Collective Performance: Modeling the interaction of habit-based actions

Supplementary Materials

Version 2: 10 September 2010

Massimo Warglien, Universitá Ca Foscari de Venezia, [warglien@unive.it]
Michael D. Cohen, University of Michigan, [mdc@umich.edu]
Daniel A. Levinthal, University of Pennsylvania [dlev@wharton.upenn.edu][*]

[*] The delight of this collaboration has been each author's firm belief that the others made the essential contributions. Accordingly, the order currently shown does not represent an equilibrium. We are grateful for the contributions of many colleagues to the ideas presented here, and especially to Markus Becker, Phanish Puranam, and Jason Owen-Smith for comments on an earlier draft.

Table of Contents

Supplement S1: Modeling The Coordination of Perceptual Categories

Supplement S2: Modeling Action in the Transform the Target Game

Supplement S3: Modeling Coordination via Mutual Observation

Supplementary model S1:

Modeling The Coordination of Perceptual Categories

As a first model instantiating our framework, we introduce a simple representation of the situation space and of how agents categorize it (premises A-E). We show how the model, despite its simplicity, can capture the inherent flexibility of habits and their sensitivity to context. We also show how agents, despite having different individual prototypes of exemplary cases or situations, can learn through minimal changes to tune their perception of the situation space and to achieve greater coordination – a phenomenon clearly underlying the development of successful collective performance (premise L).

Our model could be conceived as a simple example of a model of "appropriate behavior" (March and Olsen 1989), i.e. behavior triggered by the recognition of the type of situation in which an agent finds herself. In this first example, we will concentrate mostly on the first stage of the problem: how a situation can be identified so that the corresponding appropriate action can be performed? (See also Gavetti and Warglien 2008) We will assume that actions are tied to situation categories in a one-to-one correspondence: an assumption that we will relax in later examples.

Thus, the core of this first model is the representation of the situation space and of categorical prototypes within it. We will assume that the "situation space" can be represented literally as a space – actually the product space of multiple original quality dimensions or properties that characterize a set of situations (see section 4.1 in the text). A well-studied psychological example is the space of perceived colors (Gärdenfors 2000); in the economic context one might consider the space representing possible combinations of product fearues. A single situation can be located as a point in such space. Similarity between situations is defined (following a broad psychological literature) in terms of spatial proximity: the closer the points representing two situations, the more similar to each other they are perceived to be (Shepard 1987) (Nosofsky 1988)[1].

We further assume that agents tend to think coarsely about the situation space, aggregating broad sets of situations that have “family resemblance” in categories (Rosch and Mervis 1975). Categories are often structured around prototypes (Rosch and Mervis 1975), i.e. cases which are “best examples”, or initial examples, of the category. In our model, a prototypical situation is a case highly representative of a class of situations judged by an agent as being “of the same type”. Once they are imprinted in memory, prototypes tend to generate quite automatically (and with reduced memory load) a decomposition of the situation space into categories: a situation tends to be attributed to the category whose prototype is most similar to it.

A simple and elegant geometric model (Gärdenfors 2000) captures the essence of the above statements. If prototypes are points in the situation space and a category is defined as the set of situations to which a given prototype is most similar, the situation space will be naturally decomposed into (convex) regions that represent categories of situations around their respective prototypes (Figure S.1). This decomposition is called a Voronoi diagram, or Voronoi tessellation (Okabe et al. 1992).

[1] We do not wish to enter here the controversy on similarity measures originating with Tversky (1977). For a critical appraisal and a defense of geometric measures of similarity, see Gärdenfors (2000).

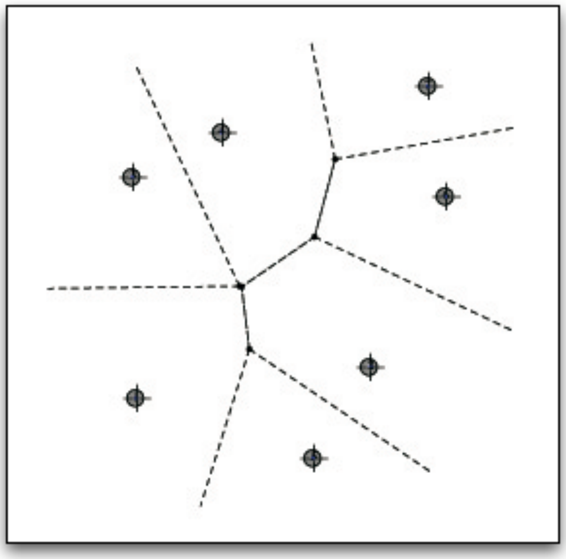


Figure S.1. A Voronoi diagram

Thinking by categories helps to explain the inherent flexibility often observed in habitual behavior and routines: as long as situations are perceived as similar, cases never actually seen previously can be smoothly handled by recognizing the category to which they belong. For example, credit authorization routines for customers with similar measures of credit worthiness will trigger similar actions that in turn generate similar loan contracts.

Moreover, it is possible to introduce much more subtle forms of flexibility by minimal extensions of this basic model. It has been repeatedly observed (Barsalou 1987) that when engaged in similarity judgments, individuals do not give unvarying weights to the different attributes along which similarity is assessed. Instead, they tend to make an adaptive use of attribute weighting. For example, in tasks in which one must discriminate between two categories, individuals tend to emphasize or increase the salience of attributes along which prototypical exemplars are most different (Nosofsky 1988).

Our geometrical model of the situation space can easily capture this adaptive use of salience by introducing weighted dimensions of similarity (Nosofsky 1988) (Gärdenfors 2000). If the weight of the dimensions of the situation space can be changed, distance and thus similarity between situations will change accordingly, and the whole decomposition of the situation space into categories will be modified.[2]

Figure S.2 provides a simple two-dimensional illustration of this effect. Five prototypical points define categories that shift in space as the relative weight (w) of the left right dimension increases. As can be clearly seen, the attribution of a situation (the black dot) to categories shifts accordingly.

The sensitivity of the categorization of situations to salience changes suggests a very simple process that makes similarity judgment – and thus action triggering – responsive to the context, providing further smoothness to habitual behavior and illustrating the "tunable" nature of the action function (see section 4.2 in the main text). Indeed, salience tends to respond to important environmental cues. For example, seeing a car accident is likely to emphasize an individual's perception of the risk dimension of driving situations, modifying the classes of situations in which a driver finds it appropriate to reduce speed. In general, sensitivity to context via salience will produce a typical fuzziness of categorical attribution in the "peripheral" regions of categories (situations which are located close to the boundaries), while the category membership of situations close to prototypes will tend to remain stable. It is important to point out that these changes can happen without any modification of the categorical prototypes

[2] Those familiar with the principles of geometry will recognize this as simple affine transformation of the state space.

