EconomEtica

Centro interuniversitario per l'etica economica
e la responsabilità sociale di impresa
promosso dalla Fondazione Italiana Accenture

N.71 August 2019

Institutions, Frames, and Social Contract Reasoning

Virginia Cecchini Manara,
Lorenzo Sacconi

Institutions, Frames, and Social Contract Reasoning

Virginia Cecchini Manara
v.cecchini@unitn.it

Lorenzo Sacconi
lorenzo.sacconi@unimi.it

Preliminary Draft

Abstract

This work aims at filling a gap in the cognitive representation of institutions, starting from Aoki's account of institutions as equilibria in a game-theoretical framework. We propose a formal model to explain what happens when different players hold different representations of the game they are playing. In particular, we assume that agents do not know all the feasible strategies they can play, because they have bounded rationality; grounding on the works by Johnson-Laird and his coauthors, we suggest that individuals use parsimonious mental models that make as little as possible explicit to represent the game they are playing, because of their limited capacity of working memory and attention. Second, we rely on Bacharach's variable frame theory: agents transform the objective game into a framed game, where strategies are "labeled" in some sense. In such a context, we argue that a social contract – given its prescriptive and universalizable meaning – may provide a shared mental model, accepted by all players, that allows agents to select a joint plan of action corresponding to an efficient and fair distribution.

JEL Classification: B52, C7, D02, D83
Keywords: institutions, shared beliefs, mental models, framing, social contract

1 Aoki and the missing point

The starting point of this work is the conceptualization of *institutions* given by Masahiko Aoki: in his book on comparative institutional analysis (2001) and related works (2010, 2011), he studies institutions through the lens of game theory, adding interdisciplinary contributions, in particular using comparative and historical information. The novelty of his work is the approach to institutions as *equilibria* rather than players or rules of the game. He defines an institution as "a self-sustaining system of shared beliefs about a salient way in which the game is repeatedly played". As such, an institution is "the product of long term experiences of a society of boundedly rational and retrospective individuals" (Kreps, 1990, p. 183).

The concept of **salience** has been widely used since its introduction by Schelling (1960), who provides an explanation of salience and focal points that relies on the idea of framing, through labeling and pattern recognition: it is a key concept in the theory of common knowledge developed by Lewis (1969); and it is underlying in the work by many other scholars, for instance in Sugden's (1995) analysis of focal points as the result of labeling functions or in Bacharach's (1993) *variable frame theory*, where an outcome is salient if it has a particular uniqueness under a particular frame. However, since labels and frames are exogenous in these frameworks, it remains quite difficult to understand where salience comes from.

The work of Aoki can offer an answer to the emergence of salience, accounting for the interplay of behavioral and cognitive dimensions both at the individual and societal level. Figure 1 describes this idea: the strategic choices made by individual agents on the basis of shared beliefs jointly construct the equilibrium state, which in turn reconfirms its summary representation. Thus the institution becomes self-sustaining and information compressed in it becomes taken for granted by the agents unless some events shaking the shared beliefs occur: "The content of the shared beliefs is a summary representation (compressed information) of an equilibrium of the game (out of the many that are theoretically possible). That is to say, a salient feature of an equilibrium may be tacitly recognized by the agents, or have corresponding symbolic representations outside the minds of agents and coordinate their beliefs" (Aoki, 2001, p. 10).

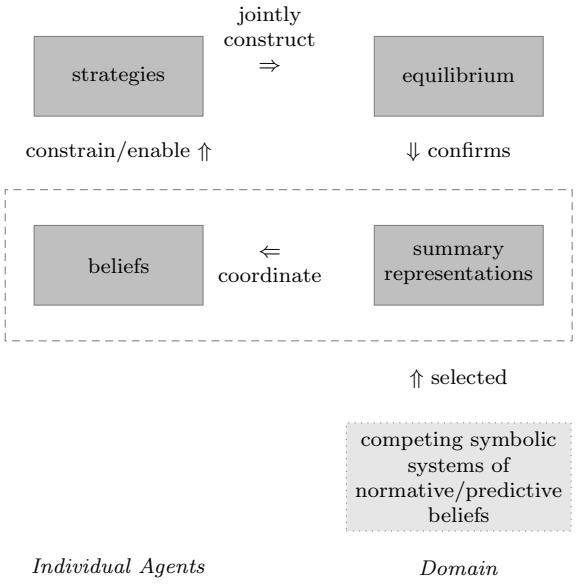

Figure 1: An institution as shared beliefs (Aoki 2001)

This view of institutions entails a dualistic nature: on one side they constrain individual choices by coordinating beliefs and therefore they drive their actions in one direction against all the others that are theoretically possible (i.e., other equilibria). On the other, an institution enables the bounded-rational agents to economize on the information processing needed for decision-making. Thus individual agents are not only constrained but also informed by institutions.

If we accept the view of institutions as equilibria, then we must admit with Aoki (2001) that explicit, codified and/or symbolic representations such as statutory laws, regulations and so on, cannot by themselves create a pattern of behavior: such representations are institutions only if the agents mutually believe in them. On the other hand, certain practices, if not formalized, can be institutions as long as the agents believe in them as relevant representations of the internal state of the domain; they cease to be institutions when the agents' beliefs in them are critically shaken. In his work, Aoki describes how bounded-rational, individual agents form their own *subjective models of the game* that they play, and discusses the mechanism of institutional change as a process of revision, refinement, and

inducement of mutual consistency of such models incorporating a (common) representation system. The goal of the present work is to give a formal description of this mechanism and to add to this picture the intuition that a social contract reasoning is able to give the starting point of the process.

In Aoki (2011) his definition of institution is reformulated in terms of "societal rules", that are "commonly cognized, salient patterns of the ways in which societal games are recursively played and expected to be played": they are endogenous outcomes of play of the societal games (see Figure 2). In this work, Aoki distinguishes the game-form (or exogenous, formal rules) and the societal rules (endogenous outcomes), and he adds that "the societal rules can be in Nash equilibrium, once they are taken for granted and as sure by all the agents, even without the complete game-form becoming their common knowledge". This is another aspect that we try to verify: we will show how it is possible to have an equilibrium even in absence of common knowledge of the game-form.

