INSTITUTIONS, FRAMES, AND SOCIAL CONTRACT REASONING

BY VIRGINIA CECCHINI MANARA♣ AND LORENZO SACCONI♦

Draft – August 2014

.

Abstract: This work aims at filling a gap in the cognitive representation of institutions, starting from Aoki's (2001, 2010) account of institutions as equilibria in a game-theoretical framework. In this view, an institution refers to that portion of agents' equilibrium beliefs common to (almost) all of them regarding how the game is actually played.
Within this framework, the inquiry about the mechanism of institutional change can be articulated in different questions: how do individual agents come to accept a specified pattern and follow it as their own cognitive frame? How does the public representation influence individual beliefs and actions? And how is it possible that different agents, with different knowledge and preferences, coordinate mutual beliefs? Which is the role of social contract in the formation of these convergent beliefs?
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 actions (strategies) that they can play, because they have bounded rationality and limits in memory and attention; grounding on the works by Johnson-Laird and his coauthors (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991), 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.
Second, we rely on Bacharach's variable frame theory (Bacharach, 2006): agents transform the objective game, representing the interaction between players as it "really is" or as the theorist describes it, into a framed game, where strategies are "labeled" in some sense, "that is, have symbolic or connotative characteristics that transcend the mathematical structure of the game" (Schelling, 1960).
Aoki (2011) suggests that "there ought to be some public representation that mediates between the equilibrium play of a societal game and individual belief formation". He refers to an "external media" or artifact that linguistically represent salient features of equilibrium plays (such as norms, rules organizations of known types, laws). Nonetheless, to have a "social order" common knowledge is not required.
In such a context, we argue that a social contract – given its prescriptive and universalizable meaning – may provide a shared mental model (Denzau and North, 1994), accepted by all players, that allows agents to select a joint plan of action corresponding to an efficient and fair distribution (Sacconi, 2010).

♣ Institute of Economics, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà 33, 56127 Pisa, Italy (e-mail: v.cecchinimanara@sssup.it)

♦ Department of Economics and Management, University of Trento, via Inama 5, 38100 Trento, Italy and EconomEtica (e-mail: lorenzo.sacconi@unitn.it)

1. The cognitive aspect of institutions

This work aims at filling a gap in the cognitive representation of institutions, starting from Aoki's (2001, 2010) account of institutions as equilibria. In this view, institutions "are not rules exogenously given by the polity, culture or a meta-game" but "rules created through the strategic interaction of agents, held in the minds of agents and thus self-sustaining" (Aoki, 2001, p. 11).

As such, institutions entail a dualistic nature: on one side they constrain individual choices *by coordinating their 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 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.*** Aoki (2011) suggests that «there ought to be some *public representation* that mediates between the equilibrium play of a societal game and individual belief formation». He refers to an «external media» or artifact that linguistically represent salient features of equilibrium plays (such as norms, rules organizations of known types, laws).

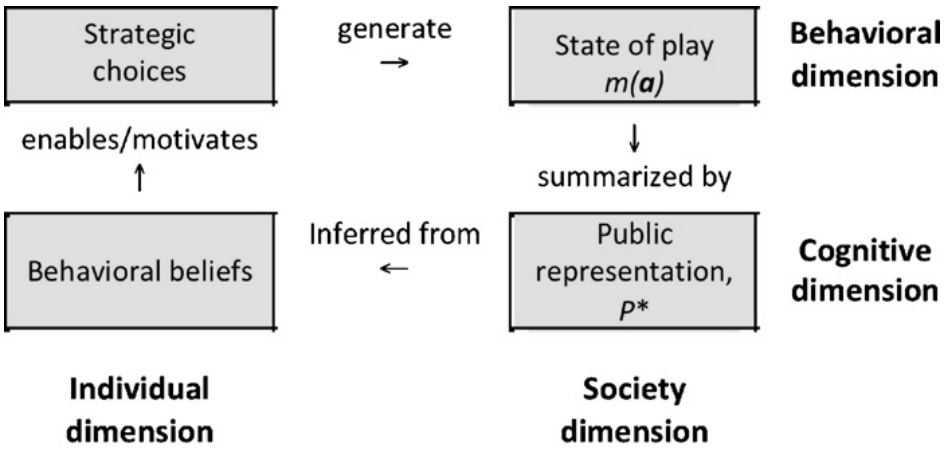

FIGURE 1. THE MEDIATING ROLE OF INSTITUTIONS IN SUBSTANTIVE FORM (SOURCE: AOKI, 2011)