– just as a result of shifts in the salience of dimensions. They may also be largely unconscious, as when changes in salience are triggered by emotions.

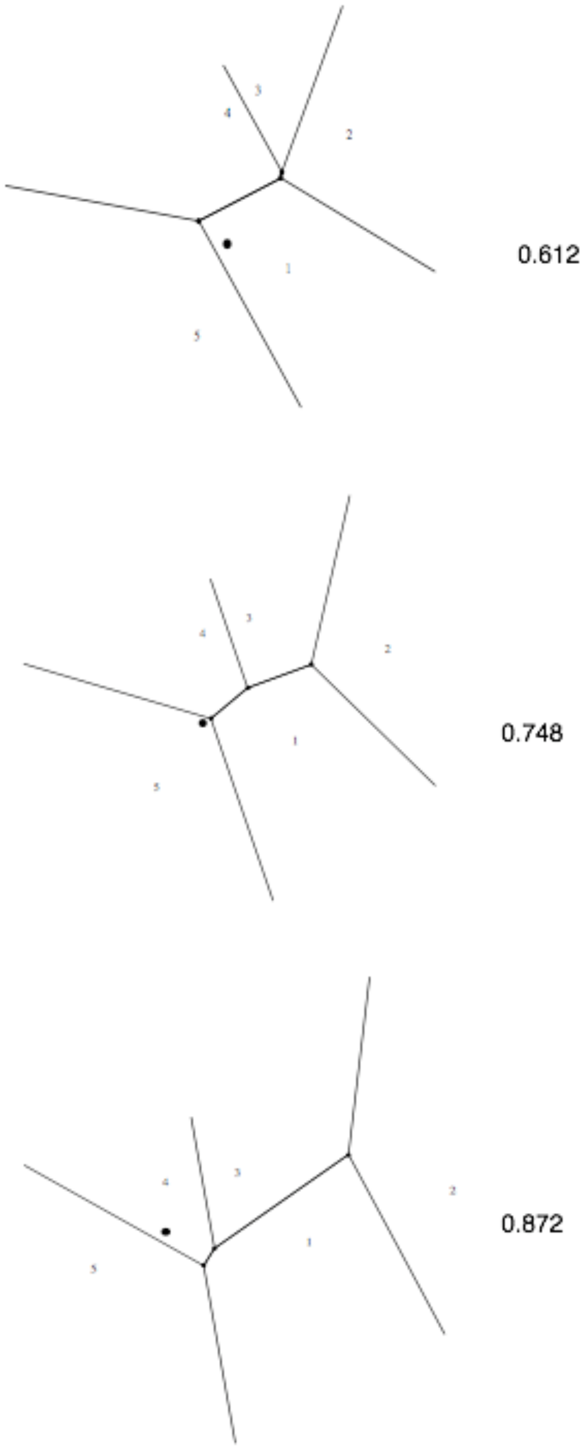


Figure S.2: Voronoi diagrams of space with five prototypes (shown by numbers). The situation (black dot) is categorized differently at each of three different levels of relative weight w for the left-right dimension.

The model above can be extended to multiple agents. Indeed, it allows us to capture two sources of heterogeneity in their perception of situations: differences in prototypes, and differences in the weight accorded to attributes. Some weights might be equal to zero, bringing within the scope of the model instances where agents consider different dimensions of a situation.

Clearly, differences in the way situations are categorized may be a fundamental source of coordination failures. We show that a basic form of incremental salience adjustment may markedly improve coordination between agents who hold heterogeneous perceptions of the situation space in the absence of any modification of their prototypes.

For sake of simplicity, imagine a problem in which there are two agents, and two actions (respectively: {A, B} and {a, b}) available to each one of them. Actions are complementary, so that they produce a coordinated outcome only when their combination is <A, a>, both agents choose the first of their two actions, or <B, b>, both agents choose the second of their two actions. Furthermore, agents have two prototypes each, which are related to actions. So when the situation is perceived as being of type α, agents will trigger, respectively, actions A and a; alternatively, when the situation is perceived as being of type β, agents will trigger, respectively, actions B and b. The problem arises from the fact that the prototypes, and therefore the category boundaries, for the situations they label 'α' and 'β,' may be different between the two agents (see Figure 3). Furthermore, agents may start with different weights for the attribute dimensions of the situation space.

Now, imagine that situations are encountered by the two agents, and at each new situation they act independently and observe whether their actions were coordinated or not. Whenever coordination failures are experienced, each agent will react by dampening the relative weight (salience) of the dimension that contributed most prominently to the coordination failure.

Figure S.3 shows how coordination failures due to the initial lack of alignment of individual categories are corrected by this simple process of "salience tuning". The graph shows the evolution of the average number of coordination failures as 500 different random situations are sequentially shown to a pair of agents, with substantial improvement occurring quite quickly.[3]

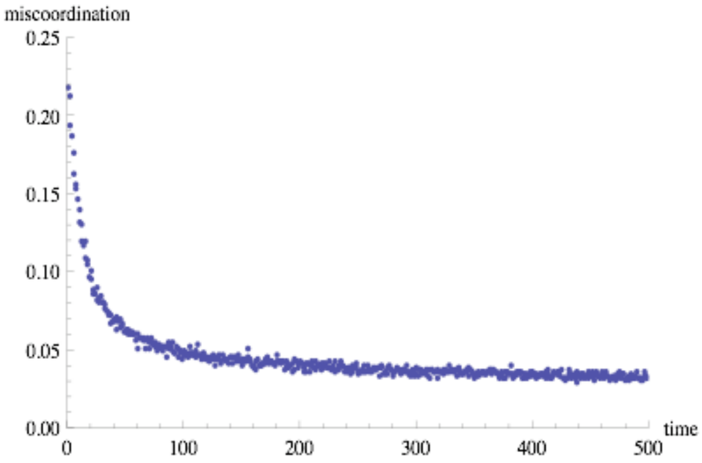


Figure S.3 Decreasing coordination failures with salience tuning.