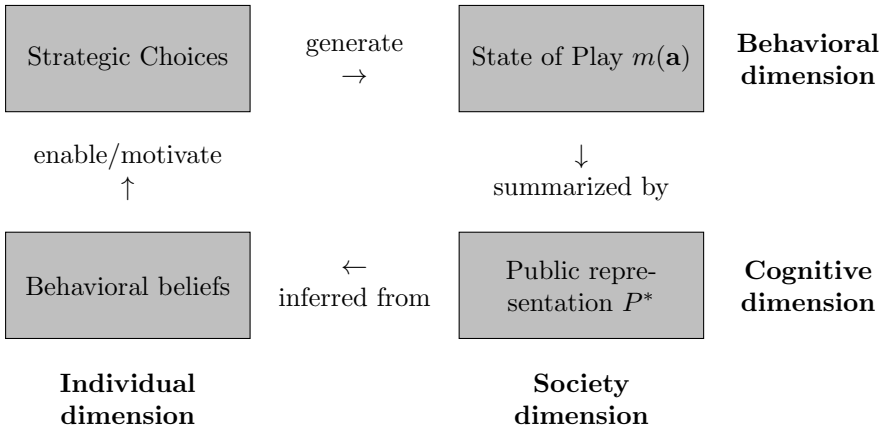

Figure 2: The mediating role of institutions on substantive form (Aoki, 2011)

But in particular what we are interested in is to understand where this "public representation" that mediates between the equilibrium play of a societal game and individual belief formation can come from: Aoki (2011) suggests that it should be an «external media» or artifact that linguistically represents salient features of equilibrium plays (such as norms, rules organizations of known types, laws).

However we argue that, notwithstanding Aoki's framework is the most complete and useful treatment of the concept of institutions, in order to analyze their emergence and stability, it still misses something: in accepting an evolutionary/conventional approach, his analysis suffers from circularity - it describes the world as it is, giving the explanation that it is like this because before it was like that, and so on. The big challenge, in our view, is to build a bridge between the description of how the world is and the prescription of how it should be.

The question then becomes: how do some strategies *become* salient? How does it happen that agents come to have certain beliefs? Our intuition is that rules and formal institutions can shape preferences and behaviors although the sole introduction of a new legal rule is not enough: therefore we are interested in studying the mechanism of transmission from formal rules to individual and collective representations that become actual beliefs and motivations to act. How do individual agents come to accept a specified pattern and follow it as their own cognitive frame? And how is it possible that different agents, with different knowledge and preferences, coordinate mutual beliefs?

We suggest that an answer to these questions can be given by incorporating into this framework the normative meaning of norms, with a particular attention to the Social Contract as a selection device.

In fact, we find one proposal in Sacconi's recent works (see for example Sacconi 2013), where a modified version of Aoki's account is presented (see Figure 3), introducing the social contract as the cognitive mechanism by which a norm may be accepted and become a shared mental model: "However, a limitation is apparent in this understanding of institutions, and it concerns the normative meaning of an institution. Institutions in the above game-theoretical definition only *ex post* tell each player what the best action is. Once the players share the knowledge that they have reached an equilibrium state, then playing their best replies is actually a prescription of prudence that confirms the already-established equilibrium. Thus, institutions tell players only how to maintain the existing, already settled, pattern of behavior. They say nothing *ex ante* about how agents should behave before the mental representation of an equilibrium has settled and a self-replicating equilibrium behavior has crystallized. Institutions only describe regularity of behavior and are devoid of genuine normative meaning and force"

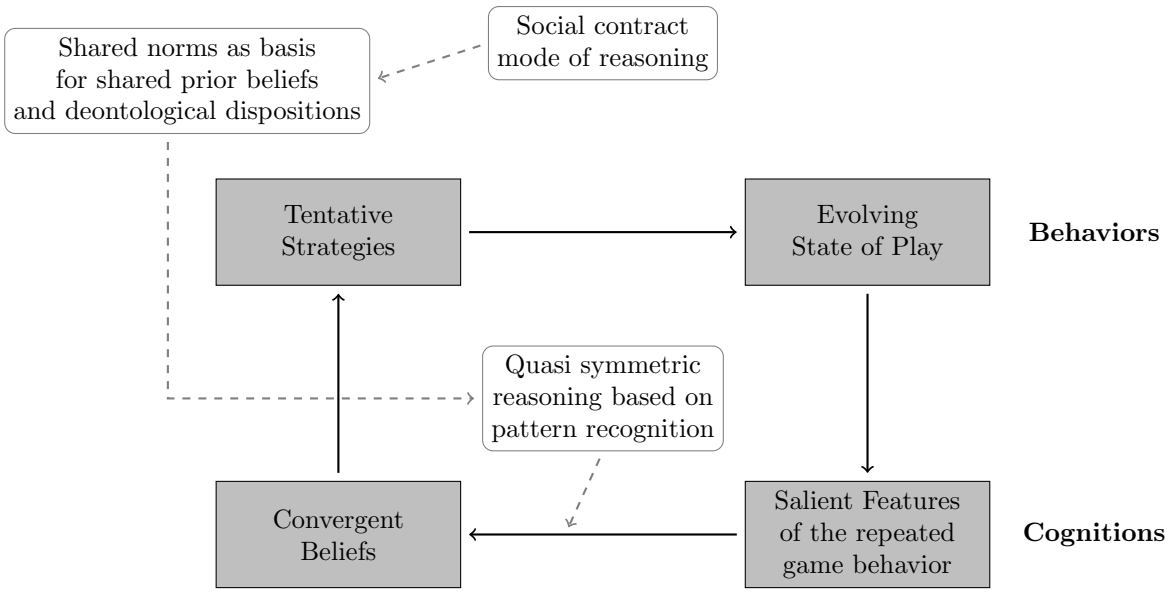

Figure 3: Aoki's modified diagram representing the recursive process of institution formation

(Sacconi, 2013, p. 121).

Binmore (2005) has shown how the social contract (a normative ethics principle) provides a source for the selection of an equilibrium in the *ex ante* problem, using the Rawlsian original position behind veil of ignorance.