In his account of institutions (see Figure 1), cognitive components (i.e., beliefs deriving from compressed mental representations of salient aspects of ongoing equilibrium play) and behavioral components *(*i.e., the iterated play of a given set of equilibrium strategies) are interlocked in a recursive scheme. The starting point is cognitive, and it consists in pattern recognition whereby given situations of interaction are framed as games of a certain form wherein players are expected to reason in a given quasi-symmetrical way. At step two, this framing of the situation induces players to entertain quasi-converging beliefs about a certain mode of playing the game. Thus, at step three, on passing from beliefs to the players' actual behavior, each player adopts a tentative strategy based on the belief that others will also adopt strategies consistent with the aforementioned mode of behavior. Hence, in step four, strategies clash and some of them prove to be more successful and based on a better prediction. By trial and error, therefore, strategies converge towards an equilibrium of the game. This may be construed as an evolutionary result because the mode of playing attracts more and more players through iterated adaptation to the other players' aggregate behaviors in the long run. At each repetition, however, this evolving equilibrium is summarily represented in its salient features by a compressed mental model resident in the players mind so the fifth step concluding the circle is again cognitive.

This circle can be recursively iterated so that the ongoing equilibrium mode of playing is repeatedly confirmed by beliefs that translate into equilibrium behaviors, which are represented summarily by mental models, and so on. At some point, this belief system reaches a nearly complete state of 'common knowledge' (Lewis, 1969; Binmore and Brandenburger, 1990) about how players interact. The resulting equilibrium is an institution: a regularity of behavior played in a domain of interaction and stably represented by the shared mental model resident in all the participants' minds. It is essentially equivalent to the notion of social norm as a 'convention.'

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.

We admit that Aoki's framework is the most complete and useful treatment of the concept of institutions, in order to analyze their emergence and stability, but it still misses something: institutions in the above game-theoretical definition only *ex post* tell each player what the best action is. 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. 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 find one proposal in Sacconi's recent works (see for example Sacconi 2012), where a modified version of Aoki's account is presented, introducing the *social contract* as the cognitive mechanism by which a norm may be accepted and become a shared mental model.

The Rawlsian social contract has been vindicated by Binmore (2005), who has shown how the social contract (a normative ethics principle) provides a source for the **selection** of an equilibrium in the *ex ante* problem. But what happens when agents exit from the original position? The ex-post game (beyond the veil of ignorance) is a context of choice different from the ex-ante game (behind the veil of ignorance)

When the veil is lifted, they return to the game of life, and are back to the initially possible set of equilibria. There is no logical basis to say that because they knew the *ex ante* solution then they also know that the egalitarian solution is the ongoing solution of the game of life *ex post*. 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 the 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.

Aoki's recursive model can be reformulated, adding a social norm that derives from social contract reasoning employed by players in order to agree on basic principles and norms when equilibrium institutions are not already established (see Figure 2).