Figure S.4 compares the initial and final categorizations of the space in a single sequence of a pair that has learned over 500 different situations. The alignment of categories (despite the persistent heterogeneity of the prototypes) can be clearly seen.[4] Further notice the rather rapid mutual adaptation process:

[3] Averages are computed over 100 different random sets of pairs of prototypes and initial weights of the space dimensions. In turn, for each set of prototypes and weights, 100 different series of 500 random situations are run.

[4] Like most learning processes, though, this one will work only within some boundary conditions. It requires that the sets of prototypes of the agents be ordered in the same way on any dimension (order-monotonicity). For example, prototypes of a category X should be North-East of those of a category Y for both agents. In other words, the model suggests the process will work, even if the prototypes are in different locations for the two actors, so long as they are similarly ordered on the dimensions that define the category space. This seems a reasonable requirement that individuals share some common structure in their representations that in turn may subsume a common coarse structure of experience.

after 20 trials coordination succeeds in about 90% of the cases (10% of errors), and after 80 trials the success rate reaches approximately 95%.

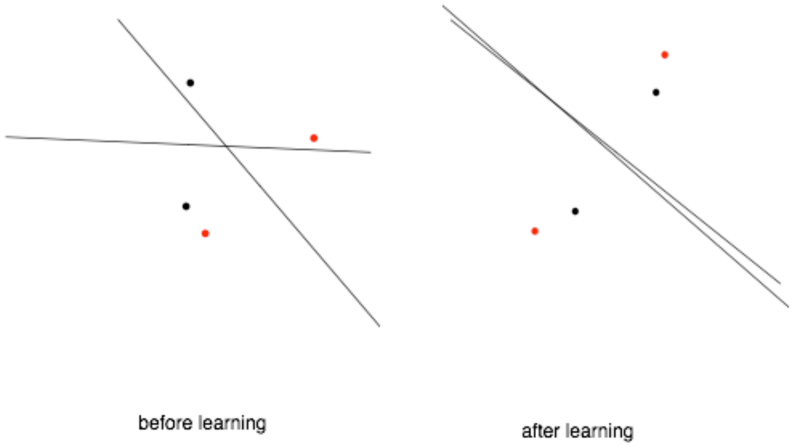


Figure S.4: Co-adaptation of two actors' conceptual spaces.

Supplementary Model S2:

Modeling acts and actions in ‘transform-the-target’

As a second illustration of modeling within the framework we consider a two-player experimental task, the ‘Transform-The-Target’ card game which has been used to establish key features of routinized activity in a controlled laboratory setting. Multiple experiments have shown that dyadic, habit-based action patterns develop within the first few deals, after just a couple minutes of play (Cohen and Bacdayan 1994; Egidi and Narduzzo 1997; Wang and Zhang 2008).

This experiment has proven to be a useful context in which to study the emergence of recurring interaction patterns among individuals. The emergent behaviors have many of the properties associated with skillful collective performance, including the rapid execution of interdependent behavior and “suboptimality”, the property that while patterns emerge to achieve the desired goal, they do not necessarily provide the most efficient solution path to the problem the actors face.

Background Task Information

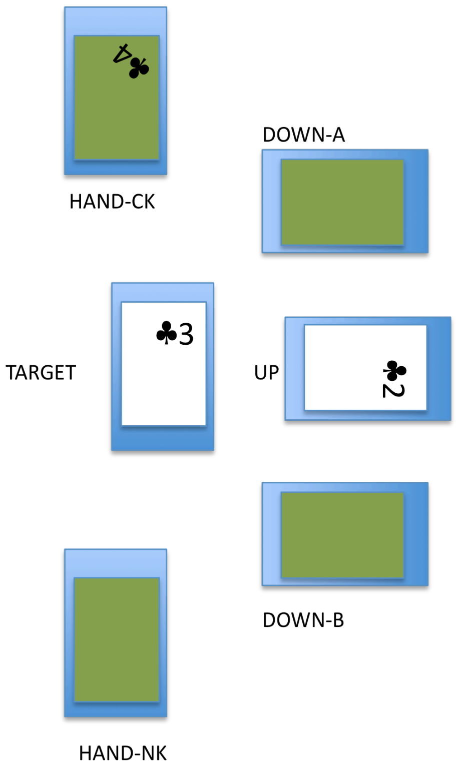
We begin with a presentation of the task details for those not familiar with Transform-The-Target (TTT). For the two-person version, the game is played with six cards, the 2, 3, and 4 of hearts and of spades[5]. The board, as shown in Figure

[5] Versions for more than two players have also been developed (Wang and Zhang 2008; Wollersheim 2009).

S2.1, has four positions while each of the two players holds one card in the positions labeled HAND-CK, at the top, and HAND-NK, at the bottom.[6] The key board position is the TARGET. The aim of play is to maneuver the red 2 into the TARGET area. Among the other three board positions, two are occupied by cards that are placed face-down (the DOWN-A and DOWN-B positions), and the third holds a card that is face-up (the UP position), and therefore visible to both players. Play proceeds in alternating turns, beginning with the player labeled 'ColorKeeper'. A turn allows a player to exchange his or her HAND card with one of the board cards, or to "pass". When a series of exchanges with the board successfully moves the red 2 into the TARGET area, the play for that deal of the cards ends. No verbal communication is allowed. In a typical experiment, a dollar might be won for completing the deal. The number of moves by both players would be counted, and each move might cost $.10. After 40 deals of the cards are played, each deal providing a different starting configuration, the net earnings would be divided between the two players.

What makes the game challenging is one further rule. The players do not have symmetric capabilities. One player, the one designated "ColorKeeper", can exchange with the TARGET only when that player's hand card, HAND-CK and the TARGET card are of the same color. The other player, designated "NumberKeeper", can exchange with the TARGET only when the HAND-NK and TARGET cards are of the same number.

[6] Board positions are labeled with capital letters.