The introduction of an *agreement*[1] among actors has in our view a double effect: on one side it is able to activate in their mind a peculiar way of reasoning that generates shared representations leading to a fair outcome, on the other side it enriches their motives to act with a new motivation, based on the sense of justice that, once developed, overcomes incentives to cheat and transforms fair behavior into each participant's best response to the other individuals' behaviors. Our work highlights the interplay of cognitive and motivational processes[2] and the fact that

[1] In the tradition of social contract theory: see Hobbes, 1651; Buchanan, 1975; Gauthier, 1986; Rawls, 1971.

[2] See for example Kruglanski et al., 2002.

"human beings are biologically adapted for participating in collaborative activities involving shared goals and socially coordinated action plans"[3], by recognizing that the (real or hypothetical) participation in an agreement is able to impact on both cognition and motivation.

The presence of a peculiar cognitive mechanism linked to social contract theory has been studied in cognitive psychology by scholars who have shown how the human mind is evolutionary shaped to reason about social contracts[4]. Some recent experimental tests have also shown how the participation in *ex ante* agreements has a strong impact on beliefs and motivations in the *ex post* behavior in games of cooperation[5].

We suggest that the normative social contract elicits a frame supporting the fair solution also *ex post*. In order to do this, an additional cognitive psychology assumption is needed: because players have cognitive limitations, they do not consider all the logical possibilities in the *ex post* game, they continue to conceive their interactions within the 'frame' in which they entered when assuming for normative reasons the perspective the original position. In particular, this frame assumes that they are equal and interchangeable and it delimits the information that an agent may consider as relevant (within the frame). Hence the only information to which the agent pays attention is the subset consistent with the frame itself.

2 Mental models in games

It is well known that individual agents have limits in attention and working memory; following a long tradition in psychology (starting from: Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991) we suggest that when they face a decision problem that involves other players and a wide space of results (possible combinations of individual acts), agents will read the situation through *parsimonious*

[3]Tomasello et al., 2005.

[4]Tooby and Cosmides, 1989.

[5]Sacconi and Faillo, 2010; Sacconi et al., 2011; Faillo et al., 2015.

mental models. In particular, we will propose that bounded rationality constraints agents to consider only small portions of big games[6]. The novelty of this approach lays in the fact that we suppose that the limits of rationality do not affect the players' ability to think strategically, that is to understand that the final outcome depends not only on their own choice or on some random move by Nature, but also by the strategic and (bounded) rational thinking of other agents.

Psychologists have studied for long time the mechanisms that allow agents to read situations through mental models, but mainly in decision problems, with scarce application to strategic contexts. On the other hand, economists have paid little attention to cognitive components of mental representations of problems, especially for those involving strategic interaction (games)[7]. Nonetheless, the behavioral and experimental literature on the nature of cognitive constraints that affect the players' mental representation of games is becoming larger, but we still lack a unified theory of mental models in games.

The main contributions have highlighted some factors that affect the representation of games[8]:

- a difficulty in managing complex, non-projective, structures of payoffs, where elements of competition and coordination are mixed, leads to simplified representations of payoffs;
- the presence of salient features may elicit the application of representations used in past situations (analogy, transfer, precedent, pattern recognition);
- the conspicuousness of some feature or explicit information affects the representations of the games through mechanisms of focussing or frame-effects;
- the description of available strategies through labeling or categorization.

[6] See Harstad and Selten (2013) and Crawford (2013) for a critical discussion of this hypothesis.

[7] But see Esponda and Pouzo (2014); Halevy et al. (2012).

[8] See for example Devetag and Warglien (2008).

Moreover, Bacharach suggested that agents do not choose on objective strategies, but on a set of act descriptions: his *variable frame theory* [9] starts with a conventional game representing an interaction between players as it "really is" or as the theorist describes it, called the objective game. This game is transformed into a *framed* game in which players do not face a decision problem with the set of objective strategies from which, objectively, they choose; but they choose from a set of act-descriptions. Act-descriptions are constructed from predicates, which belong to disjoint sets, called families.

In the absence of a proper theory that explains the formation of mental models in games, we advance some proposal for a comprehensive framework that takes into considerations the cognitive limits of agents, but without imposing that these limits affect their ability to think strategically.

The main characteristics of the mental model are the following: it is partial, in the sense that it represents only a partial interpretation of the real situation (it is a *small scale model* - Holyoak and Spellman, 1993); though partial, such representation is not arbitrary, since it preserves the structure of the original game, and finally it is parsimonious, namely it makes explicit as little as possible, because of limited working memory.

Agents form beliefs on others' behavior but they are aware that there is no common comprehension of the game, since everybody might have a different representation of it. An external signal might play an important role on the formation of individual mental models and frames: for example Legrenzi et al. (1993) have shown how people tend to *focus* on the information that is explicit in the description of a problem; however it has not been explained yet how this mechanism does not act only on the way individuals frame the game, but also on their beliefs and expectations about the others' beliefs and behavior: the inter-personal dimension of frames has not yet been investigated.

---

[9] See Gold and Sugden in Bacharach, 2006.

3 Subjective Game Models

Following Aoki's proposal, we identify a domain as a set of a finite number of agents (players) and the sets of all technologically feasible actions. In Aoki's version, this space has infinite dimension; in our framework we assume that it is *big*, in the sense that its cardinality far exceeds the computational/cognitive possibility of agents, but this does not necessarily imply that it is infinite. What we request is simply that the *feasible* space is bigger than the *conceivable* space.

We also assume that, in principle, each action is feasible for any agent, i.e. $S_i = S_j \forall i \neq j \in N$, so that the action sets of agents, S_i, are assumed to be identical, and the set of all technologically feasible action profiles S is symmetrical: $S = \times_{i \in N} S_i$. This assumption reflects the idea that all individuals are originally equal, and differences in possibility spaces are the result of culture or path dependent roles.

Consequences of action profiles are relevant to agents' welfare, and we further assume that all the agents are symmetric in having the identical payoff function. Each agent has a payoff (utility) function π_i defined on the consequence space and intends to maximize his payoffs from his action choices.