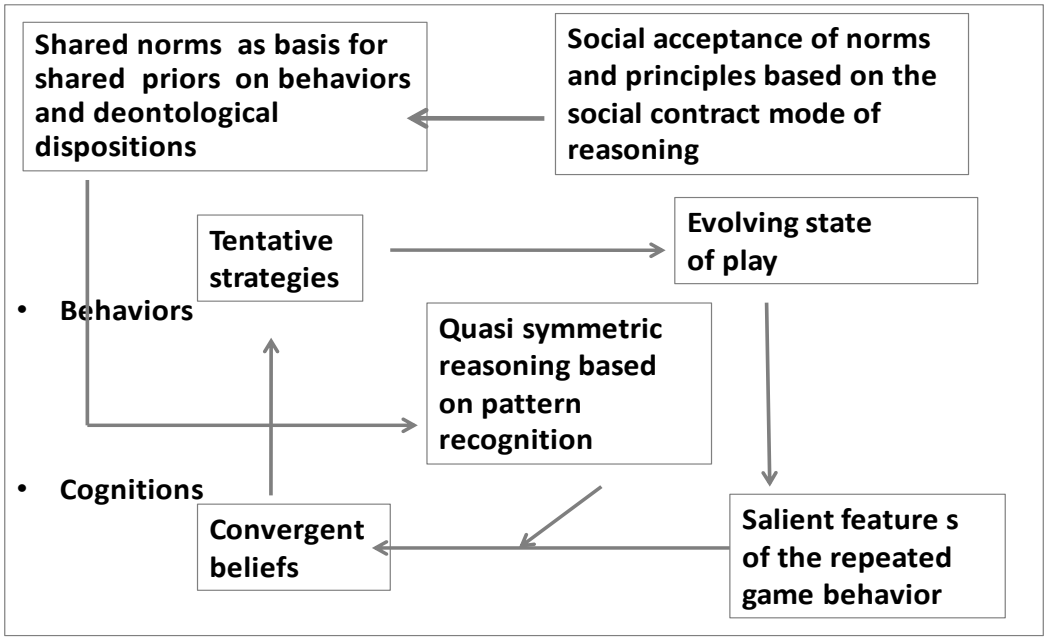

FIGURE 2, AOKI'S MODIFIED DIAGRAM REPRESENTING THE RECURSIVE PROCESS OF INSTITUTION FORMATION

The normative meaning of norms does not depend on knowledge about the ongoing behavior of other players. Instead, norms are able to justify and give first-place reasons for shared acceptance of a mode of behavior addressing all the participants in a given interaction domain before it has been established as an equilibrium point. A norm gives intentional reasons to act independently on the evolutionary benefits of adaptation in the

long run because when an individual or a group of agents in a given action domain initiate an institutional change, it cannot stem from the pressure of evolutionary forces, which unfold their attraction only in the long run. Instead, a norm enters the players' shared mental model (Denzau and North, 1994) of how the game should be played, shapes the players' reciprocal disposition to act and their default beliefs about common behaviors, and hence becomes the basis for their first coordination on a specific equilibrium. In other words, it works as the first move in a process of equilibrium selection that activates the recursive process outlined by Aoki.

We propose a formal model trying to explain this role of the social contract, looking at situations in which agents have bounded rationality and they need to coordinate with each other. First of all, we consider agents with bounded rationality: in particular, we assume that agents do not know all the feasible actions (strategies) that they can play, because they have limits in memory and attention; grounding on the works by Johnson-Laird and his coauthors (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991), 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.

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 the game.

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 been investigated enough. In particular, we claim that the participation of players in a social contract reasoning may activate a cooperative frame, signaling a pattern of behavior that is common to everybody, since it has the properties of mutual advantage.

2. Bounded Rationality in Games and Cognitive Constraints

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 mental models.

In particular, we will propose that bounded rationality constraints agents to consider only small portions of big games. The novelty of this approach stays 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.

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.

Psychologists have studied for long time this cognitive phenomenon: for example, Legrenzi, Girotto and Johnson-Laird (1993) have shown how people tend to focus on the information that is explicit in the description of a problem; the application by individuals of successful solutions, behavioral routines, or - more in general - knowledge, across different domains is usually referred to in the psychological literature as "transfer" (Gick and Holyoak, 1980). Gavetti and Warglien (2007) model individual representations of situations as clusters of features, following a long tradition in cognitive psychology; relatedly, consistent with work on categorization, their model assumes that individual memory is organized in terms of prototypical situations, or experiences, that correspond to different clusters of correlated features.

On the other hand, economists have paid little attention to cognitive components of mental representations of problems, especially for those involving strategic interaction

(games). 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 (see for example Devetag and Warglien, 2008):

- a difficulty in managing complex, non-projective, structures of payoffs, where elements of competition and coordination are mixed, lead 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.

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. In fact, as we will describe in greater detail, their limitations constrain them to see only small portions of big games: they consider subsets of strategies among the many feasible actions, activating "a small subset as a repertoire for strategic choice" (Aoki, 2001, p. 205). 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, 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.

The idea underlying the model is the conceptualization of institutions given by Aoki (2001): he states that institutions can refer to that portion of agents' equilibrium beliefs common to (almost) all of them regarding how the game is actually 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).

3. The Model

3.1 Definitions

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. *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.

The combination of actions by all agents in one period is called an *action profile*, and an actually realized action profile is the (internal) *state of the domain.* Outcomes are expressed in terms of monetary payoffs.

Each agent has a constant discount factor δ; for simplicity we assume the discount factor to be zero, meaning that agents are completely myopic within the time horizon and limited only to the current period.

We consider games that are *recurrently played in society*. At every period, agents meet a random opponent, but with a high probability to meet the same person for many rounds. After every stage of the game, agents can publicly observe the result of the interaction: they can see the *actions* chosen by all the players and all the realized *payoffs*.

We assume that each agent chooses an action in each period in order to maximize his payoffs from his action choices, even if they are bounded in their abilities to do so.