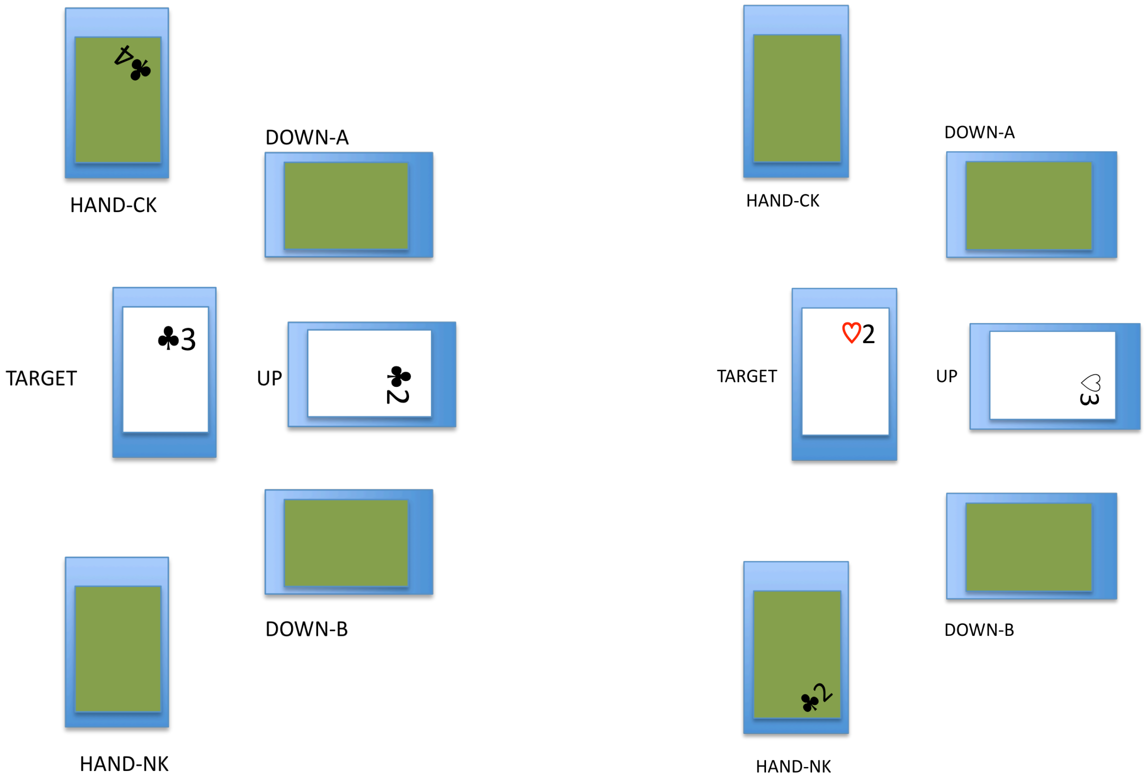


Figure S2.1. Left panel (a) showing cards visible to ColorKeeper at start of deal. Right panel (b) showing configuration of cards after NumberKeeper's finishing move.

In TTT there are 720 permutations of the six playing cards, and there are two players who might have the next turn, so the game can be in 1440 states. Of these, 240 correspond to the goal, that is, the red 2 is in the TARGET. This state space is small compared to real world problems, but big enough to keep novice players quite challenged for the duration of a 40-minute experimental session. The successive deals the players encounter, which come in a predetermined sequence, present a number of difficulties. Some are quite easy, and can be solved in just 2 or 3 smart moves. Typical hands require 4 or 5 moves. Some are quite difficult, and even good players may need 6 or 7 moves. It is not unusual for beginners to slowly take 10 or 15 steps to complete a deal that players with well-developed action patterns can do quickly in 4 or 5.

ow Nmb r	o Mov e	H ndCK	D ownA	P	D ownB	H ndNK	T ARGET	M ove Made
	K	4 ♣	3 ♥	♣	2 ♥	4 ♥	3 ♣	D ownB
	K	2 ♥	3 ♥	♣	4 ♣	4 ♥	3 ♣	D ownA
	K	2 ♥	4 ♥	♣	4 ♣	3 ♥	3 ♣	U p
	K	2 ♣	4 ♥	♥	4 ♣	3 ♥	3 ♣	U p
	K	2 ♣	4 ♥	♥	4 ♣	2 ♥	3 ♣	T arget
		3	4		4	2	2	T

	K	♣	♥	♥	♣	$\underline{♥}$	$\underline{♣}$	arget
	one	3 ♣	4 ♥	♥	4 ♣	2 ♣	2 ♥	

Table S2.1: Successive states of play in example hand of TTT. Columns indicate the player to move next, the cards occupying each of the six positions, and the move made by the player to produce the state in the following row. Underlines indicate the cards that are exchanged in producing the state in the following row.

It may help to build intuition for the patterned action we are modeling if we "walk through" play of a typical deal as listed in Table S2.1. Figure S2.1 and the first line of the Table show the positions of cards just after our example hand has been dealt. The player in the "ColorKeeper" (CK) role has the first move. She sees that the red 2 is neither in her hand, HAND-CK, which holds the black 4, nor in the TARGET or UP positions of the board. (The two DOWN cards, -A and -B, and the NumberKeeper's HAND-NK card cannot be seen by CK.) CK searches for the red 2 by exchanging her HAND-CK_card with the card at board position DOWN-B. In our example CK is fortunate in her search and now holds the red 2, and it is NK's turn. We are at row 2 of the Table. He exchanges his HAND-NK with DOWN-A (underlined). He is looking for the red 2 as well, and, since he can't see DOWN cards or HAND-CK, he doesn't know CK has found the red 2.

CK cannot finish at row 3 by putting the red 2 into the TARGET since that would change the card color in the TARGET, violating the definition of the ColorKeeper role. (For that matter, NK couldn't put the red 2 in if he held it in HAND-NK either. He needs the TARGET to contain the black 2 in order for that exchange to keep the TARGET number unchanged, as required by his role.) CK's next act is exchanging her hand card with the UP card on the board. (Yielding row

4.) After that, experienced players would typically be quick to complete the hand. NK also exchanges with UP. CK puts the black 2 in the TARGET. NK puts in the red 2, and the hand is finished after six moves. (Table S2.1, row 7, and Figure S2.1b.)

This is not actually the ideal solution for this particular deal of the cards. For example, if CK exchanges with UP on the first move, there is a good chance of finishing in four steps instead of six. But pairs of players develop their action patterns over a number of different deals they face in their early experience, and then apply them to get good results in a new situation, sometimes without noticing that still better options existed. Such occasions of smooth-though-sometimes-suboptimal action are a hallmark of habit-based action patterns. The success of TTT as an experimental instrument has been its ability to evoke just such patterns from subjects.