In particular, we consider a game defined by the *Objective Game Form*: $G^O = \{N, S, u\}$ with: $N = \{1, \ldots, n\}$ the set of agents, $S = \times_{i \in N} S_i$ the (nonempty and finite) set of feasible actions, $u : S \to \mathbb{R}$ a payoff function (outcomes are expressed in terms of monetary payoffs, we refer to them as π_i).

In order to build a model of the game in their mind, players represent in some way its structure, through a *partial mapping*: an agent reads it through his *Subjective Game Form*: $G^S = \{N^S, S^S, u^S\}$ that represents the same elements in the agent's mind.

Bounded rationality implies that the agent cannot see the whole structure of the game; nonetheless, his subjective model will not be completely arbitrary, but each agent's mind will be able to select a subset of the objective game form (a "consideration set"). The core of Johnson-Laird's theory is that in deductive reasoning we construct only "partial semantics" of a sentence. The notion of "partial mapping" that lays at the core of Johnson-Laird's theory provides cues to construct an analogous definition of mental models of games as *partial structures*,

as suggested by Devetag and Warglien (2008): in reasoning as well as decision making tasks, individuals typically focus on some information and neglect other. The problem is which elements and relations are preserved in the player's mental model.

A first possible simplification comes from considering a subset of the players involved, i.e. $N^S \subseteq N$. In our case we will consider two-persons games ($n = 2$), thus the only simplification that a player can do is to neglect the presence of the other player and choose as if the context was not strategic [10].

We will focus our attention on another kind of simplification: $S^S \subseteq S$, while we do not address the the possibility of a partial understanding of payoffs or individual utilities.

The choice behavior is twofold: first the agent reads the situation in a framed way, second he chooses his action among the one he sees [11].

A player's frame F^i is defined as the portion of the game that he can conceive, given his cognitive bounds: a *frame* is any set $F = \times_{i \in N} F_i$ such that $\emptyset \neq F_i \subset S_i \forall i \in N$. The associated *framed game* is the game $G^F = \{N, F, u\}$ with u restricted to F. [12]

Each agent is characterized by a bound ($\mathfrak{B}$) for the number of items that he can keep in his conceivable space, due to limits in attention or memory. Because of limited attention, the set of strategies in not completely known by agents: they have limited rationality and are therefore bounded to consider only a limited subset of the whole set of feasible actions. Each player has a different set available in his cognition; nonetheless, he is aware of this fact and he can expect that other players will "surprise" him acting in an unexpected way. When this happens, he learns the existence of other strategies, and his subset of conceivable actions can be enlarged, although not too much (if he focuses his attention on "new" strategies,

[10] See for example Costa-Gomes et al. 2001 or Cognitive Hierarchy Theory and Level-k reasoning.

[11] Here it is possible to see a similarity of our idea with the model *Categorize Then Choose* by Manzini and Mariotti, 2012.

[12] Similar concepts are a *retract* in Kalai and Samet, 1984 or a *block* in Myerson and Weibull, 2015.

he will forget some of the old ones); in the same way, if some strategies are not used for a long time, he will tend to exclude these non-activated strategies from his subset. Given the subset of strategies that they consider in their model, agents form beliefs and expectations about others' behavior. These beliefs are confirmed or not when choices are made.

4 A joint production model

In order to study the mechanism of frames, we are going to deal with a given two-players game in normal form, characterized by four strategies for each player. The game form has been constructed in order to have some realistic features of a context of joint production[13], where people have different background, capacity, and they can decide how much to contribute to a common goal, characterized by interdependence and complementarity. Moreover, the game is enough complicated so that its solution might not be trivial and the use of "small scale models" is justified by the complexity of the situation.

There are two agents; each one is initially endowed with a good of value g, that is the value for its generic use. Each agent can decide whether to make an investment that increases the value of his asset to $I > g$ at a cost x, with $I - x < g$, so that investing is not convenient in itself since the costs do not cover the benefits; nonetheless if the asset is used within a specific relation for which it was intended, then in this case the investment gives a higher value, because of the idiosyncratic nature of the investment.

The behavior of agents is characterized by two decisions in two different periods[14]: first, they decide whether to *invest* (i), making the value of their asset grow to I (if cooperation happens), or *not* ($\neg i$) leaving its value at g; next, each agent can choose whether to actually cooperate or not in the production stage. If an agent does not cooperate (strategy $\neg c$), he keeps his own assets for himself and

[13] As defined in Lindenberg and Foss (2011): "any productive activity that involves heterogeneous but complementary resources and a high degree of task and outcome interdependence".

[14] The model is inspired by the inter-temporal model of Grossman and Hart, 1986; Hart and Moore, 1990

does not put them into the joint production process; if instead he decides to cooperate (strategy c) he makes his personal assets available for the joint production.

The game with two players (a_1 and a_2) can be described through the extensive form in Figure 4 and its corresponding normal form (Figure 5).

