INSEAD The Business School for the World®

Working Paper

2021/41/STR

# Designing a Culture of Collaboration: When Changing Beliefs is (Not) Enough

Özgecan Koçak
Emory University, ozgecan.kocak@emory.edu

Phanish Puranam
INSEAD, Phanish.puranam@insead.edu

Appeared in *Organization Design*
Advances in Strategic Management, Emarald (2019) Vol. 40, 27-52

Organizational cultures that facilitate collaboration are valuable, but we know relatively little about how to create them. We investigate the micro-foundations of this problem using computational models of dyadic coupled learning. We find that merely altering initial beliefs about the consequence of actions (without altering the consequences themselves) can under some conditions create cultures that promote collaboration. Our results show why the right initial “framing” of a situation- established for instance through persuasive rhetoric, an inspiring vision, or careful recruitment choices- may under the right conditions be self-reinforcing, instead of becoming empty symbolism.

Keywords: Culture; Collaboration; Coupled Learning; Computational Models

Acknowledgements: The authors thank Jeho Lee, Hart Posen, Prothit Sen and participants at the fourth Asian Management Research Consortium (2018) for helpful suggestions.

## 1. Introduction

Organizational culture, defined as a system of shared assumptions, values and norms can facilitate collaboration and thus improve performance (Schein 1983, 2004; O'Reilly & Chatman, 1996; Chatman & O'Reilly, 2016). Failures of collaboration such as those observed after mergers, across internal organizational units, or in the aftermath of too rapid expansion through hiring, are often attributed to cultural divergence (e.g. Van den Steen, 2010a). Conversely, cultures are often credited with the success of organizations (e.g. Pfeffer, 1995). Since collaborative actions are often difficult to measure and verify, there are inevitable limits to the use of contracts and authority (Gibbons & Henderson, 2012). Culture can fill in the gaps in formal administrative systems to promote collaboration (O'Reilly & Chatman, 1996).

This leads naturally to the question of whether and how a culture that supports collaborative efforts can be designed. One cannot simply mandate a particular set of assumptions, values, or norms on the basis of authority, and expect them to persist. In this paper, we focus on a particular, precisely demarcated "object of intervention" in the process of setting up collaborative cultures, and then offer some initial results on how the effectiveness of interventions can be traced back to the properties of the "context of intervention". With respect to the object of intervention, we build on a conceptual framework that distinguishes between design interventions that alter individuals' **beliefs** about the consequences of their actions and interventions that alter the actual **consequences** of those actions.

While recognizing that designers can and do intervene in shaping outcomes of behavior, in this paper we take the structure of interdependence between actors (i.e. how their payoffs depend on each other's actions) as given. We then examine the conditions under which organizational designers can elicit collaborative behavior by merely shaping beliefs about payoffs - which we conceptualize to stem from the assumptions, mental models and representations that actors hold (Thagard, 2005; Rouse & Morris, 1986; Cannon-Bowers et al., 1993; Van den Steen, 2010b). These beliefs are the "object of intervention" in our analysis.

It is a well-established view in the literature that organizations constitute the psychological environment in which choice occurs (Simon, 1947) and exert a powerful influence on the beliefs of their constituent individuals (March & Simon, 1958: Chapter 3). Our analysis helps to see how far one can go towards building a culture of collaboration merely by shaping beliefs; it responds to the recent re-thinking of organizational culture presented by Chatman and O'Reilly (2016) and in particular to their call for renewed attention to the underlying psychological mechanisms that shape cultures (pg. 2). As they note, beliefs and assumptions form a bedrock layer of culture, harder to observe than norms or artifacts, but nonetheless foundational (pg. 16). The analysis in this paper provides a rigorous framework to describe the conditions under which individual beliefs aggregate into stable patterns of collaborative actions.

We build on prior influential work to identify the kinds of cultural interventions that designers can use to influence beliefs. Schein summarizes decades of his own and others' work when he writes:

> "Cultures basically spring from three sources: (1) the beliefs, values, and assumptions of founders of organizations; (2) the learning experiences of group

> members as their organization evolves; and (3) new beliefs, values, and assumptions brought in by new members and leaders." (Schein, 2004: 255).

Thus, organizational designers can try to shape beliefs in an organization: 1) by training new entrants in a top-down fashion (socialization), 2) by influencing the process of interaction and learning among the organization's members (sensemaking), and 3) by selecting organizational members on the basis of cultural fit (sorting). [1]

Clearly, these mechanisms do not operate in isolation from each other. In fact, they must be mutually reinforcing – if the beliefs that individuals bring to an organization or imbibe from its current members at the time of entry do not align with what they experience as they live and interact in the organization, the beliefs are unlikely to remain either stable or shared (Schein, 1983:8). Even brainwashing has its limits, when confronted by the constant intrusion of reality (Schein et al., 1961). The explicit focus on the processes that reinforce or weaken initial beliefs is the first distinctive feature of our analysis. We assume that the initial beliefs held by individuals are modified through feedback to their actions. To formalize this, we build on models of coupled learning that assume bounded but adaptive rationality (e.g. Lave & March, 1975; Knudsen & Srikanth, 2014).

The second distinctive feature of our analysis concerns the "context of intervention". We assume that the key element of context is the feedback that individuals get on their actions. Specifically, we assume that the structure of interdependence between individuals in organizations, which determines the feedback to individuals' joint actions, can exhibit a lot of diversity. We draw on game theoretic representations of common social dilemmas to formalize this variation (e.g. Camerer & Knez, 1996; Camerer, 2003).

---

[1] These mechanisms can also influence the values that individuals attach to their actions and outcomes, but we focus on their consequences for beliefs.

Third, while our approach builds on a tradition of using computational models to understand culture (e.g. Harrison & Carroll, 2006), it uniquely takes a micro-foundational approach to understand the links between individual beliefs and how they are shaped through interaction in a context. We model processes of interaction rather than incorporate their outcomes, such as influence of managers or peers, as exogenous parameters that affect the degree of enculturation.

Our analysis gives sharp predictions about the conditions under which efforts to shape beliefs through cultural design efforts are likely to be effective in creating collaboration, and why. The results suggest that cultural interventions aimed primarily at endowing actors with an initial set of beliefs (for instance through ***socialization*** and ***sorting***), as well as interventions that affect the process through which actors update their beliefs through interaction (i.e. ***sensemaking)***, when combined, promote the emergence of improved collaboration.

Put differently, cultural interventions that only shape beliefs and how they are revised (without altering the actual consequences of actions) may nonetheless contribute to the creation of a culture of collaboration across a range of contexts. Our results imply that the right initial "framing" of a situation- established through persuasive rhetoric and an inspiring vision, may be self-reinforcing under the right conditions, rather than merely empty symbolism. This has implications for practice in situations involving inter-organizational collaboration (e.g. between departments, or in post-merger integration) as well as the process of hiring and enculturation of new employees.

## 2. Using games to represent interdependence structures

**Designing a Culture of Collaboration: When Changing Beliefs is (Not) Enough**

Organizations exist to solve problems of collective action that prevent individuals from reaching socially optimal outcomes (Miller, 1992; Chatman & O'Reilly, 2016). The precise nature of the problem of collective action that individuals face depends on the nature of interdependence between them. Interdependence between agents in an organization implies that the returns to A of A's actions depend on B's action (and vice versa) (Puranam, Raveendran & Knudsen, 2012). It is a consequence of design choices about the division of labor (i.e. how to partition and allocate tasks), as well as the integration of effort (i.e. how to motivate and coordinate them) (Raveendran, Puranam & Warglien, 2015).

We take it as a premise that a fruitful understanding of how cultural design interventions operate cannot be achieved without reference to the underlying interdependence structure between agents. Game theory provides a useful and intuitive set of representations of interdependence structures that highlight unique challenges to collaboration (Camerer & Knez, 1997; Knez & Camerer, 1994). The utility of representations of interdependence in the form of games has been recognized by scholars across a variety of fields broadly related to organizations (e.g. Kelley & Thibault, 1978; Dawes, 1980; Ostrom,1990; Kollock,1998).[2]

A general partitioning of game structures is into those that feature challenges to cooperation, challenges to coordination, and both (Camerer & Knez, 1997). Note that it makes little sense to talk of pure cooperation or pure coordination success - the success of each presumes the other. Success at both cooperation and coordination is <u>necessary</u> for

[2] We note that adopting these representations of interdependence structure does not of course commit us to adopting the behavioral assumptions (e.g. common knowledge of game structures, rationality and previous play etc.) that game theorists often use to solve these games. We explain below how we diverge from classical game theory.

successful collaboration. However, failure in either cooperation or coordination is individually <u>sufficient</u> to generate collaboration failures.

The defining attribute of a game that features a cooperation challenge is that (at least some) players have to choose an action that produces a sub-optimal outcome for themselves, for the group to attain the best outcome (Camerer & Knez, 1997). The most famous of these games is the Prisoner's Dilemma, or the Public Goods game, which serves as the model for why groups may fail to build a public good or maintain a common pool resource, and why free-riding may occur in team tasks.

Insert Table 1 here

In Table 1, the Collaborative Outcome is achieved when both A and B cooperate and yields an individual payoff of $\delta$ (where $0<\delta<1$), but the Nash equilibrium involves both defecting (and yields an individual payoff of $0<\varepsilon<\delta$**)**. The challenge is to get agents to converge to the Collaborative Outcome, avoiding the temptations to free ride ($1-\delta$) and the opportunity cost of being exploited ($0-\varepsilon$) that drag them to the Nash equilibrium that yields payoffs of $\varepsilon$ to both.