Presentation of the Model

We show that play of the game can be represented as a computational model embodying many of the premises of section 2.1 of our main text. The ultimate goal of the game is to place a specific card (the 2 of hearts) in a specific location on the board (the TARGET) subject to constraints of information (hidden cards) and action (specialized player roles). The constraints on action also preclude any explicit communication among the players. While placing the 2 of hearts in the TARGET location is the ultimate goal, players generally develop sub-goals or intermediate ends-in-view, such as placing a card in a jointly visible area that enables action on the part of the other player. The most basic actions (premise I) of the TTT game are each player's exchanges of the card in their hand with four possible board positions, or "passing" their turn.[7] Each player therefore has five fundamental *actions*. The game aligns with the most fundamental aspect of the paper's framework: each action is a *function* that transforms the game-play *situation* and is applicable only in certain conditions (C). For an action to be available to a player, it has to be the player's turn, and, in the case of exchanges with the TARGET, the cards in that position and in the player's HAND have to jointly conform to the player's role restriction. Each action (except pass), changes three aspects of the TTT world: a board position gets a new card; the player's HAND position gets the card from the board position; and it becomes the other player's turn to move. Each action is potentially applicable in a large set of board situations, since the remaining cards can be in many different permutations among the unaffected board positions.

Each action might be invoked in the pursuit of several different near-term objectives, or *activated-ends* (F). For example, exchanging with one of the DOWN cards might be done in an effort to locate the red 2.[8] But at other times, a player might exchange for a DOWN card as part of seeking a card needed to prepare the TARGET for a subsequent finishing move.

The program fragment in Figure S2.2 shows how these components of an action can be incorporated into a simulation model.[9] The function definition is shown with 3 inputs: the current state of game play, *situation_* ; a predicate

[7] This model does not incorporate the manner of performing each action, although human players do move physical playing cards, or drag card-images on computer screens.

[8] The game is played with six cards: the 2, 3, 4 of hearts and spades. Therefore, it is sufficient to distinguish the suits by their color and the computational model refers to the cards as r2, r3, r4, b2, b3, b4.

[9] The language used is Mathematica (v6). Many other languages would be possible, of course. Mathematica has been used in view of its strength in combining functional and declarative (rule-based) programming styles. It is fundamental to our framework to represent action as a function and Mathematica provides a medium in which this approach is easily expressed.

(returning True or False) that defines a desired condition for the future, *endInViewQ_* ; and a possible board transformation, *act_* . The function operates on the board if the state of the board is consistent with the action's conditions, and the expected result of acting (E, L) would achieve the condition desired, the activated-end. For brevity, the ancillary functions *expectations*, *transform*, and *adapt* are not shown.

```
…

In[4]:=
action[situation_, activatedEndQ_, act_] :=
       Module[{s = situation, eQ = activatedEndQ, a = act},
          If[eQ[expectation[s, a]],
             {transform[s, a], adapted[eQ, s, a, True]},
             {s, adapted[eQ, s, a, False]}
      ]
 ];

In[5]:=
{resultSituation, resultActivatedEnd} =
        action[{r2, r3, b2, b4, r4, b3, ck}, ckHandB2Q, exchangeUp]

Out[1]=
   {{"b2", "r3", "r2", "b4", "r4", "b3", "nk"}, {ckHandB2Q, "attained"}}
```

Figure S2.2: Mathematica code for a function implementing *action* for the game TTT.

Also shown in Figure S2.2 is an input line, In[5], invoking the function in a particular *situation_* that is shown as a sequence of cards and an indication of which player has the next move that occupy the ordered positions {HAND-CK, DOWN-A, UP, DOWN-B, HAND-NK, TARGET, next-mover}. The *activatedEndQ_* invoked in this case is ckHandB2Q, a predicate that tests "Is the HAND-CK card the black 2?" The *act_* under consideration is exchangeUp, an exchange of HAND-CK with the UP board position. The action will return two results. The transformation is activated in this case since the *expectations* associated with the situation and action are consistent with the activated-end. Therefore, the function returns a list of two items, shown as Out[1]: (1) the *resultSituation,* which reflects the *transform* of the board positions by exchanging the cards in the UP and HAND-CK positions and advancing the next-mover, and (2) a list indicating the *activatedEnd* has been attained as the result of the function *adapted* which has been called with a parameter of True that indicates an act occurred. Thus, the action function has produced an act (G). If the transformation were not activated, the action function would have produced no act, and would have returned two different results: (1) the unchanged board situation with unchanged next-mover, and (2) a list indicating a possible revision of the *activatedEnd* resulting from the function *adapted* called with a parameter of False.

The action uses the function *adapted* to check whether its results have met its expectations, whether it has produced the desired result, or whether an inability to act signals a need for changed objectives (F, K). It uses these indicators of its performance to modify activated-ends using an updating process that is not shown here in detail. This updating allows the activated-ends to be strengthened or generalized if the action achieves expectations and activated-ends, and to be repaired if there has been a *breakdown*.(B,K). [10]

Our discussion of this computational model of TTT has not exercised the model in full play conditions. The complexity of that exposition is beyond the scope of this paper. Instead, we have described only a key portion of the model in order to demonstrate how natural it is to embody the premises of our framework

[10] In the conditions of this example, the first order expectations of these lowest level actions will always be met. The program is not allowed ever to mistakenly alter the board into an impossible configuration or lose track of which player will move next. For higher order actions, however, it is possible that expectations will not be met, mainly because of imperfect generalizations from prior experience.

in a small computer program and to represent recurring action patterns as functions.

Supplementary Model S.3:

Modeling Patterns of Participation

A third illustration relies on an entirely different modeling strategy, demonstrating how our approach can be translated into models of dynamic networks of agents.

In this case, we want to model the activation of patterns of coordinated action that are “distributed” among a group of agents. This can be considered as an elementary model of the retrieval of “routinized” behavior, in which previously learned action patterns are activated by stimuli coming from the environment, from within the group itself, or both. Models of this type illustrate some of the mechanisms that assure coherence in the collective performance of distributed systems of agents [premise L]. Furthermore, they offer insights into how the repetition of action patterns can alter the perception of the environment in terms of categories, thus extending (or contracting), in turn, the domain of application of actions themselves [premises C and K].

We consider a particularly simple system in which the central problem is for subgroups of agents to become active in circumstances where their capabilities are complementary and to avoid co-activation when their capabilities are mutually interfering or incompatible. Hence, the collective performance in this case will be patterns of co-activation of the agents varying in response to differences in environmental or contextual factors. In our model, this is represented as the alignment of binary states of the agents, which clearly captures some basic aspects of organizing (Levinthal and Warglien 1999; Milgrom and Roberts 1990; Simon 1962; Thompson 1967). The model can be considered an example of “aggregation” (Axelrod and Bennett 1993): agents have to decide how to partition themselves into mutually compatible subgroups that are also suited to environmental circumstances. For example, rescuers may have to decide how to divide into two groups to face a complex emergency, or a group of workers may have to decide who will take the day shift and who will work the nights.