3.2 Joint Production in knowledge contexts

In order to have a treatable example, we consider a special case of joint production, 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", with a special focus on knowledge creation (the context we have in

mind is one where joint effort can create a surplus, such as research in a University Department), where people have different background, capacity, and they can decide how much to contribute to the common goal (typical problem of cooperation and coordination).

Let's suppose that, in order to produce a valuable output, researchers can choose many different strategies. First of all, they have to choose how much to invest in their own human capital: this is positively related to the final output, and it is captured by the parameter v that can be high or low ($v_{LOW};\ v_{HIGH}$).

Next, they can choose whether to cooperate or not in the division of the surplus, which means that they choose how much to call for themselves (spending assets in order to steal from the other).

Disposition to cooperate can be captured through the parameter c (the "cost of fight"), that is the effort spent to fight and take value from the other. In this case we suppose that an agent who starts with low endowment will spend a part of just to steal value from the other person; while a person with high initial endowment will choose to spend more in case he meets another with high value, while he will spend less when he meets an agent with low endowment.

Among the many possible actions, we consider four possible behaviors, defined by different levels of *effort* in production and *cooperative attitude* in sharing the surplus:

A- Puts little effort in production (v_{LOW}) and he is willing to cooperate ($c = 0$)

B- Puts little effort in production (v_{LOW}) and he uses a part of his endowment to steal value from others ($0 < c < v_{LOW}$).

C- Gives a valuable help in production (v_{HIGH}) and he is willing to cooperate ($c = 0$)

D- Gives a valuable help in production (v_{HIGH}) and he uses a part of his endowment to steal value from others

$$\left(0 < c_f \; and \begin{cases} c < v_{LOW} \text{ if he meets an agent with low value} \\ v_{LOW} < c < v_{HIGH} \text{ if he meets another agent with high value} \end{cases}\right)$$

We assume that these four actions are feasible by both players in a symmetric way: $\mathrm{S} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}; \mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$.

The surplus of cooperation between two agents, i and j is thus calculated through the following formula:

$$\Pi = \left(v_i + v_j - c_i - c_j\right) * m_{ij}$$

Where m is a multiplier for cooperation: when two cooperative agents meet, the multiplier is higher than in the case of no mutual cooperation.

Values:

$$v_a = v_{LOW}$$
$$v_b = v_{LOW}$$
$$v_c = v_{HIGH}$$
$$v_d = v_{HIGH}$$

Fight costs:

$$c_a = c_{LOW}, c_b = 0, c_c = 0$$

$$c_d = \begin{cases} c_{LOW} \text{ facing an agent with } v_{LOW} \\ c_{HIGH} \text{ facing an agent with } v_{HIGH} \end{cases}$$

Multipliers:

$$m_{bb} = m_{cc} = m_{bc} = m_{COOP} \qquad otherwise: \quad m = 1$$

In order to have a treatable example, let's use the values $v_{LOW} = 6$, $v_{HIGH} = 12$, $c_{LOW} = 3$, $c_{HIGH} = 9$, $m_{COOP} = 2$ that yield to this matrix:

Π	**a**	**b**	**c**	**d**
A	**6**	**9**	**15**	**12**
B	**9**	**24**	**36**	**15**
C	**15**	**36**	**48**	**15**
D	**12**	**15**	**15**	**6**

Individual payoffs are determined as a proportion of marginal contributions plus (minus) a gain (loss) equal to the cost of fight for the agent who held it (for the other player) .

$$\pi_i = \Pi * \frac{v_i}{v_i + v_j} + f_i - f_j$$

where $f_i = c_i$ and $f_j = c_j$.

Complete Matrix

π_{ij}	a	b	c	d
A	$3,3$	$7.5,1.5$	$8,7$	$4,8$
B	$1.5,7.5$	$12,12$	$12,24$	$2,13$
C	$7,8$	$24,12$	$24,24$	$-1.5,16.5$
D	$8,4$	$13,2$	$16.5,-1.5$	$3,3$

FIGURE 3, THE COMPLETE MATRIX WITH 3 NE IN PURE STRATEGIES

3.3 Subjective Game Models

We assume that "agents hold subjective, compressed views regarding the structure of the game they play - the subjective game models - and revise them in interactive and innovative ways when they face large external shocks and/or cognitive crises that the internal dynamics of the objective game endogenously generate" (Aoki 2001, p. 183).

In particular, if a game is defined by the *objective game form* $G_O = \{N, S, u\}$ where N is the set of agents, S the set of all technologically feasible actions and u the consequence function, the agent read it through the *subjective game form* $G_S = \{N_S, S_S, u_S\}$ that represents the same elements in the agent's mind. The assumption we make here is that $S_S \subseteq S$.