Games that feature a pure coordination challenge lack barriers to cooperation because each player's payoff is maximized when the group's payoff is maximized. Matching (Schelling, 1960) and Stag-Hunt (Skyrms, 2004) are two important instances of pure coordination games. The matching game popularized by Schelling (1960) is one where both players are better off as long as their actions match (both pick action 1 or action 2), but it does not matter which particular actions they choose (Table 2). Standardizing on one among multiple equivalent choices is an instance - what matters is that we reach agreement, but what we agree on is not important. There are thus two Collaborative

Outcomes possible in this case and yield the same average payoff (1.0). There is no incentive to deviate if each player knows for sure what action the other player will take (i.e. each Collaborative Outcome is a Nash equilibrium). The difficulty lies in finding and converging to one, to avoid the mis-coordinated choices that lead to payoffs of $0<\delta<1$.

Insert Table 2 here

The "stag-hunt" game represents another interdependence structure in which the key challenge to collaboration arises from the difficulty of coordinating on the best combination of choices (Table 3). The game gets its name from a vignette which describes the advantages to two hunters to collaborate and hunt stag (action 1), rather than individually hunt hare (action 2). The Collaborative Outcome in Table 3 arises when both A and B pick action 1 and yields an average payoff of 1. However, each hunter also faces an opportunity cost ($\delta-1$, where $0<\delta<1$) if they show up alone to hunt stag while the other chooses to hunt hare (Camerer, 2003; Skyrms, 2004). There are two Nash equilibria in this game, the Collaborative Outcome as well as the case where both players choose action 2 (yielding an average payoff of $0<\varepsilon<\delta$). The challenge in this case is to get the agents to avoid the "wrong equilibrium" and converge to the right one.

Insert Table 3 here

Adoption of a new organizational practice features this interdependence structure. All parties are better off if all parties adopt the practice, but there are penalties to unilateral adoption. To the extent there is doubt about others adopting, the outcome may be that none adopts. Note that unlike the cooperation problem captured in a public goods problem, there is no gain from withholding collaboration (in this case, adoption) when others are known to be acting collaboratively (i.e. pick the high payoff-action in stag-hunt, or any particular action in matching). Rather, the uncertainty about what others will do is the key constraint on successful collaboration in pure coordination games like matching and stag-hunt.

Finally, there are also mixed motive games that feature both cooperation and coordination challenges. The Battle of the Sexes- (also popularized by Schelling, 1960) is a prototype for this kind of game (Table 4). Like the matching game, players prefer to match actions rather than not; however, they are not indifferent between the actions, as each player prefers to match on a different action. While A may choose to coordinate on action 1, B prefers action 2. The name of the game comes from the story used to motivate it – in which a husband and wife prefer to spend the evening together, but one prefers opera, the other prefers sport. For the pair to coordinate successfully, at least one must make a choice that leads to a sub-optimal outcome ($0<\delta<1$) for themselves. This is the cooperation challenge. At the same time, the uncertainty around what the other agent will do, which typifies the coordination challenge, is also present. Mis-coordination yields payoffs of $0<\varepsilon<\delta$ to both agents. In Table 4, there are thus two Collaborative outcomes (and each is a Nash equilibrium) yielding an average payoff of ½($\delta$+1), but each agent prefers a different one.

Insert Table 4 here

Bargaining is the prototypical organizational situation captured by Battle of Sexes. While both parties would like to strike an agreement (i.e. coordinate on an equilibrium), they may prefer different equilibria.

These four simple interdependence structures are by no means exhaustive of the possible two-player, two choice games one can imagine (e.g. Rappoport & Guyer, 1966; Kilgour & Fraser, 1988 report 726 possible games). However, these are the basic interdependence structures that analysts of organizations from a range of disciplinary backgrounds have found most useful to describe a range of commonly occurring organizational phenomena - i.e. shirking and free riding in teamwork (Public Goods), bargaining (Battle of Sexes), adopting a convention among many equivalent ones (Matching), and adopting a costly but useful new practice (Stag-Hunt) (Camerer & Knez, 1997). They also capture variation along two fundamental dimensions of social interaction– the need to subjugate individual interest to achieve a collectively optimal outcome (Public goods and Battle of sexes) and the need to anticipate other's actions (Battle of sexes, Matching, and Stag Hunt). That is why we focus on these.

## 3. Cultural design interventions: shaping beliefs (vs. shaping payoffs)

We take as a premise that a designer's goal is to get the members of an organization to pursue collaborative actions. However, given the limits of observability and enforcement, rather than directly constrain or order the required action, we focus on cultural interventions that shape the beliefs of agents about the value of taking different actions.

We conceptualize each actor as having beliefs[3] about what she should do, expressed as probabilities – such as, "I should choose action 1 with probability .7 and action 2 with probability .3". An agent's beliefs, or equivalently *action propensities*, are likely to be based on her knowledge, assumptions, habits, values, and the norms or expectations that she perceives her context to present. Initial beliefs (i.e. beliefs that have not yet been adjusted through feedback) are referred to as *priors*. The payoff from actions taken, not

[3] We choose to use the term 'belief' because it corresponds best to the game theoretic notion of beliefs as action propensities. The more accurate psychological term is 'attitude'.

known to the agent initially but revealed through interaction, serve as inputs to adjust her beliefs.

Cultural interventions may either modify beliefs or payoffs (or both). We explain below the difference between the two, and also why we choose to focus in this analysis only on cultural design interventions that shape beliefs about the consequences of actions, and not on interventions that affect outcomes of actions (i.e. the payoffs).

Interventions involving a change to payoffs have been extensively studied in a range of literatures. Researchers studying repeated public goods games finds that indefinitely repeated interaction changes payoffs to make cooperation more attractive even for self-interested players who share an accurate understanding of the situation (Kreps, 1996; Camerer, 2003). For instance, the prospect of indefinite repetition with a tit-for-tat strategy changes a one shot public goods game into a stag-hunt game of pure coordination (Camerer & Knez, 1993). Collaborative outcomes can also be sustained in public goods games by increasing rewards for cooperation or penalties for defection (Miller, 1992; Ostrom 1998). This can be achieved in organizations through compensation and incentive systems that increase the rewards to cooperation and penalties for defection above and beyond the payoffs dictated by the unaltered interdependence structure.

In one shot stag-hunt games, the penalties for unilateral collaboration can be reduced through a “modularity” strategy. Modular designs involve creating positive payoffs from contributing to collaborative activities even when others do not reciprocate (Baldwin & Clark, 1997; Baldwin & Clark, 2000; Langlois, 2002; Sanchez & Mahoney, 1996). Another structural intervention to change pay-offs in coordination games is to introduce an order of action. Leaders or managers can do this by acting first, that is, leading by example (Güth et al., 2007; Van Den Steen, 2010b). The possible courses of action and resulting payoffs are then simplified for the followers. This can be useful in mixed motive games like Battle of Sexes.

The subjective payoffs that actors experience can also be altered independently of the material payoffs. This can be achieved through normative control, that is, by engineering values and norms that motivate cooperative behaviors and by creating symbolic rewards for a moral orientation to the organization (Kunda, 1992). For instance, guilt or shame may lower the payoffs from the non-cooperative action in a public goods game, and from the safe action in a stag hunt. A sense of social identity and duty may raise the psychological payoffs from cooperating (Pillutla & Chen, 1999; Akerlof & Kranton, 2000).

Goal framing theory suggests that cognitive frames can alter the perceptions individuals have about the value of payoffs, for instance making them perceive a Battle of Sexes game as a Matching game (Foss & Lindenberg, 2013; Lindenberg, 2013; Lindenberg & Foss, 2011). Sorting is a powerful mechanism to ensure that players with the "correct" evaluation of payoffs interact with each other (Van den Steen, 2010b). Actors who value norms of cooperation are also likely to be willing to sanction others for non-cooperation. Thus, they not only perceive the payoffs to actions differently than actors who do not share these values, but through their willingness to impose penalties on others for norm violation, effectively change the (material) payoffs for others. [4]

[4] The collective action problem involved in punishing norm violation is itself a different game, and as such our analysis can be applied to understand how interventions that shape priors and their evolution through feedback affect behavior in such games.

**Designing a Culture of Collaboration: When Changing Beliefs is (Not) Enough**

In our analysis, we choose ***not*** to focus on design interventions that change payoffs through any of the channels noted above. Such interventions to change payoffs, whether arising through changes to incentives or through changes to "states of mind" (Barnard, 1938) are a) already well studied and b) amount to changing the interdependence structure. For instance, when repeated games turn public goods games into stag-hunts or normative goal frames turn battle of sexes into matching games, the resulting interdependence structure has changed.

Instead we model cultural interventions that affect beliefs (as action propensities) only and assume that the feedback that agents get as a result of their actions always reflects the payoffs seen in the interdependence structure. We can view our analysis as informative about how much of cultural integration can be accomplished by shaping beliefs alone, and when changes to payoffs (either through cultural or other design instruments- such as task re-design or incentive structures) may be necessary. It can also indirectly tell us when changes to payoffs are sufficient (i.e. when no major intervention to alter beliefs is necessary to obtain collaborative action). In our analysis, we therefore focus on design interventions that affect agents' initial beliefs and the process of changing beliefs as a function of feedback. Interventions that can accomplish such changes include socialization (that shape priors), interventions to sensemaking (that shape the degree of adjustment to feedback), and interventions through sorting (that can do both). In the next section, we explain how these design tools may be employed.