To make it simple, we assume that agents can take only two effective states, which we label 1 and -1, thus determining two subgroups (the agents in state 1 and the agents in state -1). However, agents can also be temporarily in an “undecided” state, labeled 0. The system models each agent as having links to many of the others. These “connections” represent complementarities and trade offs between individual agents. These may be due to multiple factors: complementarities of skills or locations, perhaps even relationships of empathy or animosity. Altogether, these connections summarize the propensity of a pair of agents to activate themselves jointly (positive connections) or separately (negative connections). For example, agents with complementary abilities will, ideally, have positive propensity to joint activation (taking the same state). There is no reason to assume that there is only one kind of complementarity/conflict relationship. A pair of individuals may have some complementary skills, but at the same time hold other incompatible abilities or feelings. Furthermore, different tasks or different environments may put different demands on such skill sets resulting in the possibility of multiple patterns of activation among the same set of actors. We assume that agents link a set of situations to a pattern of activation associated with it. These associations thus implicitly define categories of situations linked to specific activation patterns. In general, such associations are expected to be learned (e.g. by associative or Hebbian learning (Hertz et al. 1991)), but for sake of simplicity we will directly specify them into our model. We will also assume that such connections are symmetric (each pair of agents have the same perception of their complementarities/tradeoffs) or, more reasonably, they show randomly distributed deviations from the symmetric case for any pair of agents, (with no loss of qualitative results (Hertz et al. 1991)). In this model, agents rely on observing others' behavior to modify their own and thereby coordinate with others, rather than exploiting explicit communication – a tacit process resembling soccer players in the field or drivers in traffic rather than participants in a meeting.[11] Every agent looks at others' states and aggregates that information. We assume that agents attempt to improve their overall level of fit (or reduce their level of frustration) with other agents. The resulting behavior is very similar to what you observe at dinnertime during conferences: everybody looks for a table where the sum of interesting/pleasant persons maximally exceeds the sum of undesirable ones (with each diner, of course, having her own weighting of the desirability of the others). In our model, agents proceed in much the same way. If an agent sees another agent, with whom she has positive connections, having her same state, she will compute a positive input; if she sees another agent, with whom she has negative connections, having her same state, she will compute a negative input; and vice versa when agents' states are discordant. The agent will adopt a value of its state corresponding to the sign of the "weighted majority" of the other agents' states.[12] Thus, the network of agents is subject to shifting of state variables in response to shifts in the state variables of the agents to which

[11] See Hutchins 1995, Warglien and Marchiori 2005, Gavetti and Warglien 2007 for ways to model explicit communication in this type of network models.

[12] More formally, each agent i will update its state according to:

$$x_i = \mathrm{sgn}(\sum_{j \neq i} w_{ij} x_j)$$

where xi is the state of an agent i. w is the matrix of the connections, and represents the sum of all underlying layers, and wij is the value of a connection between agents i and j.

one is connected. It can be shown that such networks will converge by subsequent adaptations to a stable point in which no agent further modifies its state (Hopfield 1981; Hertz et al. 1991; Axelrod and Bennet 1993). Common images to represent such a process of convergence are those of a ball rolling down a surface until it finds the bottom of some basin in the surface – or symmetrically one of a hill climbing path, where going up leads inevitably to a (local) peak. In general, there can be many such stable points, representing different stable patterns of behavior. This enables the model to represent multiple action patterns within a single structure. Indeed, an arbitrary initial state of the environment will trigger an adaptive process through which agents will rapidly converge towards one of the stable patterns implied by the connection structure of the agents' network. Thus, any state of the environment will be implicitly categorized – and a corresponding pattern of activation will be triggered. Clearly, the process of categorization and activation is distributed among agents (there is no central coordinator), and the network of connections acts as sort of collective memory.

We are now ready to further explore this basic model. In what follows, we show how a group of agents can smoothly respond to variable environments, and can categorize the environment and evoke an appropriate activation pattern even in conditions in which inputs are incomplete, or some agents have been removed from the group, thereby demonstrating the robustness of the patterns of action and the corresponding categorization of the situation.

Subsequently, we show that, under simple assumptions on how the experience of action payoffs are taken into account by agents, successful patterns of action may tend to annex new states of the world to their domain of application even without previously experiencing such states (a kind of "categorical imperialism” by successful patterns of action). Finally, we show how categories may disappear (or appear) as the result of such processes, and how discontinuous change in behavior may thereby arise out of an underlying continuous change in the structure of the network.

We introduce a simple role structure in the group of agents: some agents have interface roles, interacting directly with the environment, while others are working “inside” the organization (e.g. “back-office” workers). In particular, we model a group of seven agents, four of them interacting directly with the environment (shown in Figure S.3.1 as shaded circles), while three are on the “inside” (unshaded). Continuous lines in the figure represent positive relationships, while dashed lines represent negative connections.

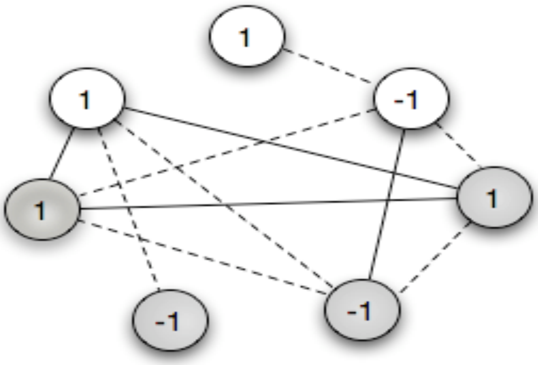


Figure S.3.1: Network of agents in a stable configuration, about here.

The environment can take different configurations, each of them resulting from the combination of four binary (e.g. present/absent, or low/high) features. For example, if a feature is “high” it will take a ↑ value, if it is “low” it will take a ↓ value. A full environmental context, that we conventionally mark by angle brackets, can be specified by an ordered list, such as <↑, ↓, ↑, ↑>. Thus, there are 2^4=16 possible environmental configurations. Interface agents are feature detectors: each one of them detects a feature of the environment and reflects it in its initial state (e.g. if a feature has a ↑ state it takes a 1 value, if the feature is ↓ it takes a -1 value).