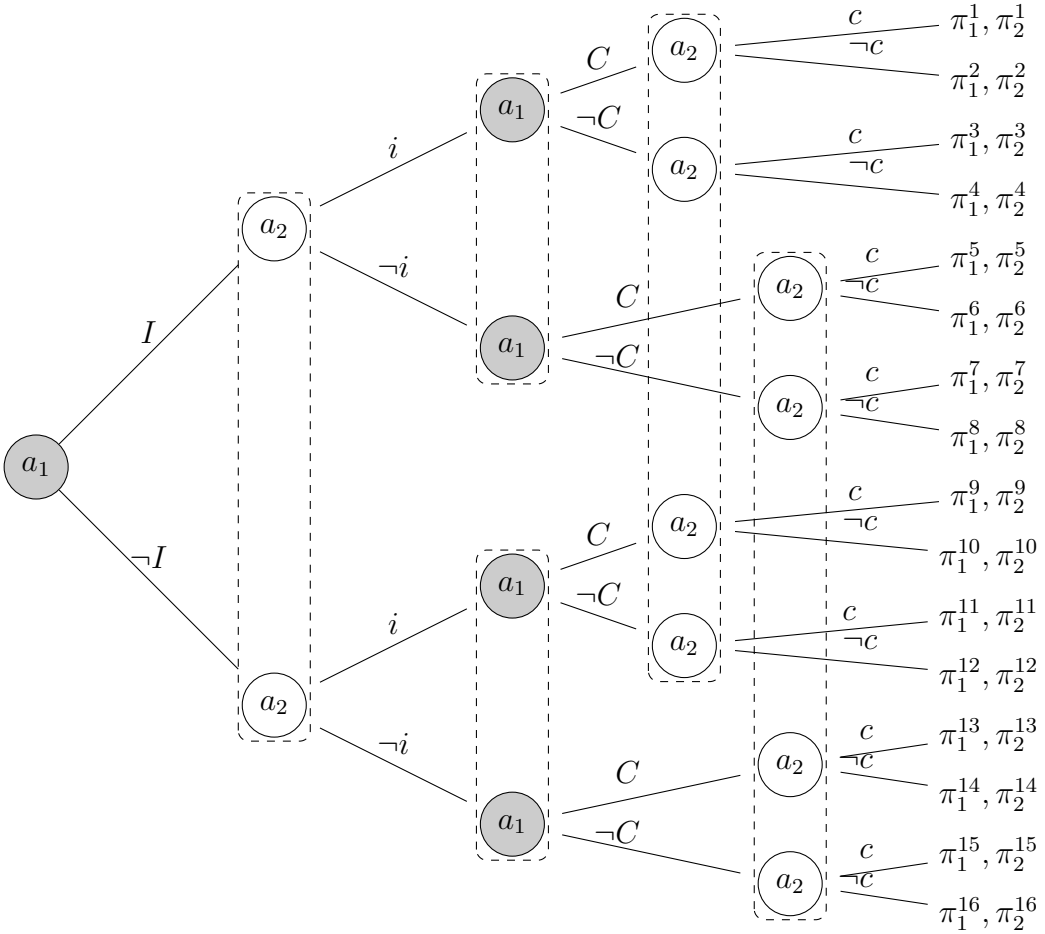

Figure 4: The game in extensive form.

Joint production uses the assets that are brought together by cooperative choices, and outputs reflect the complementarity of resources.

We denote with k_i the assets that player i puts in common; with r_i the assets he keeps for himself. Note that the costs of the investment can only be private.

		Player $\mathbf{a_2}$			
		i, c	$i, \neg c$	$\neg i, c$	$\neg i, \neg c$
Player $\mathbf{a_1}$	I, C	π_1^1, π_2^1	π_1^2, π_2^2	π_1^5, π_2^5	π_1^6, π_2^6
	$I, \neg C$	π_1^3, π_2^3	π_1^4, π_2^4	π_1^7, π_2^7	π_1^8, π_2^8
	$\neg I, C$	π_1^9, π_2^9	π_1^{10}, π_2^{10}	π_1^{13}, π_2^{13}	π_1^{14}, π_2^{14}
	$\neg I, \neg C$	π_1^{11}, π_2^{11}	π_1^{12}, π_2^{12}	π_1^{15}, π_2^{15}	π_1^{16}, π_2^{16}

Figure 5: The game in normal form.

For each strategy, we can list what the agent decides to put in common (k_i) and what he keeps as private (r_i).

strategy	k_i **common**	r_i **private**
i, c	I	$-x$
$i, \neg c$	0	$I - x$
$\neg i, c$	g	0
$\neg i, \neg c$	0	g

Table 1: Assets put in common and kept private for each strategy

		Player $\mathbf{a_2}$			
		i, c	$i, \neg c$	$\neg i, c$	$\neg i, \neg c$
Player $\mathbf{a_1}$	i, c	$I + I$	I	$g + I$	I
	$i, \neg c$	I	0	g	0
	$\neg i, c$	$g + I$	g	$g + g$	g
	$\neg i, \neg c$	I	0	g	0

Table 2: Inputs of production

We apply a very naïve production function, where the assets that were pooled in common (k_1 and k_2, summarized in Table 2) are used as inputs:

$$\Pi = f(k_1, k_2) = \gamma(k_1 + k_2).$$

The parameter γ captures interdependence and complementarity of investments: it has the highest value when both invest and cooperate, a bit smaller when one

invests and the other cooperates, and still smaller when at most one invests and both cooperate. Finally, it has no effect ($\gamma = 1$) in all the other cases.

Individual payoffs are determined by:

$$\pi_i = \rho_i \cdot \gamma(k_1 + k_2) \tag{1}$$

When cooperation happens, joint work entails some *coordination costs* so that what is produced does not completely enter the payoffs, but it is weighted by the parameter $\rho_i \leq 1$: costs are null when nobody cooperates, medium when only one does, high when both work together.

Using real values for the parameters discussed above[15], a possible scenario is given by the payoffs shown in Figure 6, where strategies are labeled $\{A, B, C, D\}$ for both players (a_1 and a_2).

$\pi_{1,2}$	A	B	C	D
A	$11.8, 11.8$	$-2.2, 5.5$	$1, 6$	$-2.2, 6.5$
B	$5.5, -2.2$	$2, 2$	$6.5, 3.6$	$2, 3$
C	$6, 1$	$3.6, 6.5$	$3.6, 3.6$	$1.2, 4.5$
D	$6.5, -2.2$	$3, 2$	$4.5, 1.2$	$3, 3$

Figure 6: The game in normal form, complete matrix (Objective Game Form).

[15] $I = 7, x = 5, g = 3; \gamma = \{4, 3, 2, 1\}; \rho = \{0.6, 0.9, 1\}$.

5 Frames and game representations

We use the complete matrix above as the *objective game form* and introduce an element of bounded rationality.

The objective game form G^O is characterized by $N = \{1, 2\}$, $S_i = \{A, B, C, D\}, i = 1, 2$ and π_i as listed in Figure 6 for each possible outcome of the game, i.e. for any element $\mathbf{s}$ of the set of feasible action profiles $S = S_1 \times S_2$.

If we impose a bound on players' ability to conceive items of the objective game form, this implies that the number of elements that they can represent in their mind cannot exceed a given threshold $\mathfrak{B}$: whenever this bound is smaller than the cardinality of the objective set, players are constrained to consider a subset of the possible patterns of behavior and their consequences.