## 4. Modeling culture design as the development of beliefs that support collaboration

In classical game theory, agents know the pay-offs that they themselves can expect and also the payoffs that their counterparts expect under different combinations of their actions. Common knowledge of these payoffs coupled with common knowledge of rationality is typically necessary to provide a basis for accurate beliefs as to which action is likely to yield a better outcome (Cremer, 1993).

Our treatment of beliefs as action propensities diverges from this usage. We do not make any strong assumption about the knowledge or cognitive capacities of agents. First, we do not assume that agents' beliefs are valid. That is, we do not assume that they correctly reflect the payoffs. We allow for the possibility that cultural intervention can rely on creating invalid beliefs about payoffs that may lead to collaborative action. Second, we assume that beliefs are probabilistic. Agents may recognize that their knowledge about the payoff structure, perceptions of other player's intentions, or any other relevant assumptions that underlie beliefs are subject to errors. Therefore, they are not likely to have only absolute beliefs (action propensities that are 0 or 1). Third, we do not assume that beliefs are stable. Instead we assume that they are updated by learning from feedback on outcomes for self. Because the feedback that an agent receives from her actions will depend on the actions that the other agent took, agents may never perceive the 'true' payoffs of the interdependence structure.

This approach requires a study of the dynamics of how initial non-absolute beliefs (priors) are updated through interaction, and when the resulting beliefs are a) stable b) support collaborative action, as well as d) how rapidly these are formed. This can be done by using a formalization based on computational models of mutual adjustment (Lave & March 1975; Lounamaa & March, 1987; Knudsen & Srikanth, 2014; Puranam & Swamy, 2015; Lee & Puranam, 2016). Accordingly, we model the effects of the instruments of cultural design - such as sorting, sense-making, and socialization- as changing the priors and adjustment rates in such interactions.

We consider two agents, A and B who face an interdependence structure, **I** (which can be one of the four that we have discussed- public goods, battle of sexes, stag-hunt or matching). At t=0, A and B each has "priors"- initial assumptions about the payoffs and the possible actions of other agents. These translate into a probability distribution over the two actions available to them. We denote by $p_{i,k,t}$ the probability of agent i taking action k at time t.

***Interventions that shape priors***

Organizations can shape the priors of agents ($p_{i,k,0}$) through socialization. Van Maanen and Schein (1977: 4) explain what socialization involves:

> "To come to know an organizational situation and act within it implies that a person has developed some commonsensical beliefs, principles, and understandings, or in shorthand notation, a perspective for interpreting one's experiences in a given sphere of the work world. As Shibutani (1962) suggests, it provides the individual with an ordered view of the work life that runs ahead and guides experience, orders and shapes personal relationships in the work setting, and provides the ground rules under which everyday conduct is to be managed. Once developed, a perspective provides a person with the conventional wisdom that governs a particular context as to the typical features of everyday life."

Collective socialization efforts such as basic training or boot camp in military organizations or onboarding of new employees and management training are thus designed to shape the priors of entire cohorts of recruits (VanMaanen & Schein 1977). Leaders also actively shape priors. As van Ginkel and van Knippenberg note: “group leaders are likely to both voice and enact their own task representations in group task performance, and thus shape group members' task representations in the process. In group interaction group members are thus likely to develop task representations that are similar to their leader's task representations.” (2012: 96). In all these instances, the result is to shape the initial beliefs that agents “bring to the table” – about the structure of interdependence as well as how others are likely to act- when they commence interacting with each other.

***Interventions that shape the process of belief revision***

Once A and B take actions, they receive payoffs that depend on their choices and the interdependence structure they face (which, we assume, does not change). In keeping with the idea that culture can work when formal systems fail, we assume that the agents cannot directly observe each other’s actions, precluding verifiability. In general, knowledge-related work hinders mutual observability of action. Differences in technical background, organizational and physical location also serve to constrain communication and visibility of actions across agents (e.g. Dougherty, 1992).

Based on feedback received, the agents then update their probabilities of taking each action via *reinforcement learning* (Thorndike, 1911). This is a learning process with robust psychological validation in task environments with information constraints, and is used in a variety of models in psychology, computer science, organization science, and evolutionary biology (Bush & Mosteller, 1955; March, 1991; Domjan, 2010; Nowak &

Sigmund, 1993; Sutton & Barto, 1998; Erev & Roth, 1998; also see Puranam et al., 2015 for a recent review of the evidence).

A common approach to model an agent's reinforcement learning is the exponential (recency weighted) average rule (Sutton & Barto, 1998). We use this to model updating of the probability $p_{i,k,t}$ of agent i taking action k at time t as follows:

$$p_{i,k,t} = p_{i,k,t-1} + \phi_i [r_t - p_{i,k,t-1}] \qquad (1)$$

where $\phi_i \, \varepsilon [0, 1]$ is the adjustment/learning rate (Sutton & Barto, 1998) and $r_t$ is the payoff received. With the new values of $p_{i,k,t}$ A and B choose again, and the process can repeat for T periods. This formulation captures three key behavioral regularities in reinforcement learning that have substantial empirical evidence for them (e.g. Erev & Roth, 1998; Puranam et al., 2015):

a) Thorndike's Law of effect – which also implies exploration in choice (i.e. the ***probability*** of taking an action increases in rewards from taking it)

b) Power law of practice (i.e. eventual flattening out of learning curve; note that p increases/decreases slower as it approaches 1 or 0)

c) Watson's Law of recency (i.e. recent outcomes have different (greater) weight (phi) than outcomes in the distant past)

We have selected interdependence structures (Tables 1-4) to be such that payoffs are always in the range [0,1], ensuring that the updated values of $p_{i,k,t}$ also lie in this range. This is a convenience to aid interpretation as the joint probabilities can directly be interpreted in terms of the chances for successful collaborative action. In general, it is possible to allow for any payoff values and update propensities for choices that then translate into probabilities as proportions of propensities (e.g. Erev & Roth, 1998; Sutton

& Barto, 1998; Puranam & Swamy, 2015). Given binary choices, our simpler formulation in terms of probabilities being updated is sufficient (also see Bush & Mosteller, 1955).

The process we model is adaptive to feedback on own past actions rather than forward looking. As with many reinforcement learning models, it has a (in this case, tunable) degree of recency salience – the extent to which recent feedback is important relative to past feedback (via $\phi_i$). It requires no assumptions about common knowledge of game structure or behavioral rules used by others. We can say agents tend to take actions that look best in their own belief systems at any given point in time, but the tendency is not absolute (agents can deviate from what looks optimal), and beliefs can be unstable and erroneous.

This process of mutual adjustment that involves periods of interaction followed by adjustment of assumptions followed by interaction again, captures the essence of the process of sense-making (Louis 1980, Weick et al., 2005). For instance:

> "… sense making can be viewed as a recurring cycle comprised of a sequence of events occurring over time. The cycle begins as individuals form unconscious and conscious anticipations and assumptions, which serve as predictions about future events. Subsequently, individuals experience events that may be discrepant from predictions. Discrepant events, or surprises, trigger a need for explanation, or post-diction, and, correspondingly, for a process through which interpretations of discrepancies are developed. Interpretation, or meaning, is attributed to surprises. Based on the attributed meanings, any necessary behavioral responses to the immediate situation are selected. Also based on attributed meanings,

understandings of actors, actions, and settings are updated and predictions about future experiences in the setting are revised. The updated anticipations and revised assumptions are analogous to alterations in cognitive scripts." (Louis, 1980: 241). [5]

The key design parameter in this sensemaking process is $\phi_i$ - the learning rate of each agent that can be independently increased or decreased. This can be accomplished by (differentially) changing the sensitivity of agents to payoffs without changing the payoffs themselves. Differences in the attitude towards feedback, the extent to which the priors are re-iterated every period by the designer, or even the frequency with which agents are allowed to change actions can produce differences in the extent of adjustment of beliefs and behavior to feedback, which is captured by $\phi_i$. Differences in status and roles may also lead to differences in adjustment rates among agents. Not only do leaders convey credibility in their commitment to a course of action (as we showed is critical for socialization), but also remain resolute and unchanging in the course of action they have chosen. For instance, Bolton et al. (2013): argue that leaders should avoid changing direction and therefore should not seek too much feedback from others in the organization.

[5] Plausibly, socialization also covers ongoing learning through sensemaking, where the adjustment to assumptions follows from feedback on actions (see Van Maanen & Schein, 1977, Weick et al., 2005 for accounts of socialization where these are treated jointly; Gioia and Chittipeddi, 1991and Maitlis and Lawrence 2007 for discussion of "sensegiving" that covers both the influence of leaders on priors as well as on the subsequent learning process). We limit the effect of socialization to priors in order to examine its lasting impact on beliefs with and without the aid of sensemaking.

They model resolute leaders as trusting their own initial judgments more than rational leaders would and discounting subsequently learned information.

Both priors and adjustment rates can also be tuned through designed sorting of agents. In a recent study, Srivastava et al. (2017) find that organizational members that do not achieve a high level of cultural fit early on are asked to leave; conversely those that socialize but then reduce their fit leave voluntarily. They find that both voluntary and involuntary selection increases homogeneity in language use (their measure of cultural fit). Analogously, we assume it is possible to hire, retain and exit members of an organization on the basis of their priors or propensity to adjustment to feedback.

