4}(PR) \\ &= (1-p)(1-r)E_{S_1}(PR) - p(1-r)E_{S_2}(PR) - prE_{S_3}(PR) - (1-p)rE_{S_4}(PR) \\ &= \frac{(1-p^2)(1-r^2)}{4} - \frac{p^2(1-r^2)}{4} - \frac{r^2p^2}{4} - \frac{r^2(1-p^2)}{4} = \frac{1-2r^2-2p^2+2r^2p^2}{4} \end{aligned}$$

□

Proposition A.6. *When Coordination by Default prevails, there is fog of conflict if*

$$\hat{r}^2 \cdot \hat{p}^2 > 1/6 \tag{4}$$

and fog of cooperation if the opposite (strict) inequality holds.
When Defection by Default prevails, there is fog of conflict if

$$\hat{r}^2 + \hat{p}^2 - \hat{r}^2 \cdot \hat{p}^2 > 1/6 \tag{5}$$

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

Proof. By Proposition A.1, the expected payoff to a party under the benchmark is 1/6. Given a threshold pair $(\hat{r}, \hat{p})$, Proposition A.5 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 1/6. If $\hat{r} + \hat{p} > 1$ and Cooperation by Default applies, the expected payoff in (2) is lower (or greater) than 1/6 when (4) (or its opposite) holds. The argument is similar using (3) when Defection by Default applies. □

Frames and fog. Beyond the special case of $\hat{r} = \hat{p} = x$ discussed in Section 3 and shown in Figure 7, Proposition A.6 characterizes which frames generate which kind of fog. Its main message is conveyed in Figure 16. Note that the unit square in Figure 16 differs importantly from those in Figures 4 and 6: instead of depicting the situations created by one particular frame as in the earlier figures, now Figure 16 illustrates the space of all possible frames—each frame is determined by a threshold pair $(\hat{r}, \hat{p})$ in $(0, 1)^2$.

In Figure 16 cooperating by default prevails when the parties' shared frame is a threshold pair $(\hat{r}, \hat{p})$ above the diagonal, while defecting by default prevails when it is below. The

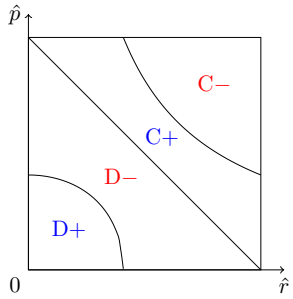


Figure 16: Frames engendering fog of cooperation (+) or fog of conflict (−).

curve above the diagonal separates the cooperating-by-default region into the frames $(\hat{r}, \hat{p})$ generating a fog of conflict (marked −) versus those generating a fog of cooperation (marked +). Similarly, the curve below the diagonal separates the defecting-by-default region into fog of conflict (marked −) versus fog of cooperation (marked +). Consistent with the special case of $\hat{r} = \hat{p} = x$ depicted in Figure 7, moving from northeast to southwest within a rule in Figure 16 improves payoffs continuously; on the other hand, crossing the boundary from Cooperation by Default into Defection by Default causes a discontinuous drop in payoffs.

Proposition A.7. *Suppose $x > 1/2$. Full Cooperation on (H, H) can be supported by a subgame perfect equilibrium based on a Nash-reversion strategy if and only if*

$$\delta \geq \frac{2 - 2x}{2 - 2x + x^4} \tag{6}$$

Proof. Cooperation on (H, H) is a dominant strategy for the three static situations S_1, S_2, S_4. Thus, we need to compare the short-term temptation against the long-term benefit only for

S_3. Assuming $\hat{r} = \hat{p} = x$, the payoff matrix perceived by the agents in S_3 is depicted on the left of Figure 17. (Payoffs are rescaled by a factor of 4.)

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

S_3

	H	L
H	$(1+x)x$	$-(1-x)(2-x)$
L	$(1-x)(2-x)$	$-(1+x)x$

S_2

Figure 17: The one-shot games associated with S_3 (left) and S_2 (right) for $\hat{r} = \hat{p} = x$.