To draw an example, let's consider the case in which the threshold is given by $\mathfrak{B}_1 = \mathfrak{B}_2 = 4$, namely the number of items that each player can conceive is limited to four. This implies that, even if they can be aware[16] of all their possible actions, they need to restrict their attention to a subset of them in order to evaluate their consequences and outcomes. Thus two possible subjective game forms are given in Figure 7 and Figure 8.

$\pi_{1,2}$	A	B	C	D
A	$11.8, 11.8$	$-2.2, 5.5$		
B	$5.5, -2.2$	$2, 2$		
C				
D				

Figure 7: Player 1's subjective game form, $F^1 = \{A, B\} \times \{A, B\}$

Players choose their action on the basis of their models of the world, in order to achieve the highest payoff, given that they expect the other player to do the same, within the frame that they conceive. Time consists of an infinite sequence of periods, each denoted by t, within each of which agents choose and implement actions. We assume here that the characteristics of the domain will be stationary over all periods. Each agent has a constant discount factor δ; for simplicity we

[16]Modica and Rustichini(1994), Karni and Viero (2013), Li (2009).

$\pi_{1,2}$	A	B	C	D
A				
B	$5.5, -2.2$			$2, 3$
C	$6, 1$			$1.2, 4.5$
D				

Figure 8: Player 2's subjective game form, $F^2 = \{B, C\} \times \{A, D\}$

$\pi_{1,2}$	A	B	C	D
A	$11.8, 11.8$	$-2.2, 5.5$	$1, 6$	$-2.2, 6.5$
B	$5.5, -2.2$	$2, 2$	$6.5, 3.6$	$2, 3$
C	$6, 1$	$3.6, 6.5$	$3.6, 3.6$	$1.2, 4.5$
D	$6.5, -2.2$	$3, 2$	$4.5, 1.2$	$3, 3$

Figure 9: *Objective* and *Framed* Nash Equilibria (in gray and light gray)

assume the discount factor to be zero, meaning that agents are completely myopic within the time horizon and limited only to the current period. After each stage, players can observe the behavior of their opponents and the outcome in terms of payoffs for all the players.

What does this representation of mental models imply in terms of behavior, beliefs and equilibria?

First, observe that if agents have limited representations of the game, the conceivable games are many (for example there are 36 possible 2x2 games within the 4x4 original matrix, each of them having one or more Nash Equilibria). Thus when bounded rationality is included in the picture, the number of Nash Equilibria is different than with perfect rationality, as there are many strategy profiles that are sustainable in equilibrium within the framed games.

As shown in Figure 9, the following strategy profiles are Nash equilibria of the objective game in pure strategies: $(A, A); (B, C); (C, B); (D, D)$ but there are many other equilibria that can be sustained within small framed games: (A, A); (B, B); (B, C); (B, D); (C, B); (C, C); (C, D); (D, B); (D, C); (D, D).

When people economize on their models of the world, they can get stuck in points that would not be chosen if all the possibilities were considered. And if participants in human interaction hold the same mental model and coordinate their

$\pi_{1,2}$	A	B	C	D
A	$11.8, 11.8$	$-2.2, 5.5$	$1, 6$	$-2.2, 6.5$
B	$5.5, -2.2$	$2, 2$	$6.5, 3.6$	$2, 3$
C	$6, 1$	$3.6, 6.5$	$3.6, 3.6$	$1.2, 4.5$
D	$6.5, -2.2$	$3, 2$	$4.5, 1.2$	$3, 3$

Figure 10: The result (B,C) can be justified under several frames, for example: $F = \{A, B\} \times \{B, C\}$ and $F = \{B, C\} \times \{C, D\}$

choices on an equilibrium of that game, which is confirmed at each stage, these behaviors can be sustained for long time within a community or a group with the same views of the world.

Still, another possibility arises: players might have different representations of the game, but these different subsets of the objective game partially overlap on an action profile that is an equilibrium in all the different models they have in mind. In this case, the actual behavior of other players reconfirms beliefs and expectations that an agent has, given his own mental model, and nobody has a reason to change his behavior or beliefs, although players hold different views of the game, because they have compatible ways of playing, although the beliefs on the off-the-play path differ.

Figure 10 shows how the same outcome can be justified under different representations: equilibrium does not require common knowledge of the game form. Agents having the same mental model will confirm their beliefs observing actual behavior, but this can happen also to agents with different mental representations of the game: different mental models can bring to the same equilibrium.

Moreover, in most cases agents experience a disequilibrium, not only in behaviors but also in beliefs: players hold different views of the world, i.e. they consider different subsets of all the feasible actions, and equilibrium choices differ: in this case, when they play, they observe an outcome that contradicts their mental model and thus they are induced to change their behavior or cognition. This process of interactive revision of game representations is driven by observation: when players observe an outcome that does not match their mental model, they will react adding the opponent's action that was actually played by the other

into their frame. In doing this, they will replace the action they expected from the other with the observed one, thus keeping a 2x2 representation of the game, incorporating actions done by other players through a dynamic adaptation[17].

Consider for example a situation in which two players, 1 and 2, play the game with the following subjective game models: $F^1 = \{A, B\} \times \{A, B\}$ and $F^2 = \{B, C\} \times \{A, D\}$. Suppose that, given F^1, player 1 chooses action B while, given F^2, player 2 chooses action D. Then the chosen strategy profile $(B, D) \notin F^1$, namely it is not coherent with player 1's mental representation of the game, and this observation will lead him to revise his own subjective game form, incorporating the action chosen by the other player and eliminating the one he was expecting, i.e. B, from the subset of strategies available to his opponent: $B \notin F_2^1; D \in F_2^1$. Now his revised game form is given by $F^1 = \{A, B\} \times \{A, D\}$ and his choice is again B. The second player needs not revise his own model, since the result of interaction was compatible with his starting representation of the game. Now they play again and they observe the strategy profile: $(B, D) \in F^1 \wedge F^2$: they have reached an equilibrium since they now have no incentives to change their behaviors or beliefs until they meet another player whose choice contradicts their model.