In sum, given the analytical structure we adopt, we can study cultural design interventions in terms of attempts to influence the assumptions of agents at the beginning of interaction (i.e. priors), and modifying the adjustment rates of agents; these represent the effects of interventions to socialization, sorting and sensemaking processes to form the organization of agents with particular priors and/or adjustment rates.

We note that our approach stands in contrast to that taken in typical game theoretic treatments (e.g. Gibbons, 1992) in which players have knowledge (or at least unbiased beliefs) of game structure (including the possibility of indefinitely repeated play), discount rates and know that the others know that they know and so on (i.e. common knowledge). The work on self-confirming equilibria (e.g. Fudenberg & Kreps, 1995; Fudenberg & Levine, 1993) is more closely related to our approach, in that it does not assume common knowledge of game structure. However, unlike this approach, we do not assume visibility on other's actions for focal actors (e.g., "fictitious play") in order to derive the self-confirming equilibrium. Instead, we let agents interact per our assumptions (which preclude mutual visibility of actions), and study if stable sets of beliefs and actions emerge (which may or may not be equilibria in the traditional sense).

## 5. Results: The efficacy of cultural design interventions that shape beliefs

In this section, we report results from computational experiments that vary priors ($p_{i,k,0}$) and adjustment rates ($\phi_i$) in theoretically meaningful ways across the four interdependence structures described in Tables 1-4. In all game structures, $1>\delta>\varepsilon>0$, where $\delta$ lies in (0,1) and $\varepsilon$ in (0, $\delta$), both drawn from uniform distributions. Thus, each game structure represents a family of games with the same interdependence structure, with realized values of $\delta$ and $\varepsilon$ determining the exact game.

We observe the effects of the manipulations of $p_{i,k,0}$ and $\phi_i$ on stability and "collaborativeness" of beliefs, as well as on outcomes (cumulative payoffs). We let pairs of agents interact for T periods, and average results across N=10,000 pairs of agents, each playing a different game from within the same game structure, to smooth out stochastic fluctuations. A period in our model is an interval of time during which beliefs have no opportunity to be revised. Its chronological duration may be quite small or large, nor does the duration have to be identical across periods. Our model describes any situation in which beliefs are periodically revised (or refined) on the basis of feedback (with the periods being of possibly long / short and unequal duration).

The model parameters that implement the interventions (socialization, sensemaking, and sorting) are presented in Table 5. These are constant across the four games. We compare the result of interventions to baseline ("un-designed") benchmarks. We believe that flat priors, where agents are indifferent between the two alternative actions

($p_{i,k,0}$=0.5), are a reasonable baseline for all games. Zero adjustment rates are a possible baseline. More realistically, adjustment rates can be non-zero even in the baseline case, though hard to specify precisely. Therefore, we compute results for various levels of adjustment rates (Table 6).

Before we turn to the consequences of different design interventions, we first illustrate the dynamics in our models for each interdependence structure over time. Figure 1 shows payoffs and beliefs (averaged across 10,000 pairs) over T=1000 periods, when priors are flat ($p_{i,k,0}$=0.5) and with symmetric adjustment rates of $\phi_i$ =0.1. Figure 1 shows that it is useful to consider the average payoffs and beliefs as well as variation around them. For instance, three of the game structures reach asymptotic average beliefs by 1000 periods (Stag Hunt takes much longer; in excess of 20,000 periods). However, there remains a lot of cross-sectional variation in payoffs at T=1000 across pairs of agents (and implicitly, specific payoff values within a game) indicating that the asymptotic average payoffs and beliefs are less reliably achieved in such cases.

For instance, in Figure 1 in Matching, the average beliefs are 0.50 even though payoffs approach 1.0 and standard deviation in payoffs approaches zero. (The expected payoffs under collaboration in this game is 1.0, zero for one-sided collaboration). This indicates that while nearly all pairs converge to a collaborative outcome, they may converge randomly to one or the other equilibria (i.e. within each sub-sample of pairs that reach a particular collaborative outcome, the average beliefs are either 1.0 or 0.0, and standard deviation in beliefs is near zero). In Stag Hunt, beliefs and payoffs at T=1000 are at around 0.75 and 0.80 respectively, with standard deviation in performance at about 0.25 (the expected average payoff under collaboration in this game is also 1.0, and 0.25 for one-sided collaboration). Thus, we may say that the specific set of parameters in Figure 1 produce a more effective “culture of collaboration” (because beliefs strongly favor one of the collaborative outcomes) more reliably and swiftly (because lower standard deviation in payoffs is attained much faster) in Matching than in Stag-Hunt.

In contrast, in Public Goods game, beliefs asymptote symmetrically to about 0.36, with average payoffs around 0.42 with an asymptotic standard deviation of around 0.24 (expected average payoffs under collaboration and one sided-collaboration are both 0.5). In Battle of Sexes, the beliefs of the first mover asymptote to 0.58, of the second to 0.42, and payoffs reach 0.51, with an asymptotic standard deviation of 0.31 (expected average payoffs under collaboration are 0.75 and is 0.25 under one sided collaboration). With this set of parameters, we could say the resulting culture is marginally more collaborative in Battle of Sexes than in Public Goods game (though neither could really be described as collaborative in absolute terms) but this outcome is less reliably attained in the former.

**a. Changing priors symmetrically**

Before considering our results for changing priors in detail, it is useful to briefly summarize what we know from prior work about how changing priors can aid collaboration. It is a well-known property of games with multiple equilibria that even when collaborative actions are an equilibrium, it may not be attained by forward looking agents who are fully informed of the structure of interdependence (e.g. in Matching, Stag-Hunt or Battle of Sexes). Instead, symmetry needs to be broken among equally plausible multiple coordination points. This can be accomplished spontaneously when certain equilibria are salient because they are focal points, as in the examples described by Schelling (1960). However, in the absence of a focal coordination point, someone or something that breaks the symmetry is required. Communication, particularly by leaders, is known to be capable of accomplishing this.

Studies in experimental game theory show that agents can try to create salience through ‘cheap talk’ – so called because the communication is not tied to payoffs (Farrell, 1987). Cooper et al. (1989) show in experiments with a one-shot symmetric Battle of the Sexes game, that pre-game communication helps coordination. They also show that it works much better if it is one-way rather than two-way communication. Drawing on such results, Foss (2001) argues that leaders are able to resolve pure coordination problems by creating common beliefs in their organizations. They break up 'epistemic stalemates' because their cheap talk is credible. “Thus if everybody believes that everybody else will move to a superior equilibrium when such a change has been communicated by the leader, and if it is common knowledge that players condition their behavior on the leader and that the leader himself gains by the change, then the leader’s communication is very credible indeed" (p 378).

Consistent with these arguments, Brandts et al. (2015) find that cheap talk by elected leaders is more effective than that by randomly assigned leaders in a turnaround game. Indeed, symmetry breaking by a leader may not require the leader to do anything. If their priors and the salience of their position is known, other agents can converge to their preferred action (De Kwaadsteniet & Van Dijk, 2010). Their salience is sufficient when the priors are known.

Our analysis however makes no assumptions about the interdependence structure being known to the agents, or of the agents being able to communicate to convey their intentions to each other. Instead, we assume that agents simply begin with (possibly erroneous) initial beliefs that they then adjust based on the feedback they receive. We assume that the organizational designer is external to the game; the designer can affect priors and adjustment rates (without necessarily being aware of the interdependence structure) but not the interdependence structure itself.

As Figures 2-5 show, we find that relative to flat priors (the baseline) aligning priors towards actions that yield collaboration (e.g. setting $p_{i,k,0}$=1.0) can aid the formation of assumptions that yield collaborative actions only in some interdependence structures-namely Matching and Stag-hunt. It is not effective when the interdependence structure is the Public Goods Dilemma or Battle of Sexes-where (at least some) players have to choose an action that produces a sub-optimal outcome for themselves, for the group to attain the best outcome.

Insert Figures 2-5 here

Creating convergent priors that encourage agents to pick collaborative actions is useful only in Matching and Stag-hunt (Figures 3 and 4), because in these games payoffs reinforce these priors but not in Public Goods game or Battle of Sexes (Figures 2 and 5). To see this, consider that a steady state in beliefs being attained implies $p_{i,k,t} = p_{i,k,t-1}$. This in turn implies that for all ϕ >0, $r_t = p_{i,k,t-1}$ (see equation 1). Setting priors $p_{i,k,0}$=1.0 is sufficient for this condition to be satisfied in Matching and Stag Hunt, leading to steady states; but this will not be the case in Battle of Sexes or Public Goods games.

**Insert Table 6 here**

Table 6 shows that at different adjustment rates {0.1,0.5,0.9}, the effect of aligning priors (relative to flat priors) is strongly visible only in Matching and Stag-hunt (compare within rows). Note that clashing priors never offer improvement over flat priors. The erosion of priors through interaction is more extreme in Public Goods games than in Battle of Sexes even at low adjustment rates because an improvement in payoffs will result after a deviation from collaborative priors in Battle of Sexes only if both agents deviate from the collaborative outcome, whereas in Public Goods, divergence will lead to an increase in pay-off if only the focal agent deviates. When payoffs do not reinforce the priors in Public Goods or Battle of Sexes, deviation from the collaborative outcome is unavoidable because the probability of taking the collaborative action falls below one. This causes agents to find "better" payoffs for themselves by deviating from the choices suggested by priors, making convergence to collaborative actions even harder. This is easier in Public Goods dilemma than in Battle of Sexes. [6] Thus, ***just aligning priors towards collaborative action is an effective intervention when the interdependence structures feature pure coordination challenges (but not when challenges to cooperation exist).*** If the payoffs from collaboration do not reinforce the priors, agents will explore non-collaborative actions, which in the case of Public Goods and Battle of Sexes will lead the agents away from collaborative choices.