Agents' connections associate patterns of activation with specific sets of environmental contexts. While many pattern of activation can be stable configurations of a given network of connections, for simplicity we will consider a case in which only two activation patterns are stable. We will call the activation patterns α and β and use A and B for the corresponding environment categories (the set of states of the environment associated respectively with α and β). We first show that the patterns of behavior and the associated categorization of the environment encoded in the agents' weighted network are robust and can respond to variation in inputs, incomplete information, and even the deletion of group members. These are all features that we associate with the flexibility and smoothness of collective performance.

The connections shown in Figure S.3.1 guarantee that the agents will respond to all possible states of the environment by reproducing one of the network's two stable patterns of action, {1, 1,-1, 1,-1,-1, 1} and {1,-1,-1, 1,-1, 1, 1}.[13] This means that all states of the environment will be categorized accordingly. If, for example, the state of the environment < ↓, ↓, ↑ ↑> , which we have pre-defined for our illustration, is presented to the network, it will evoke a pattern of action {1, 1,-1, 1, -1, -1, 1}, the one shown in Figure 7. The same network pattern will occur with the initial state of the environment <↑, ↑, ↓, ↑>. Thus the two states of the environment belong to the same category A. Conversely, state

[13] Braces denote patterns of action and bold characters the state of agents that are interfaces with the environment. Interface agents may change their values as the organization adjusts to the environment.

<↑, ↓, ↑, ↓> will be associated with the pattern {1,-1,-1, 1,-1,1,1} and will belong to a different category, B.

Not only can this model of collective performance respond to all possible inputs by smoothly converging to one of its stored patterns of activation, it can also robustly survive the incomplete presentation of environmental states. Table S.3.1 reports the robustness of collective performance as a function of the number of environmental features missing from the input received by interface agents. For a given network and each possible input, we compare the implicit classification of the environmental state reached when all environmental features are present with the one achieved when a number of features are randomly set to 0. Robustness is measured as the fraction of instances in which the same implicit classification is achieved despite missing information. As in Figure S.3.1, we use a network of seven units (four of which were interface agents); in this variant of the model, three stored patterns are designated to represent prior learning (and equally spanning the input space). The simulation was repeated 100 times for each input, and each time the missing features were randomly determined.

(a) # of missing features	(b) Robustness
1	0.86
2	0.81
3	0.77

Table S.31: Robustness of collective performance with varying features.

An even more remarkable form of robustness is that the patterns of activation will be preserved even if members of the seven-agent group are deleted at random. Table S.3.2 reports the robustness of collective performance as a function of the number of deleted agents. In this case, robustness can be measured using an indicator of successful partial reconstructions: surviving agents can only reconstruct the parts of the original stable pattern that do not involve the deleted agents. Since as fewer agents are considered the chances of random success increase, the third column in the table shows the difference between the actual success rate and the chance success rate (measuring the "marginal contribution" of the surviving connection structure).

(a) Number of agents deleted	(b) Relative Frequency of successful partial reconstructions	(c) (b) – random chance of successful partial reconstruction
1 agent	0.96	0.93
2 agents	0.96	0.90
3 agents	0.94	0.88
4 agents	0.78	0.53

Table S.3.2: Robustness of collective performance with varying individuals

Interesting modeling issues arise by allowing some dynamics in the relative weight of activation patterns. A given network of connections associated with multiple activation patterns can be thought of as the aggregation of multiple layers of connections, each one associated with a single pattern. Dynamically this can be conceived as resulting from the fact that each time some nodes of the network are simultaneously active in a given situation, they reinforce connections among them. Statically, one can engineer the network by summing networks associated with each activation pattern. Thus, by experience or by design, each pattern of activation carries its own "marginal contribution" to the overall network of connections.

Until now, we have assumed that patterns were equally weighted in determining the value of the connections between agents, and that, in addition, these weights were constant. However, relaxing this assumption may be useful. For example, Gilboa and Schmeidler (2001) have suggested that experienced payoffs may change the relative weight of cases in individual memories, thus altering over time the behavior of decision makers. We can follow their suggestion in order to explore how the expression of patterns of activation may affect the way the environment is subsequently categorized – altering in turn the domain in which particular activations will be expressed.

Suppose, for example, that each time agents carry out a pattern of activation associated with a given situation they modify the relative weight of the pattern in proportion to the experienced payoff. There is no need at this stage to specify which specific reinforcement process is at work; it suffices that higher average payoffs imply an increasing relative weight for a given pattern of activation. For simplicity we consider a setting in which there are two modal environment configurations, the most frequent ones. This allows us to ignore details of how small variations in the environmental configurations are experienced. In the same vein, we assume – as shown by simulations not reported here – that other configurations have frequency low enough that they do not significantly affect, the dynamics of the network.[14]

We consider the situation with the following properties:

[14] Of course, this is an extreme assumption, but we make it only to simplify the logic of our exposition. Analogous results can be obtained in less extreme settings that relax this assumption, as well as with different reinforcement processes.

- the first network pattern (say α), when evoked by its modal state of the environment A* (belonging to category A) , gets a stronger reward than the second pattern (say β) when the latter is expressed in response to its own modal state of the environment B* (belonging to category B)

- both payoffs are positive

- whenever α is expressed in the environment configuration B*, the payoff is lower than the one generated by action β, though still positive. Symmetrically, β is less good than α when expressed in A*

- each prototypical state occurs with approximately equal frequency over time.

As a result of the reinforcement process, the relative weight of the (A, α) case will increase over time. The relative weight of (B, β) will conversely diminish. The dynamics of the network of agents can be usefully portrayed with the help of the now familiar representation of a "landscape" curve. Figure S.1.2 captures the evolution of three important features of the agents' network behavior as the relative weights of the first case, λ, and of the second one, $(1-\lambda)$, are modified. The curves show two basins of attraction corresponding to the categories A and B (the portions of the diagram leading a rolling ball respectively to α or β). These represent the number of configurations of the environment categorized by each pattern. As λ grows, there are transitions in the width of such basins. In other words, at some critical values of λ, the category associated with pattern α annexes new states of the environment (those formerly associated with β). This implies that as it succeeds, and thus raises its relative weight, the pattern of action α extends its domain of application to new environmental configurations. It is remarkable that some configurations of the environment can shift their categorical attribution without ever having been experienced – as a mere outcome of the "categorical imperialism" of successful patterns of activation. In effect, they change attribution because they have some similarity to other environments that were experienced as rewarding.