We call S_i the player i's frame and we impose, for example, that its cardinality cannot exceed 4, because of cognitive constraints, namely that each player will consider only 2 strategies available for himself and 2 strategies for the opponent.

Each player might hold a different frame among the 36 possible representations of 2x2 games, as in Figure 4.

Player i's frame: $S_i = \{A, B; a, b\}$

$\pi_{i,j}$	a	b	c	d
A	3, 3	7.5, 1.5	8, 7	4, 8
B	1.5, 7.5	12, 12	12, 24	2, 13
C	7, 8	24, 12	24, 24	−1.5, 16.5
D	8, 4	13, 2	16.5, −1.5	3, 3

Player j's frame: $S_j = \{B, C; a, d\}$

$\pi_{i,j}$	a	b	c	d
A	3, 3	7.5, 1.5	8, 7	4, 8
B	1.5, 7.5	12, 12	12, 24	2, 13
C	7, 8	24, 12	24, 24	−1.5, 16.5
D	8, 4	13, 2	16.5, −1.5	3, 3

FIGURE 4, SUBJECTIVE GAME MODELS

4. Our Results

After each stage, players can observe the behavior of their opponents and the outcome in terms of payoffs for all the players. Many cases are possible:

- players hold the same mental model and coordinate their choices on an equilibrium of that game, which is confirmed at each stage;
- players have different representations of the game, but these different subsets of the big 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 incentive to change his behavior or cognition, 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;

- 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.

4.1 Number of Equilibria

If agents have limited representations of the game, the conceivable games are many, each of them will have one or more Nash Equilibria: therefore the number of sustainable equilibria expands.

When Bounded Rationality is included in the picture, the number of Nash Equilibria is bigger than with Perfect Rationality. Figure 5 shows this fact, limiting attention to the case of equilibria in pure strategies.

Framed Nash equilibria

$\pi_{i,j}$	*a*	*b*	*c*	*d*
A	3, 3	7.5, 1.5	8, 7	4, 8
B	1.5, 7.5	12, 12	12, 24	2, 13
C	7, 8	24, 12	24, 24	−1.5, 16.5
D	8, 4	13, 2	16.5, −1.5	3, 3

FIGURE 5, FRAMED NASH EQUILIBRIA

4.2 Compatible Frames

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, see Figure 6 for an example.

The same outcome can be justified by different representations: equilibrium does not require common knowledge of the game form.

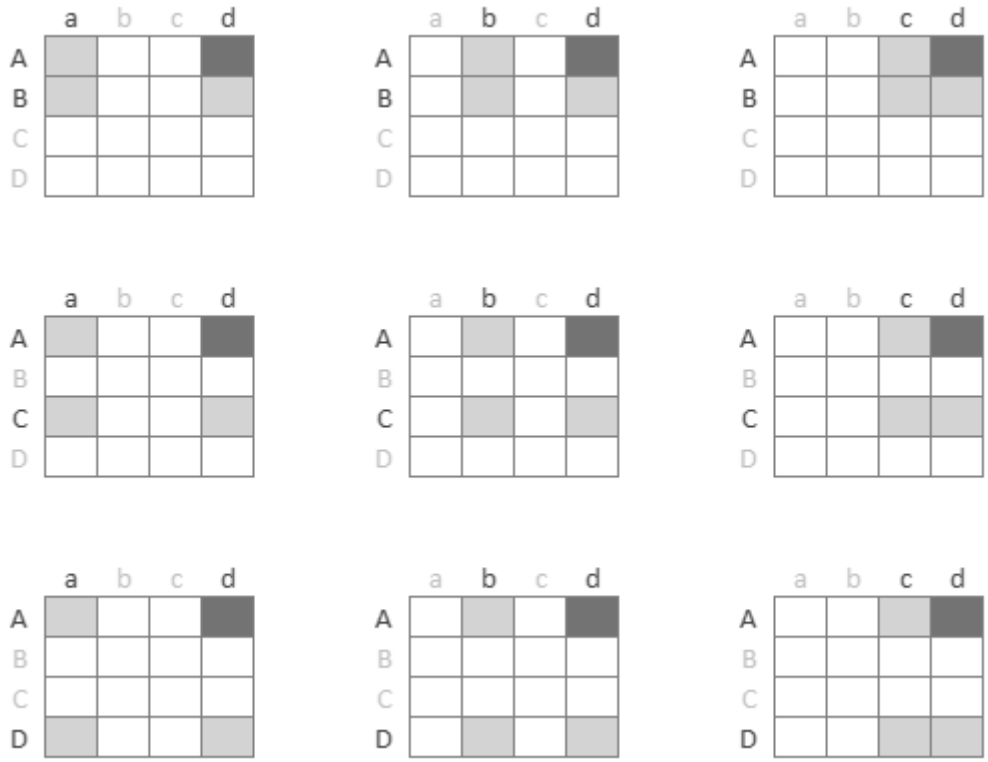

FIGURE 6, EXAMPLE OF COMPATIBLE FRAMES SUSTAINING THE EQUILIBRIUM (A,d)