### b. Changing adjustment rates symmetrically

The second intervention we consider is symmetric changes to adjustment rates for agents. We find that increasing adjustment rates symmetrically up to a degree is either

[6] We can get priors to be useful in Public Goods or Battle of Sexes games by curtailing exploration in choice (i.e. make assumptions absolute and unchangeable), and/or by setting adjustment rates to zero. Both seem unrealistic.

helpful or does not hurt across all games except Battle of Sexes (See Figures 2-5). Collaborative beliefs arise in Matching and Stag Hunt as learning rates increase. Interestingly, we also find that in Public Goods games, average beliefs become somewhat more collaborative with an increase in adjustment rate, though the variance in beliefs also rises (note high intra- as well as inter- agent variance in beliefs, Table 6, looking within the middle column for flat priors). In Battle of Sexes, average beliefs remain about the same though variance increases with adjustment rates (Table 6), and uniquely performance declines with an increase in adjustment rate (Figure 5).

Why is Battle of Sexes different? It is the only interdependence structure where payoffs under collaboration are not symmetric. This implies that greater symmetric sensitivity to payoffs can still have asymmetric effects on behavior of agents. A well-known pathology of high learning rates in situations of interdependence is the greater risk of learning from false negatives relative to chances of learning from true negatives or true positives (Lave & March, 1975; Puranam & Swamy, 2015). “False” in this context means feedback that discourages choosing actions that lead to collaborative outcomes, conditional on oneself selecting such an action. Increasing adjustment rates has a strong destabilizing effect in Battle of Sexes because a) it encourages the agent with the lower payoff from the collaborative outcome to switch and b) subsequently causes even the other agent to switch away from collaboration (due to the false negative generated).

In a Public Goods game, the switch away from collaborative outcome is more likely to be coordinated when it occurs, because payoffs are symmetric. When both agents switch to non-collaborative actions and receive the low payoff associated with this outcome, they are subsequently likely to switch back to collaboration in lock-step again. This is only probabilistically true of course, since the chance of picking the collaborative action is low to begin with in this game structure unless agents start with aligned priors. We can summarize this result as follows: ***symmetric increases in adjustment rates are weakly beneficial for collaboration performance in interdependence structures that reward collaboration symmetrically (and can be harmful when payoffs are asymmetric).***

**c. Asymmetric changes to adjustment rates and priors**

Following Lave and March (1975) who investigated matching games, a third intervention we consider is the impact of asymmetric adjustment rates. We plot the cumulative performance in each game when one of the agents has an adjustment rate of zero, for all levels of adjustment rate for the other agent ("asymmetric"), for cases where the agent's priors are either flat, convergent or clashing. In Figures 2-5, $\phi_A = 0$, and $\phi_B$ increases from 0.1 to 1 (note that asymmetric adjustment is undefined when both agents adjust at zero rate).

We find that the utility of asymmetric adjustment (relative to the case of symmetric adjustment by both agents) depends on priors: relative to the baseline (symmetric adjustment and flat priors) it is useful with clashing or aligned priors in Matching and Battle of Sexes (Figures 3 and 5), under aligned priors in Public Goods (Figure 2), and in Stag Hunt (Figure 4). Notably, asymmetric adjustment is never useful with flat priors.

Why should asymmetric adjustment combined with aligning or clashing priors be effective at improving collaboration across ***all four*** interdependence structures, when neither asymmetric adjustment by itself, or aligning or clashing priors by itself is effective? The answer lies in how asymmetries interact with the priors held by agents. Conflicting or aligned priors in Matching and Battle of Sexes ensures that the non-adjuster is anchored on a collaborative choice, while the other agent converges to her (see Figure 5). Once we have priors aligned towards the collaborative actions in Public Goods or Stag Hunt, the same condition (i.e. of fixing the non-adjustor to a collaborative action) obtains. ***Asymmetries in adjustment are beneficial, when at least one of the agents has a prior that promotes collaborative action, and the other agent adjusts "towards" this agent.***

## 6. Concluding remarks

Shaping the "states of mind" of an organization's members – the beliefs and assumptions about the consequences of their actions- is an important aspect of the design of cultures that promote collaboration (Barnard, 1937; Schein 2004). We have introduced a conceptualization and an accompanying modeling framework to study the conditions under which the beliefs of individuals interacting within distinct interdependence structures can be shaped to produce sustainable collaboration.

Our conceptualization distinguishes sharply between the "object of intervention" (i.e. beliefs) and the "context of intervention" (i.e. interdependence structure) and is our first contribution to the literature on the design of culture integration. We recognize that this decomposition of the problem - into what is designed (beliefs about payoffs) and what is not (actual payoffs) need not be the only possible one. As we have noted, an extensive prior literature has examined how payoffs can be altered in situations of interdependence to enable collaboration, and some of those interventions are also channeled through cultural changes. Nonetheless, we believe there is utility to our approach because it allows further clarity on when cultural interventions that merely shape beliefs are sufficient to produce collaborative action and when changes to payoffs are necessary. [7]

Our second contribution is the general point that cultural interventions that shape beliefs are not all created equal; their relative efficacy depends on contextual factors in a complex but predictable manner. Specifically, interventions that only shape the initial beliefs that interacting individuals bring into a collaborative situation, may be very effective when the key challenge to collaborative action is the difficulty of predicting what others will do (i.e. "epistemic interdependence" -leading to coordination problems; see Puranam et al., 2012). However, when the key challenge lies in the conflict between doing what is good for one's self versus what promotes the collaborative outcome, then shaping priors is insufficient to create a collaborative culture. In a similar spirit, interventions that affect rate of adjustment of beliefs to feedback are useful in some contexts but not others. Increasing the rate of symmetric adjustment is useful only when collaboration produces symmetric rewards. Asymmetric increases in adjustment rates, on their own, are never useful. In general, our analysis shows when the framing created by socialization will be reinforced by sensemaking vs. when it unravels as a consequence of it.

Our third contribution is to point to the remarkable power of interventions that jointly affect both priors and adjustment processes. We found that **combining asymmetry in adjustment with aligned priors is the best performing cultural intervention in every game structure.** Without altering actual payoffs, it is possible to always improve collaborative outcomes, at least modestly, "merely" by shaping beliefs (and the processes that shape beliefs). While just changing priors or just changing the adjustment process is

[7] If changes to payoffs were possible, our results suggest it is best to always convert every game to Matching or Stag-hunt.

effective only in some game structures, the combination is sufficient (though not necessary) across all. This result is more striking when one considers that each game structure is really a family of (infinitely) many games with specific payoff values that all obey a specified ordered and bounded relation.

Fourth, we find that even within interdependence structures, the nature of pre-intervention beliefs can matter. For instance, in situations captured by interdependence structures that resemble Matching or Battle of Sexes games, it may matter whether the interacting agents are drawn from different organizational units (in which case they are likely to have conflicting priors) or if they are both new hires (in which case they may have flat priors). In the former case, differences in adjustment rates are adaptive, but not in the latter.

A broader related point our analysis highlights is the necessity to use cultural interventions in a complementary fashion. While socialization can pre-dispose agents to collaborate and sorting can sustain collaborative behaviors by making likeminded agents encounter each other, feedback on actions through the sensemaking process will change beliefs sooner or later if there is any learning. Our analyses of interdependence structures with conflicts of interest show this clearly: the effect of cultural elements learned during socialization or set-up through sorting do not last when contradicted by experience generated through sensemaking.

Our results have implications for creating collaborative cultures across situations that pose conflicts of interest and those that feature pure coordination challenges. For instance, Puranam and Vanneste (2016) document different kinds of synergies in post-merger integration or between divisions within a corporate portfolio and note that some are more prone to conflicts of interest than others. The former includes situations of "one-sided" synergies (where one side gains while the other loses under the collaborative outcome) and of asymmetric dependence of one side on another. Rationalizing excess capacity is a classic instance, with one side having to give up something for aggregate gain. Our results suggest that cultural interventions to shape priors alone are likely to have modest effects at best in such contexts where there are conflicts of interest. Other interventions, involving incentives, the re-structuring of tasks, or possibly other aspects of culture (such as sorting out those with inappropriate values) are necessary. However, if asymmetry in adjustment rates can be combined with priors aligned towards collaboration, then one can make progress towards a "culture of collaboration" across a range of interdependence structures.

While we have kept the design of payoffs outside the scope of our study so that we could study the effect of beliefs under given and unchanging structures of interdependence, further studies may also want to consider a more complex relationship between beliefs and payoffs. By this we mean something beyond the point that payoffs themselves may be altered through cultural interventions that alter what is valued (e.g. by altering frames of perception, changing people's definitions of self and their identification with groups, or creating the risk of sanction by defining what is acceptable). Rather, one can imagine that psychological payoffs themselves change from the experience of successful (or unsuccessful) collaboration, above and beyond the material payoffs. The righteous pride of being a cooperative citizen or the guilt of failing others' trust can arise from a history of interactions and change payoffs dynamically, rather than being fixed within a given game. We believe this is a fruitful area for further investigation.