6 On the origin of frames

Aoki (2001, p. 201) suggested the possibility "that a summary representation can emerge even before the game can precisely locate a corresponding equilibrium, and even precedes it and guides the agents to find it" . How do agents form this summary representation? There are several ways in which a frame can be formed in the individual's mind.

Of course it is possible that there is an initial focusing mechanism due to education and past experience[18]: agents learn how to act in society first in small groups, where interactions are much simpler and they are taught the basic rules in

[17] See Gavetti and Levinthal, 2000 for a similar result in non-strategic contexts.

[18] As in Piaget's theory of development.

a parsimonious way, so that they tend to conceive the world under a representation that is given by culture in prototypical situations. In these contexts, the subjective game form is given by some authority and newbies learn it as it is told to them, and one of the possible equilibria is played, which re-confirms their beliefs about the world, thus becoming an institution in Aoki's sense.

Their beliefs will be transferred[19] to the other games they play, but when they exit these little worlds and they face new situations, playing with agents who do not share their mental model, they will experience a disequilibrium: not only in actions, but in their own cognitions. And this leads them to discover new possibilities.

Thus agents may enter the interaction with their own frame in mind for any valid reason, due to previous interactions, transfer mechanisms, education, prototypes and many other factors and then evolution leads them to converge on a particular pattern of behavior.

But we suggest that a *normative* system of beliefs, preceding the evolution of the corresponding equilibrium, can become accepted by all agents in the relevant domain, entering their shared mental model of how the game *should* be played and hence becoming the basis for their coordination on a specific equilibrium.

A norm with normative meaning and content will not simply select one equilibrium among the many possible, but by its prescriptive and universalizable meaning will be able to justify its shared acceptance by all participants.

Such a norm "logically" comes before than any exogenous institution is overimposed on a given action domain, or before any institution has yet emerged.

7 Social Contract Reasoning

The best justificatory account for norms, entailing ex ante shared acceptance, is the social contract modelwhere norms result from a voluntary agreement in a hypothetical original choice situation.

In Rawls (1971) the principles of justice for the basic structure of society are

[19]Gick and Holyoak, 1980; Holyoak and Thagard, 1996; Knez and Camerer, 2000.

the object of an original agreement, that is understood as a purely *hypothetical situation* characterized so as to lead to a certain conception of justice (justice as fairness).

The original position of equality corresponds to the state of nature in the traditional theory of the social contract, where no one knows his place in society, his class position or social status, nor does any one know his fortune in the distribution of natural assets and abilities, his intelligence, strength, and the like. The principles of justice are chosen behind a veil of ignorance.

We propose that reasoning under the veil of ignorance can foster a frame in the mind of agents, through a cognitive mechanism that generates peculiar beliefs and expectations which are able to persist in the agent's mind and sustain the selected equilibrium also in the "game of life" (Binmore, 2005), even in absence of common knowledge.

In particular, we claim that an agreement among players is able to activate a peculiar way of reasoning that is compatible with constraints on bounded rationality that we discussed above, but on the other side is able to generate shared representations leading to a fair outcome, since it has the properties of mutual advantage.

At the cognitive level, the role of an agreement beyond the veil of ignorance is to activate a "symmetric" mental model: symmetry becomes focal and only symmetric frames are conceived.

This role of the social contract is explained through its main characteristics: impersonality, impartiality and prescriptivity.

The first step is the application of the principle of *impersonality*: since the selected equilibrium must not depend on personal and social positions, the veil of ignorance symmetrizes players by assigning them the same strategy space.

This requires that for any conceivable action, this might be thought as possible for *any* player, i.e.:

$$\forall s \in F_i^i \Rightarrow s \in F_j^i$$

When players consider themselves to be unequal, the veil symmetrizes them

by assigning them the same strategy space:

$$F_i = F_j \forall i, j \in N$$

But this might be problematic in requiring too much for a bounded rational agent, as it requests to take into consideration too many strategy profiles, thus exceeding his cognitive threshold, $\mathfrak{B}$.

Players can still have a limited model of the world, but impersonality requires agents to be equal and interchangeable, so that only subjective game models that are symmetrical (that are invariant under the players' position exchange) can be activated in the agents' mind.

The second step is given by the principle of *impartiality*: players must agree on an outcome under the hypothesis that each player has equal probability of finding himself in the position of each of the possible two roles

This requires that strategy profiles with asymmetrical outcomes are not taken into consideration. Only payoffs that are symmetrical (that are invariant under the players' position exchange) can be accepted.

Through the application of these two principles, agents create a summary representation of the game that considers the *diagonal* of the objective game form: this restriction is compatible with their cognitive bounds as it leads them to consider a number of outcomes that does not exceed their computational threshold.

Note that all these payoffs pair on the diagonal are Pareto ranked; only a subset of them are equilibria but from the perspective of the social contract the collective features prevail. A Pareto ranking is obviously a selection criterion: the best outcome from Pareto point of view is chosen, and a unique equilibrium point is selected, the one Pareto dominant .

Finally, *prescriptivity* comes into the picture. Since the reasoning that leads to consider the diagonal of the game has prescriptive and universalizable meaning able to justify its shared acceptance by all participants, it enters their shared mental model of how the game should be played and hence becomes the basis for their coordination on a specific equilibrium.

This is the opposite of the typical *naturalistic fallacy*: it isn't the case that an "is" entails an "ought"; on the contrary it is *because* players had a normative

reason to act according to fairness under the though experiment of the veil of ignorance, that then they hold a mental model within which it is normal to think that the players act according to the social contract prescription

So, in conclusion, Social Contract reasoning works as an equilibrium selection device, in Binmore's sense but with a cognitive interpretation, as *no evolutionary interpretation* is needed.

8 References

Aoki, M. (2001) Toward a Comparative Institutional Analysis, Cambridge, MA: MIT Press.

Aoki, M. (2010), Corporations in Evolving Diversity: Cognition, Governance, and Institutions, Oxford University Press.

Aoki, M. (2011), "Institutions as Cognitive Media between Strategic Interactions and Individual Beliefs", Journal of Economic Behavior and Organization, 79(1–2), pp. 20-34.

Bacharach, M. (1993), Variable Universe Games, in K. Binmore. A. Kirman and P. Tani (eds.), Frontiers of Game Theory, MIT Press.