4.3 Interactive Revision of Game Representation

In most cases agents experience a disequilibrium, not only in behaviors but also in beliefs. In particular, we suggest that when they 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 (See Gavetti and Levinthal, 2000 for a similar result in non-strategic contexts).

5. Social Contract

5.1 Basic Ideas – Social Contract Reasoning

Agents may enter the interaction with their own frame in mind, due to previous interactions, transfer mechanisms, education, prototypes and many other factors. But how is it possible to converge on a particular frame that sustains cooperation? We will suggest that the social contract (Hobbes, 1651, Buchanan, 1975, Gauthier, 1986, Rawls, 1971) might play a key role in this.

We follow Rawls (1971) accepting that "the original position is the appropriate initial status quo which insures that the fundamental agreements reached in it are fair" and therefore "no one should be advantaged or disadvantaged by natural fortune or social circumstances in the choice of principles". In our example, players ex ante pose themselves under a veil of ignorance assuming to be equal and interchangeable. It has already been shown (Binmore, 2005) how a Rawlsian Social Contract is able to solve the normative equilibrium selection problem through a decision procedure that satisfies elementary conditions of impersonality, impartiality and empathy. Under the ethical assumption of the veil of ignorance, an egalitarian solution is agreed.

In this work we want to show that this solution can enter the players' minds generating peculiar beliefs and expectations which are able to persist also in the "game of life": if players are able to consider the game under the ethical assumptions of justice as fairness, placing themselves in the original position, *ex post* (i.e. after they exit the hypothetical initial situation of equality), they will continue to conceive the interaction within the frame in which they entered *ex ante*.

The main intuition is that reasoning under the veil of ignorance can foster the "right" frame in the agents' minds (i.e. the one bringing to cooperation), through a cognitive mechanism that (Binmore, 2005), even in absence of common knowledge.

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, which is able to broaden the number of strategies that are taken into consideration: for any conceivable

action, this might be thought as possible for any player – we have an expansion of the considered subset of strategies. Nonetheless, this does not require to evaluate the complete 4x4 matrix, but allows agents to create a summary representation of the game that considers the diagonal, because their cognitive frame, acquired through the veil of ignorance, forces them to adjust their model in order to admit symmetry in actions.

Then, prescriptivity comes into the picture. When an agent is in front of a very big set of strategies, instead of focusing on the ones he already used before in his previous interactions (a kind of path dependence in mental models), he might partition this big set of actions into subsets, that we will call categories. The steps involved in this process are three: first, divide the space into small subsets; second, give a label to each of them; and third, choose on which to concentrate.

Categorization can be driven by different factors, and we want to suggest that social contract reasoning might induce a categorization driven by fairness considerations (Rawls, 1971).

The prescriptivity of social contract reasoning might help agents in choosing a particular subset of actions that can be described (or labeled) as fair for the joint production in the context of a social contract under veil of ignorance.

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 & Organization, 79(1-2), pp. 20-34.

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

Binmore, K. and Brandenburger, A. (1990), "Common Knowledge and Game Theory." In Ken Binmore, ed., Essays in the Foundation of Game Theory, 105–150. Oxford: Basil Blackwell.

Buchanan, J. (1975). The Limits of liberty. Between Anarchy and Leviathan, Chicago: Chicago University Press..

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.

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.

Gavetti, G. and Warglien, M. (2007), "Recognizing the New: A Multi-Agent Model of Analogy in Strategic Decision-Making", Strategy Unit Working Paper No. 08-028.

Gick, M. L., and Holyoak, K. J. (1980), "Analogical Problem Solving", Cognitive Psychology, 12, pp. 306-355.

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 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.

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

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

Lewis, D. (1969), Convention. A Philosophical Study. Cambridge, MA: Harvard University Press.

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.

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

Sacconi, L. (2012), "Ethics, Economic Organisation and the Social Contract", EconomEtica working paper, 41