**Table 1. Public Goods (Prisoners Dilemma) payoff matrix**

| | **B chooses action 1** | **B chooses action 2** |
|---|---|---|
| **A chooses action 1** | δ, δ | 0, 1 |
| **A chooses action 2** | 1, 0 | ε, ε |

**Table 2. Matching payoff matrix**

| | **B chooses action 1** | **B chooses action 2** |
|---|---|---|
| **A chooses action 1** | 1, 1 | δ, δ |
| **A chooses action 2** | δ, δ | 1, 1 |

**Table 3. Stag hunt payoff matrix**

| | **B chooses action 1** | **B chooses action 2** |
|---|---|---|
| **A chooses action 1** | 1, 1 | 0, δ |
| **A chooses action 2** | δ, 0 | ε, ε |

**Table 4. Battle of the Sexes payoff matrix**

| | **B chooses action 1** | **B chooses action 2** |
|---|---|---|
| **A chooses action 1** | 1, δ | ε, ε |
| **A chooses action 2** | ε, ε | δ, 1 |

In all cases 1>δ>ε>0, where δ lies in (0,1) and ε in (0, δ).

**Table 5. Model parameters for cultural interventions on beliefs**

| Intervention | Model parameters |
|---|---|
| Modifying priors<br>(e.g. through socialization and sorting) | *Flat* priors (Baseline): $p_{A,k,0} = p_{B,k,0} = 1/2$<br><br>*Aligned* priors: $p_{A,k,0} = p_{B,k,0} = 1$<br><br>*Clashing* priors: $p_{A,k,0} = 1$, $p_{B,k,0} = 0$ |
| Modifying adjustment process<br>(e.g. through changing sensemaking, sorting) | *Symmetric* increase (Baseline): increasing $\phi_i$ from 0 to 1.<br><br>*Asymmetric* increase: $\phi_A = 0$, and $\phi_B$ increases from 0 to 1. |

**Table 6. Final period beliefs, their variation and stability at T=100**

**Clashing Priors**

| Phi | | A's Beliefs | B's Beliefs | Inter-agent sd+ | Intra-agent sd* |
|---|---|---|---|---|---|
| **0.1** | M | 0.52 | 0.48 | 0.31 | 0.15 |
| | SH | 0.63 | 0.63 | 0.24 | 0.13 |
| | BOS | 0.58 | 0.42 | 0.12 | 0.10 |
| | PG | 0.36 | 0.36 | 0.10 | 0.12 |
| **0.5** | M | 0.50 | 0.49 | 0.50 | 0.15 |
| | SH | 0.96 | 0.96 | 0.18 | 0.18 |
| | BOS | 0.58 | 0.42 | 0.24 | 0.21 |
| | PG | 0.40 | 0.40 | 0.21 | 0.19 |
| **0.9** | M | 0.51 | 0.51 | 0.50 | 0.11 |
| | SH | 0.99 | 0.99 | 0.09 | 0.17 |
| | BOS | 0.57 | 0.44 | 0.34 | 0.32 |
| | PG | 0.46 | 0.46 | 0.33 | 0.31 |

**Flat Priors**

| Phi | | A's Beliefs | B's Beliefs | Inter-agent sd+ | Intra-agent sd* |
|---|---|---|---|---|---|
| **0.1** | M | 0.50 | 0.50 | 0.32 | 0.13 |
| | SH | 0.64 | 0.64 | 0.24 | 0.17 |
| | BOS | 0.58 | 0.42 | 0.12 | 0.07 |
| | PG | 0.36 | 0.36 | 0.10 | 0.07 |
| **0.5** | M | 0.50 | 0.50 | 0.50 | 0.12 |
| | SH | 0.97 | 0.97 | 0.16 | 0.07 |
| | BOS | 0.58 | 0.43 | 0.24 | 0.20 |
| | PG | 0.39 | 0.39 | 0.21 | 0.18 |
| **0.9** | M | 0.50 | 0.50 | 0.50 | 0.08 |
| | SH | 0.99 | 0.99 | 0.07 | 0.04 |
| | BOS | 0.56 | 0.43 | 0.34 | 0.32 |
| | PG | 0.46 | 0.46 | 0.33 | 0.30 |

**Aligned Priors**

| Phi | | A's Beliefs | B's Beliefs | Inter-agent sd+ | Intra-agent sd* |
|---|---|---|---|---|---|
| **0.1** | M | 1.00 | 1.00 | 0.00 | 0.00 |
| | SH | 1.00 | 1.00 | 0.00 | 0.00 |
| | BOS | 0.63 | 0.47 | 0.14 | 0.13 |
| | PG | 0.36 | 0.36 | 0.10 | 0.16 |
| **0.5** | M | 1.00 | 1.00 | 0.00 | 0.00 |
| | SH | 1.00 | 1.00 | 0.00 | 0.00 |
| | BOS | 0.58 | 0.43 | 0.24 | 0.21 |
| | PG | 0.40 | 0.40 | 0.21 | 0.20 |
| **0.9** | M | 1.00 | 1.00 | 0.00 | 0.00 |
| | SH | 1.00 | 1.00 | 0.00 | 0.00 |
| | BOS | 0.57 | 0.44 | 0.34 | 0.32 |
| | PG | 0.45 | 0.45 | 0.33 | 0.31 |

+ Average inter-agent standard deviation indicates the difference in beliefs across dyads.

* Average intra-agent standard deviation indicates the fluctuation from period to period in agents' beliefs.

Figure 1: Illustrative system metrics at T=1000 (phi=0.1, flat priors)

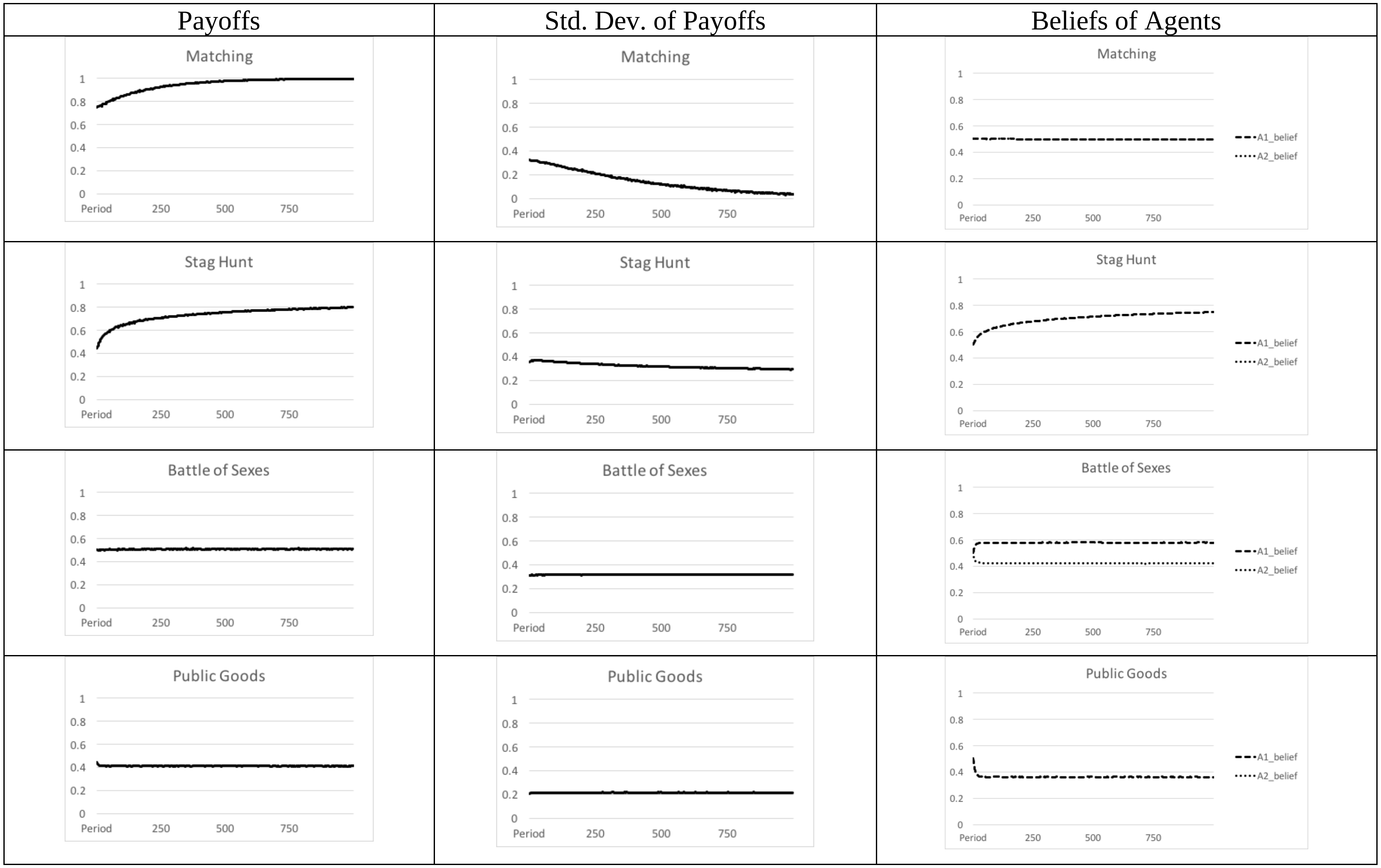


Figure 2. Cumulative Performance (y) vs. adjustment rates (x) in the Repeated Prisoners Dilemma Game (at T=100)