Bacharach, M. (2006). Beyond individual choice: teams and frames in game theory. Princeton University Press.

Binmore, K. (2005) Natural Justice, Oxford: Oxford University Press.

Buchanan, J. (1975), The Limits of Liberty: Between Anarchy and Leviathan, Library of Economics and Liberty [Online] available from http://www.econlib.org/library/Buchanan/buchCv7.html.

Costa-Gomes, M., Crawford, V. P., and Broseta, B. (2001), Cognition and behavior in normal-form games: An experimental study. Econometrica, 69(5), 1193-1235.

Crawford V. (2013), Boundedly Rational versus Optimization-Based Models of Strategic Thinking and Learning in Game, Journal of Economic Literature, 51 (2), pp. 512-527.

Denzau, A. and North, D. (1994), Shared Mental Models: Ideologies and Institutions, Kyklos, 47(1), pp. 3-31.

Devetag, G. and Warglien M. (2008), Playing the wrong game: An experimental analysis of relational complexity and strategic misrepresentation, Games and Economic Behavior, 62, pp. 364-382.

Esponda, I. and Pouzo, D. (2014), An Equilibrium Framework for Players with Misspecified Models, Working Paper.

Faillo M. Sacconi, L. and Ottone, S. (2015), The social contract in the laboratory. An experimental analysis of self-enforcing impartial agreements, Public Choice (2015), 163:225–246.

Gauthier, D. (1986), Morals by Agreement, Oxford: Clarendon Press.

Gavetti, G. and Levinthal, D. (2000), Looking Forward and Looking Backward: Cognitive and Experiential Search, Administrative Science Quarterly, Vol. 45, No. 1, pp. 113-137.

Gick, M.L., and Holyoak, K.J. (1980) 'Analogical problem solving', Cognitive Psychology, 12, 306-355.

Grossman, S. J., and Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. The Journal of Political Economy, 691-719.

Halevy, N., Chou, E. Y. and Murnighan, J. K. (2012), Mind games: The mental representation of conflict, Journal of Personality and Social Psychology, 102 (1), pp. 132-148.

Harstad, R. and Selten, R. (2013), Bounded-Rationality Models: Tasks to Become Intellectually Competitive, Journal of Economic Literature, 51 (2), pp. 496-511.

Hart, O., and Moore, J. (1990). Property Rights and the Nature of the Firm.Journal of political economy, 1119-1158.

Hobbes, T. (1651), Leviathan: Or the Matter, Forme, and Power of a Common-Wealth Ecclesiasticall and Civill, (ed. by Ian Shapiro , Yale University Press; 2010).

Holyoak, K. J., and Thagard, P. (1996). Mental leaps: Analogy in creative thought. MIT press.

Holyoak, K.J. and Spellman, B.A. (1993), Thinking, Annual Review of Psychology, 44 (1), pp.265-315.

Johnson-Laird, P. (1983), Mental Models, Cambridge University Press.

Johnson-Laird, P. and Byrne, R. (1991), Deduction, Lawrence Erlbaum Associates.

Kalai, E., and Samet, D. (1984). Persistent equilibria in strategic games, International Journal of Game Theory, 13(3), 129-144.

Karni, E. and Vierø M. (2013), "Reverse Bayesianism": A Choice-Based Theory of Growing Awareness, American Economic Revie, 103 (7), pp. 2790-2810.

Knez, M., and Camerer, C. (2000). Increasing cooperation in prisoner's dilemmas by establishing a precedent of efficiency in coordination games. Organizational Behavior and Human Decision Processes, 82(2), 194-216.

Kreps, D. M. (1990), Game Theory and Economic Modelling, Oxford: Claredon Press.

Kruglanski, A. W., Shah, J. Y., Friedman, R., Fishbach, A., Chun, W. Y., and Sleeth-Keppler, D. 2002. A theory of goal systems. Advances in Experimental Social Psychology, 34: 331–378.

Legrenzi, P. Girotto, V. and Johnson-Laird P.N. (1993), Focussing in reasoning and decision making, Cognition, 49, pp. 37-66.Lewis 1969

Li, J. (2009), Information Structures with Unawareness, Journal of Economic Theory, 144 (3), pp. 977-993.

Lindenberg, S. and Foss, N. (2011), Managing Joint Production Motivation: The Role of Goal Framing and Governance Mechanisms, Academy of Management Review, Vol. 36, No. 3, 500-525.

Manzini, P., and Mariotti, M. (2012). Categorize then choose: Boundedly rational choice and welfare. Journal of the European Economic Association, 10(5), 1141-1165.

Modica, S., and Rustichini, A. (1994). Awareness and partitional information structures. Theory and Decision, 37(1), 107-124.

Myerson, R. and Weibull, J. (2015), Tenable Strategy Blocks and Settled Equilibria, Econometrica, 83 (3), pp. 943-976.

Rawls, J. (1971) A Theory of Justice, Oxford: Oxford University Press.

Sacconi, L. (2013), Ethics, Economic Organisation and the Social Contract, in A. Grandori (ed), Handbook of Economic Organization: Integrating Economic and Organization Theory, Edward Elgar Publishing, pp. 112-136.

Sacconi, L. and Faillo, M. (2010), Conformity, Reciprocity and the Sense of Jus-

tice. How Social Contract-based Preferences and Beliefs Explain Norm Compliance: the Experimental Evidence, Constitutional Political Economy, 21 (2), 171–201.

Sacconi, L., Faillo, M., and Ottone, S. (2011), Contractarian Compliance and the Sense of Justice: A Behavioral Conformity Model and Its Experimental Support, Analyse und Kritik, 33(1), 273-310.Schelling 1960

Sugden, R. (1995), A Theory of Focal Points, The Economic Journal, 105 (430), pp. 533-550.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll H. (2005), Understanding and sharing intentions: The origins of cultural cognition, Behavioral and Brain Sciences, 28, pp. 675-735.

Tooby, J. and Cosmides, L. (1989). Evolutionary psychology and the generation of culture, Part I. Theoretical considerations. Ethology and Sociobiology, 10, 29-49.