The short-term temptation $\mathrm{ST}(S_3)$ for S_3 is the difference in payoffs from playing L instead of H in situation S_3 when the other party plays H:

$$\mathrm{ST}(S_3) = (2-x)^2 - x^2 = 4(1-x)$$

The long-term benefit $\mathrm{LB}(S_3)$ is the discounted sum of the (expected) incremental payoffs from sustaining cooperation in S_3. Since S_3 occurs with probability x^2, we find

$$\mathrm{LB}(S_3) = \frac{\delta}{1-\delta}\left[x^2 \cdot \left(x^2 - (-x^2)\right)\right] = \frac{2\delta x^4}{1-\delta}$$

Imposing $\mathrm{LB}(S_3) \geq \mathrm{ST}(S_3)$ gives

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

which yields (6). □

Proposition A.8. *Suppose $x < 1/2$. Full Cooperation on (H, H) can be supported by a subgame perfect equilibrium based on a Nash-reversion strategy if and only if*

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

Cooperation by Default on (H, H) can be supported by a subgame perfect equilibrium based on a Nash-reversion strategy if and only if

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

Proof. Consider Full Cooperation. Because cooperation on (H, H) is a dominant strategy only in the static situation S_1, we need to compare the short-term temptation against the long-term benefit across the other three situations S_2, S_3 and S_4. Under the assumption $\hat{r} = \hat{p} = x$, the payoffs for S_2 and S_4 are the same so we restrict attention to S_3 and S_2 in

the following. Their payoff matrices are depicted in Figure 17.

The short-term temptations in S_3 and in S_2 (or S_4) are $\mathrm{ST}(S_3) = 4 - 4x$ and $\mathrm{ST}(S_2) = \mathrm{ST}(S_4) = 2 - 4x$, respectively. Because $\mathrm{ST}(S_3) \geq \mathrm{ST}(S_2)$ for any x, Full Cooperation across all situations obtains if and only if the long-term benefit across S_2, S_3 and S_4

$$\mathrm{LB}(S_2S_3S_4) = \frac{\delta}{1-\delta}\left[x^2 \cdot \left(x^2 - (-x^2)\right) + 2x(1-x) \cdot \left(x + x^2 - (-x - x^2)\right)\right] = \frac{2\delta(2x^2 - x^4)}{1-\delta}$$

from cooperation in the three situations S_2, S_3 and S_4 exceeds the (higher) temptation $\mathrm{ST}(S_3)$. Rearranging $\mathrm{LB}(S_2S_3S_4) \geq \mathrm{ST}(S_3)$ yields (7).

Consider now improved cooperation, when Defection by Default in the static model is upgraded to Cooperation by Default in the repeated interaction. Because the two rules offer matching prescriptions over the two consonant situations (cooperation in S_1, defection in S_3), it suffices to check that the long-term benefit

$$\mathrm{LB}(S_2S_4) = \frac{\delta}{1-\delta}\left[2x(1-x) \cdot \left(x + x^2 - (-x - x^2)\right)\right] = \frac{2\delta(2x^2 - 2x^4)}{1-\delta}$$

from cooperation in S_2 and S_4 exceeds the short-term temptation $\mathrm{ST}(S_2) = \mathrm{ST}(S_4) = 2 - 4x$ in each dissonant situation. Rearranging $\mathrm{LB} \geq \mathrm{ST}(S_2)$ yields (8). □

Proposition A.9. *Suppose that the parties perceive any two games $(r_1, p_1) \neq (r_2, p_2)$ in $\mathcal{G}$ as distinct. Then cooperation on (H, H) across all games in $\mathcal{G}$ can be supported by a subgame perfect equilibrium based on a Nash reversion strategy if and only if*

$$\delta \geq \frac{12}{13} \tag{9}$$

Proof. Recall the payoff matrix for a game $G(r, p)$ from Figure 2. Let the *short-term temptation*

$$\mathrm{ST}(r, p) = (1-p)(1-r) - pr = 1 - p - r$$

be the difference in payoffs from choosing L versus H when the other party plays H in the one-shot game $G(r, p)$.

Let the *long-term benefit* LB be the discounted sum of the incremental payoffs from sustaining cooperation against defection across all games in $\mathcal{G}$. Because cooperation on (H, H) is a dominant strategy for $G(r, p)$ when $r + p \geq 1$, it suffices to consider the complementary event $D^- = \{(r, p) : r + p < 1\}$. Then

$$\mathrm{LB} = \left(\frac{\delta}{1-\delta}\right) E\left[PR - (-PR) \cdot \mathbb{1}_{D^-}\right] = \left(\frac{\delta}{1-\delta}\right) E\left[2PR\mathbb{1}_{\{R+P<1\}}\right] = \frac{1}{12}\left(\frac{\delta}{1-\delta}\right)$$

Cooperation on (H, H) can be supported across all games only if the long-term benefit LB is never smaller than the short-term temptation $\mathrm{ST}(r, p)$ for all (r, p) in D^-; that is, only if $\mathrm{LB} \geq \mathrm{ST}(r, p)$. Rearranging this expression, we find

$$\delta \geq \frac{12(1-p-r)}{1+12(1-p-r)}$$

that holds for any (r, p) in D^- if and only if (9) holds. □

5.2.3. Fictitious exemplars. We provide a simple example to illustrate the tradeoffs behind using stories as fictitious exemplars. Let the exemplar sets be $E_\ell = \{0, 0.2\}$ and $E_h = \{0.6, 1\}$, with $\bar{e}_\ell = 0.1$ and $\bar{e}_h = 0.8$. The middle ground is $[0.2, 0.6]$. The current threshold is $t = 0.45$: the followers play Defection by Default and $V(t) \approx 0.272$.