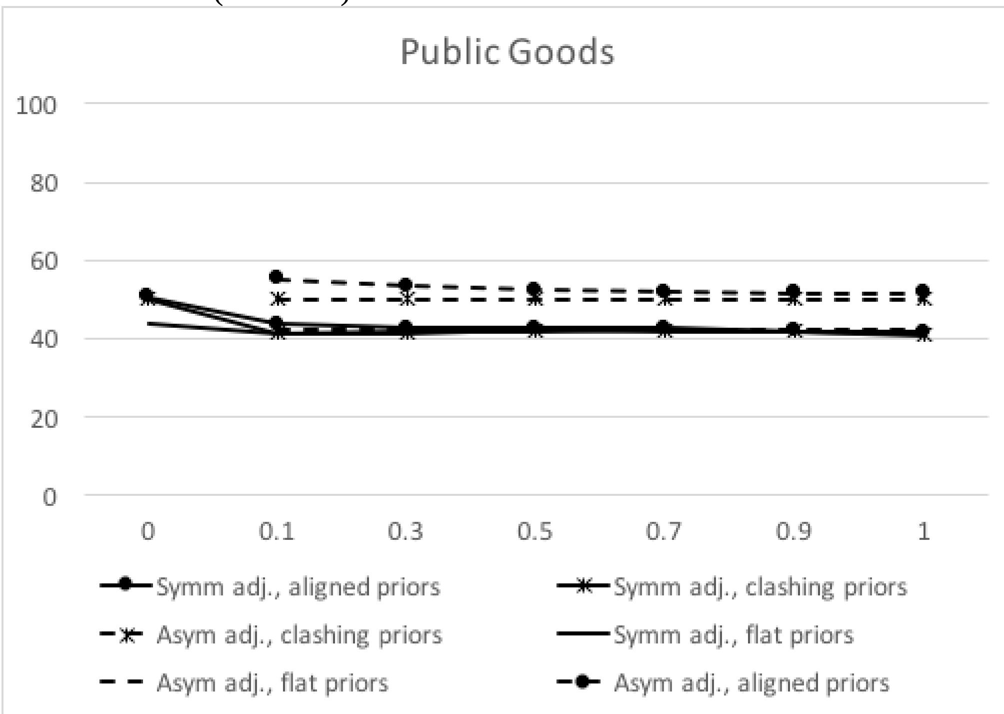



Figure 3. Cumulative Performance (y) vs. adjustment rates (x) in the Repeated Matching Game

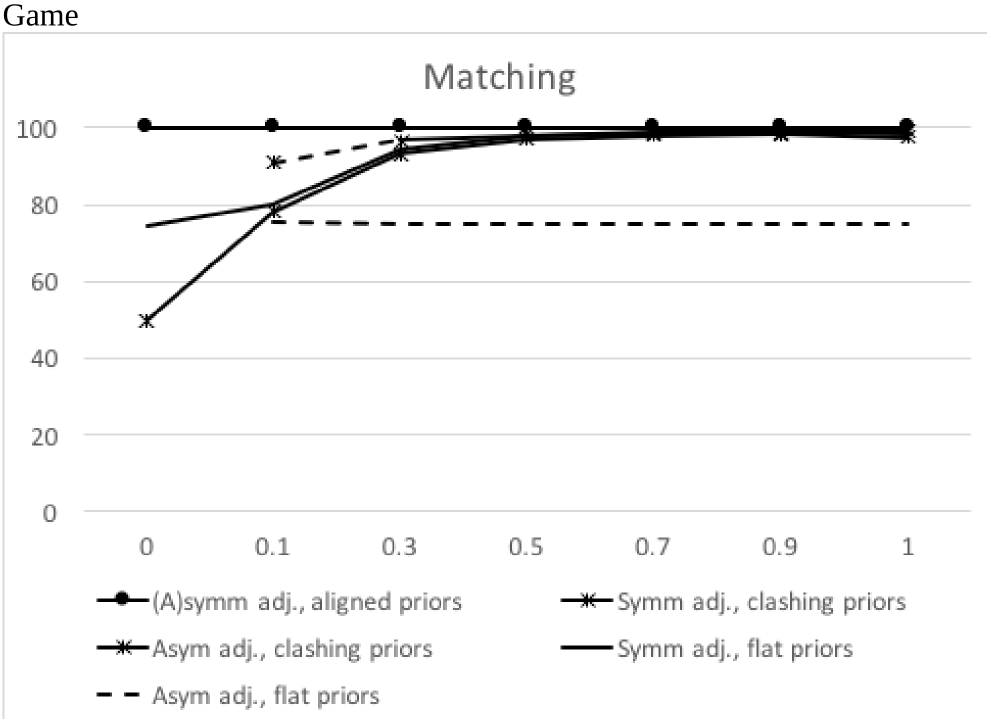


Figure 4. Cumulative Performance (y) vs. Adjustment rates (x) in the Repeated Stag Hunt Game

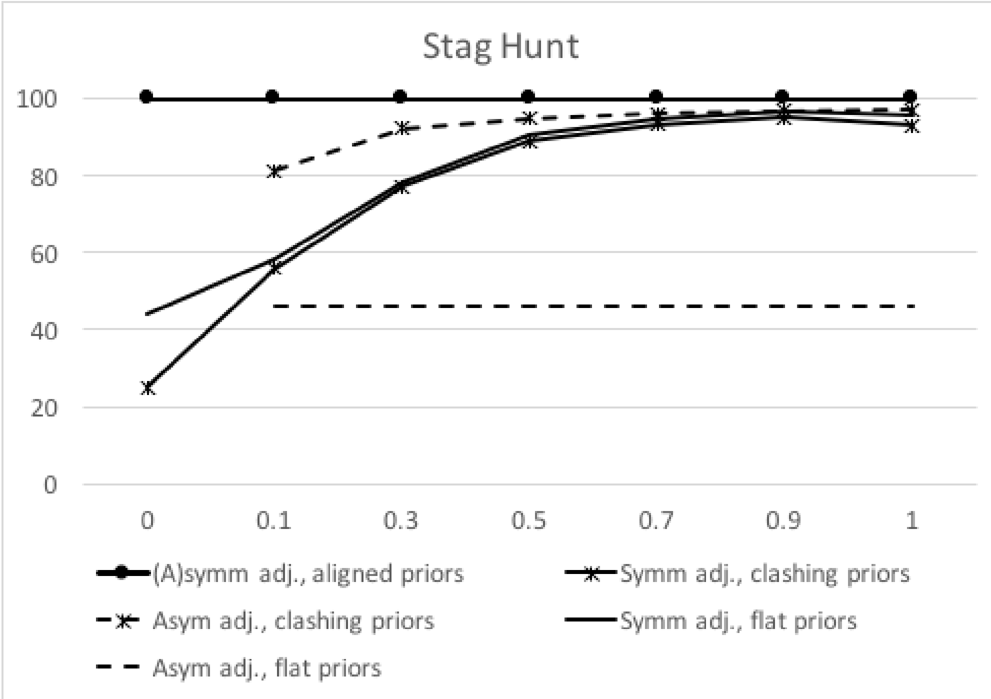
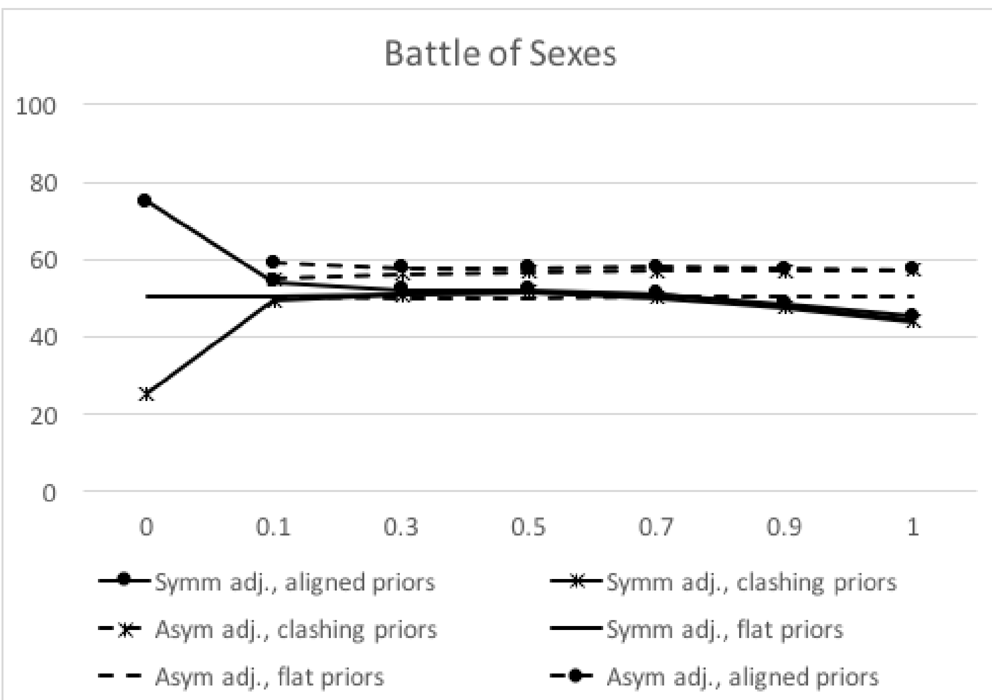



Figure 5. Cumulative Performance (y) vs. adjustment rates (x) in the Repeated Battle of Sexes Game



## References

Akerlof, G. A., & Kranton, R. (2000). "Economics and identity." *The Quarterly Journal of Economics* 115(3): 715-753.