The depth of the curves represents a measure of the "frustration" or "energy" of the system.[15] The more propensities of agents are satisfied, the lower the frustration level will be. As can be seen from the Figure, as λ grows, the energy of the α pattern decreases while that of β increases. Changes in energy are continuous and a linear function of λ. Finally, the value of the curve at the top between the two valleys offers a graphical representation of the "energy wall" separating the two activation patterns. The higher the wall, the more a pattern is, so to speak, protected from the attraction exerted by the other. But, as the wall disappears, as for λ=.8 in the fourth panel of Fig.S.3.2., the pattern β with highest energy loses any attractive force and its whole basin of attraction landslides towards α. The outcome is the disappearance of a stable pattern of action, missing any domain (or category) of environments supporting its activation.

[15] A conventional measure of "frustration" or "energy" in networks such as the one we model here is:

$$H = -\frac{1}{2}\sum_{ij} w_{ij} x_i x_j$$

, where x(i) and w(i,j) are interpreted as above. It is easy to see that H will diminish (increase) each time that agents that have positive (negative) "propensity" connections are in the same group – and vice versa for the case in which they are in different states. See Axelrod and Bennet (1993) for a slightly different measure with similar properties.

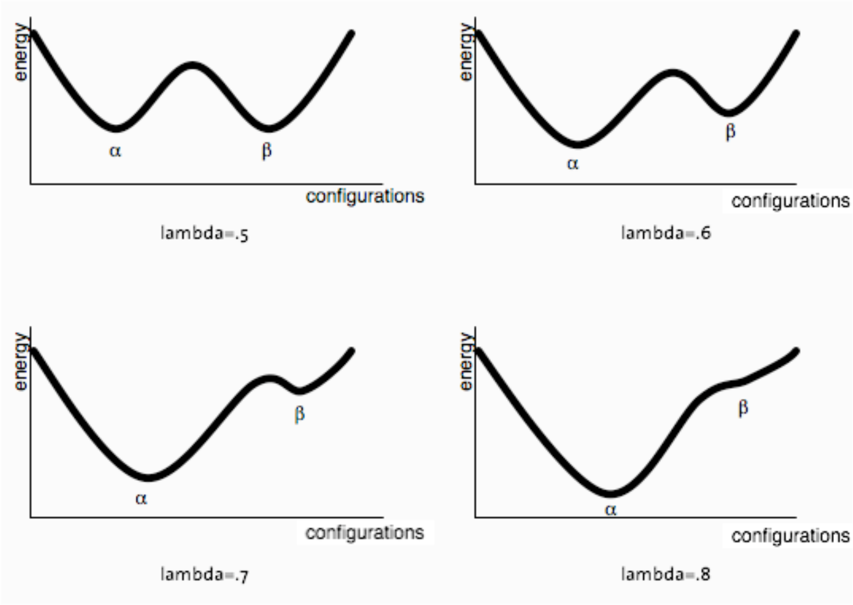


Figure S.3.2. Frustration levels for two activation patterns at varying levels of λ (lambda), the relative weight of pattern α

Works CIted

Axelrod, Robert., and D.S. Bennett

1993 “A Landscape Theory of Aggregation”. British Journal of Political Science 23: 211-233.

Barsalou Lawrence W.

1987 “The instability of graded structures: implicaitons for the nature of concepts”. In U. Neisser (ed.) Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization. Cambridge: Cambridge University Press.

Cohen, M. D., and P. Bacdayan

1994 “Organizational Routines Are Stored as Procedural Memory: Evidence From a Laboratory Study.” Organization Science 5:554-568.

Egidi, Massimo, and Alessandro Narduzzo

1997 “The emergence of path-dependent behaviors in cooperative contexts.” International Journal of Industrial Organization 15:677-709.

Gärdenfors, Peter

2000 Conceptual Spaces: The Geometry of Thought. Cambridge, MA.:The MIT Press.

Gavetti, Giovanni, and M. Warglien

2007 "Recognizing the New: A Multi-Agent Model of Analogy in Strategic Decision-Making" Strategy Unit Working papers, Harvard Business School, Cambridge, MA.

Gilboa, Isaac, and D. Schmeidler

200. A Theory of Case-Based Decisions. Cambridge: Cambridge University Press.

Hertz, John, A. Krogh, and R. Palmer

1991 Introduction to the theory of neural computation. Reading, MA: Addison Wesley.

Hopfield, John J.

1982 "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the National Academy of Sciences, 79: 2554–2558.

Levinthal, Daniel, and M. Warglien

1999 "Landscape Design: Designing for Local Action in Complex Worlds." Organization Science 10: 342-357.

March, James G., and J.P. Olsen

1989 Rediscovering Institutions: The Organizational Basis of Politics. New York : Free Press/Macmillan.

Marchiori, Davide, and M. Warglien

2005 Constructing shared interpretations in a team of intelligent agents: The effects of communication intensity and structure. In T. Terano, H. Kita, T. Kaneda, K. Arai, H. Deguchi (eds), Agent-Based Simulations: From Modeling Methodologies to Real-World Applications: 58-71. Tokyo: Springer Verlag.

Milgrom, Paul, and J. Roberts

1990. “The economics of modern manufacturing.” American Economic Review 80: 511-528.

Nosofsky, Robert M.

1988 “Similarity, frequency, and category representations.” Journal of Experimental Psychology: Learning, Memory and Cognition 14 54-65.

Okabe, Atsuyuki, B. Boots, and K. Sugihara

1992 Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. New York : John Wiley & Sons.

Rosch, Eleanor, C.B. Mervis

1975 “Family resemblances: Studies in the internal structure of categories.” Cognitive Psychology 7(4): 573-605.

Shepard, Roger N.

1987 “Toward a universal law of generalization for psychological science.” Science 237: 1317-1323.

Simon, Herbert A.

1962 “The Architecture of Complexity.” Proceedings of the American Philosophical Society 106: 467-482.

Thompson, James D.

1967 Organizations in Action. New York : McGraw-Hill.

Wang, Jian-An, and Gang Zhang

2008 “Knowledge, Routines and Performance in Collective Problem Solving.” Acta Psychologia Sinica 40:862-872.

Wollersheim, Jutta

2009 “An Empirical Investigation of the Role of Dynamic Capabilities in Mergers and Acquisitions.” Rotterdam, Netherlands.