The leader is considering a story from the middle ground. We represent her uncertainty about its interpretation by assuming that the followers will eventually attribute to the story a value x that is uniformly distributed over the entire middle ground $[0.2, 0.6]$. Once the interpretation is settled, the followers behave according to the new threshold.

If the story shifts the new threshold above $1/2$, the followers switch to Cooperation by Default and achieve a radical change for the better. Otherwise, the change enacted by the story is incremental and V decreases. Given an interpretation x, the new threshold is at $t' = (1.3 + 0.5x)/3$. Therefore, $t' > 0.5$ if and only if $x > 0.4$: there is an equal chance of radical change ($x > 0.4$) or incremental change ($x < 0.4$). The expected payoff of telling a story

$$\frac{1}{2}\mathrm{E}\left(V(t')|x > 0.4\right) + \frac{1}{2}\mathrm{E}\left(V(t')|x < 0.4\right) \approx 0.516$$

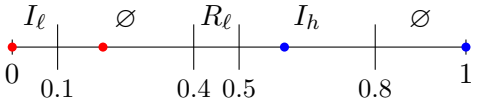
is higher than the current value. So the leader finds it optimal to tell a story, even though there is a risk that the followers interpret the story differently from her intended goal.

5.2.4. Acting now or later. We provide a simple example to illustrate a real-option tradeoff when using exemplars. We use the same initial values as in the previous example, with exemplar sets be $E_\ell = \{0, 0.2\}$ and $E_h = \{0.6, 1\}$.

The leader can act only in the first two periods of her tenure and she can add at most one exemplar per period. In each period, exemplars are uniformly and independently distributed on $[0, 1]$. We assume that the leader can add an exemplar an exemplar $x < 1/2$ only to E_ℓ and an exemplar $x > 1/2$ only to E_h. The leader maximizes her expected discounted payoff over an infinite horizon, using the discount factor ρ.

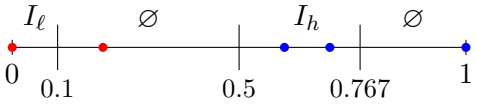
Consider the static one-period problem. The optimal strategy for the leader depends on the value of x: if $x > 0.8$, the leader takes no action (because adding x to E_h would

cause an incremental change that is detrimental); if $0.5 < x < 0.8$, she pursues incremental change and adds x to E_h; if $0.4 < x < 0.5$, she pursues radical change and adds x to E_ℓ; if if $0.1 < x < 0.4$, she takes no action; and if $0 < x < 0.1$, she pursues incremental change and add x to E_ℓ. We summarize the optimal strategy as follows, where I_k and R_k denote incremental and radical change by adding an exemplar to E_k (for $k = \ell, h$), $\varnothing$ denotes no action, and the dots represent the existing exemplars.



The expected value for this first-period strategy given the initial threshold $t = 0.45$ is $V_1 \approx 0.355$.

Consider now that infinite-horizon problem (recall that, for simplicity, the leader can act only in the first two periods). Suppose that in the first period the exemplar $x = 0.7$ becomes available: then the leader must choose whether to add it to E_h and pursue incremental change, or to ignore it and wait for another exemplar to arrive in the next period. If the leader adds $x = 0.7$ to E_h, the new threshold moves to $t' \approx 0.433$ and $V(t') \approx 0.320$. The short-term improvement in V affects the leader's optimal strategy in the second period; in particular, *there are no longer occurrences of x for which radical change is beneficial.* The optimal strategy in the second period can be summarized as follows, including the new exemplar at $x = 0.7$.



The expected value for this second-period strategy given the initial threshold $t' \approx 0.433$ is $V_2 \approx 0.326$.

Consider the two-period problem when $x = 0.7$ has occurred in the first period. If the leader adds $x = 0.7$ to E_h in the first period, this changes her optimal strategy in the second period and her discounted payoff is $0.320 + (0.326\rho)/(1-\rho)$. If she takes no action and waits one period using the forthcoming occurrence in the best possible way, her expected discounted payoff is $0.272 + (0.355\rho)/(1-\rho)$. She prefers to postpone action and preserve the option of radical change in the second period if $0.320 + (0.326\rho)/(1-\rho) < 0.272 + (0.355\rho)/(1-\rho)$; that is, if $\rho > 0.623$.