Baldwin, C. Y., &. Clark, K.B. (2000). *Design rules: The power of modularity*. MIT press.

Barnard, C. I. (1968). *The Functions of the Executive*. Harvard University Press.

Brandts, J.C., David J. W., & Roberto, A. W. (2015). "Legitimacy, Communication, and Leadership in the Turnaround Game." *Management Science* 61(11): 2627-2645.

Bush, R. R., & Frederick, M. (1955). *Stochastic Models for Learning*. New York: John Wiley & Sons.

Camerer, C. **(**2003). *Behavioural Game Theory: Experiments in Strategic Interaction.* Princeton University Press.

Camerer, C., & Knez, M. (1997). "Coordination in organizations: A game-theoretic perspective." *Organizational Decision Making*, 158-188.

Cannon, B., Janis A., & Eduardo, S. (2001). "Reflections on shared cognition." *Journal of Organizational Behaviour*, 22: 195-202.

Chatman, J. A., & O'Reilly, C. A. (2016). Paradigm lost: Reinvigorating the study of organizational culture. *Research in Organizational Behavior*, *36*, 199-224.

Cooper, R., Douglas, DeJong, D., Forsythe, R Robert, F., & Ross, T.W. (1989). "Communication in the battle of the sexes game: some experimental results." The RAND Journal of Economics, 568-587.

Crémer, J. (1993). "Corporate culture and shared knowledge." *Industrial and Corporate Change* 2(3): 351-386.

Dawes, R. M. (1980). "Social dilemmas." *Annual Review of Psychology* 31(1): 169-193.

Kwaadsteniet, E.W., & Van Dijk, E. (2010). Social status as a cue for tacit coordination. *Journal of Experimental Social Psychology*, 46, 515-524

Domjan, M. **(**2010). Principles of learning and behavior, 6th edition. Cengage /Wadsworth

Dougherty, D. (1992). "Interpretive barriers to successful product innovation in large firms." *Organization Science* 3(2): 179-202.

Erev, I., & Roth, A. E. (1998). “Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria.” *American economic review*, 848-881.

Farrell, J. (1987). "Cheap talk, coordination, and entry." *The RAND Journal of Economics*, 34-39.

Foss, N. J. (2001). Leadership, beliefs and coordination: An explorative discussion. *Industrial and corporate change*, 10(2), 357-388.

Foss, N. J., & Lindenberg, S. (2013). "Microfoundations for strategy: A goal-framing perspective on the drivers of value creation." *The Academy of Management Perspectives* 27(2): 85-102.

Fudenberg, D., & Kreps, D.M. (1995). "Learning in extensive-form games I. Self-confirming equilibria." *Games and Economic Behavior,* 8(1): 20-55.

Fudenberg, D., & Levine, D.K. (1993). "Self-confirming equilibrium.*" Econometrica: Journal of the Econometric Society*, 523-545.

Gibbons, R. (1992). *Game Theory for Applied Economists*. Princeton University Press.

Gibbons, R., & Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization Science*, *23*(5), 1350-1364.

Gioia, D. A., & Chittipeddi, K. (1991). "Sensemaking and sensegiving in strategic change initiation." *Strategic Management Journal* 12(6): 433-448.

Güth, W., Levati, M., Sutter, M., & van Der Heijden, E. (2007). "Leading by example with and without exclusion power in voluntary contribution experiments." *Journal of Public Economics* 91(5): 1023-1042.

Harrison, J. R., & Carroll, G.R., (2006). *Culture and demography in organizations*. Princeton University Press.

Kelley, H. H., & Thibaut, J.W., (1978). *Interpersonal relations: A theory of interdependence*. John Wiley & Sons

Kilgour, D. M., & Fraser, N.M. (1988). "A taxonomy of all ordinal 2× 2 games." *Theory and Decision* 24(2): 99-117.

Knez, M., & Camerer, C. (1994). "Creating Expectational Assets in the Laboratory: Coordination in ‘Weakest-Link’Games." *Strategic Management Journal* 15(S1): 101-119.

Knudsen, T., & Kannan, S. (2014). "Coordinated exploration: Organizing joint search by multiple specialists to overcome mutual confusion and joint myopia." *Administrative Science Quarterly* 59(3): 409-441.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, *24*(1), 183-214.

Kreps, D.M. (1996). "Corporate culture and economic theory." *Firms, Organizations and Contracts, Oxford University Press, Oxford*, 221-275.

Kunda, G. (1992). *Engineering Culture: Control and Commitment in a High-tech Corporation*. Temple University Press.

Langlois, R. N. (2002). "Modularity in technology and organization." *Journal of Economic Behavior & Organization* 49(1): 19-37.

Lave, C. A., & James, G. M. (1975). *An Introduction to Models in the Social Sciences* New York: Harper & Row

Lee, E., & Puranam, P. (2016). “The implementation imperative: Why one should implement even imperfect strategies perfectly.” *Strategic management journal*, *37*(8), 1529-1546.

Lindenberg, S., & Foss, N. J., (2011). "Managing joint production motivation: The role of goal framing and governance mechanisms." *Academy of Management Review* 36(3): 500-525.

Louis, M. R. (1980). "Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings." *Administrative Science Quarterly*, 226-251.

Lounamaa, P. H., & March, J. G. (1987). "Adaptive coordination of a learning team." *Management Science* 33(1): 107-123.

Maitlis, S., & Lawnernce, T.B. (2007). "Triggers and enablers of sensegiving in organizations." *Academy of management Journal* 50(1): 57-84.

March, J. G. (1991). "Exploration and exploitation in organizational learning." *Organization Science* 2(1): 71-87.

Miller, G. J. (1992). *Managerial Dilemmas. The Political Economy of Hierarchy*. New York: Cambridge University Press.

Nowak, M., & Sigmund, K. (1993). "A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game." *Nature* 364(6432): 56-58.

O'Reilly, C. A., & Chatman, J.A. (1996). “Culture as social control: Corporations, cults, and commitment.” In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior: An annual series of analytical essays and critical reviews*, Vol. 18, pp. 157-200). US: Elsevier Science/JAI Press.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* Cambridge: Cambridge University Press.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American political science review*, *92*(1): 1-22.

Pfeffer, J. (1995). Producing sustainable competitive advantage through the effective management of people. *The Academy of Management Executive*, *9*(1), pp.55-69.

Pillutla, M. M., & Chen, Xiao-Ping. (1999). "Social norms and cooperation in social dilemmas: The effects of context and feedback." *Organizational Behavior and Human Decision Processes* 78(2): 81-103.

Puranam, P., Stieglitz, N., Osman, M., & Pillutla, M.M. (2015). Modelling bounded rationality in organizations: Progress and prospects. *The Academy of Management Annals*. 9(1): 337-392.

Puranam, P., Raveendran, M. & Knudsen, T. (2012). “Organization design: The epistemic interdependence perspective." *Academy of Management Review* 37(3): 419-440.

Puranam, P., & Swamy, M. (2016). "How initial representations shape coupled learning processes." *Organization Science* 27(2): 323-335.

Puranam, P., & Vanneste, B. (2016). *Corporate strategy: Tools for analysis and decision-making*. Cambridge University Press.

Rappaport, A and Guyer, M. (1966). "A taxonomy of 2 x 2 games." *General Systems* 11: 203-214.

Raveendran, M., Puranam, P., & Warglien, M. (2015). “Object salience in the division of labor: Experimental evidence." *Management Science* 62(7): 2110-2128.

Rouse, W. B. & Morris, N.M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, *100*(3), p.349.

Sanchez, R., & Mahoney, J. T., (1996). "Modularity, flexibility, and knowledge management in product and organization design." *Strategic Management Journal* 17(S2): 63-76.

Schein, E.H., Schneier, I., & Barker, C.H., (1961). *Coercive Persuasion*. New York: W.W. Norton & Co.

Schein, E.H. (1983). "The role of the founder in creating organizational culture." *Organizational Dynamics* 12(1): 13-28.

Schein, E. H. (1990). "A general philosophy of helping: Process consultation." *MIT Sloan Management Review* 31(3): 57.

Schein, E. H. (2004). *Organizational Culture and Leadership*. John Wiley & Sons

Schelling, T.C. (1960). *The Strategy of Conflict*. Cambridge, Mass.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Srivastava, S. B., Goldberg, A., Manian, V.G., & Potts, C. (2017). "Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations." *Management Science Forthcoming*

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. Vol. 135. Cambridge: MIT Press.

Thagard, P. (2005). *Mind: Introduction to Cognitive Science*, MIT Press, Boston MA.

Thorndike, E.L. (1911). *Animal intelligence: Experimental studies*. Macmillan.

Van den Steen, E. (2010a). "Culture clash: The costs and benefits of homogeneity." *Management Science* 56(10): 1718-1738.

Van den Steen, E. (2010b). "On the origin of shared beliefs (and corporate culture)." *The RAND Journal of Economics* 41(4): 617-648.

van Ginkel, W. P., & van Knippenberg, D. (2012). "Group leadership and shared task representations in decision making groups." *The Leadership Quarterly* 23(1): 94-106.

Maanen,V., Eastin, J., & Schein, E.H. (1977). "Toward a theory of organizational socialization." *MIT Working Paper Series*.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). "Organizing and the process of sensemaking." *Organization Science* 16(4): 409-421.