Phenom Cogn Sci
DOI 10.1007/s11097-013-9324-0

Minds as social institutions

Cristiano Castelfranchi

Abstract I will first discuss how social interactions organize, coordinate, and specialize as "artifacts," tools; how these tools are not only for coordination but for achieving something, for some outcome (goal/function), for a collective work. In particular, I will argue that these artifacts specify (predict and prescribe) the mental contents of the participants, both in terms of beliefs and acceptances and in terms of motives and plans. We have to revise the behavioristic view of "scripts" and "roles"; when we play a role we wear a "mind." No collective action would be possible without shared and/or ascribed mental contents. This is also very crucial for a central form of automatic mind-reading (mind ascription). Second, I will argue that often what really matters is the ascribed/prescribed, worn, mind not the real, private one. We have to play (like in the symbolic play) "as if" we had those mental contents. This social convention and mutual assumption makes the interaction work. The ascribed beliefs and goals are not necessarily explicitly there; they might be just *implicit* as inactive (we act just by routine and automatically) or *implicit* as potential. The coordination and social action works thanks to these "as if" (ascribed and pretended) minds, thanks to those conventional constructs. Our social minds for social interactions are coordination artifacts and social institutions.

Keywords Social frames · Theory of mind · Social institutions · Coordination · Extended mind

This work has been developed within the European Network for Social Intelligence (SINTELNET); a preliminary version has been presented at the Second Workshop of the European Network on Social Ontology (ENSO), Rome, Italy, September 21–23, 2011

"For sure I lie! The problem is that you believe me!" Shakespeare

C. Castelfranchi (✉)
ISTC- CNR, via San Martino della Battaglia 44, 00185 Rome, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it
URL: http://www.istc.cnr.it/group/goal

C. Castelfranchi
LUISS University, Viale Romania, 32, 00197 Rome, Italy

Published online: 30 May 2013

Minds *and* social artifacts

Social[1] interactions organize, coordinate, and specialize (we memorize them, apply them, and transmit them) as artifacts or tools.[2] They set up conventions, roles, scripts, procedures, or rules which allow *a drastic reduction of subjective uncertainty*, *of coordination costs*, *communication, and negotiation costs*, and support trust in interactions (Luhman 1991). They also reduce the *cognitive costs* of prediction, decision-making, and planning (Schank and Abelson 1977; Gilbert 1998); they simplify learning, and the transmission and capitalization of knowledge.

As Garfinkel (1963) correctly claimed, the first problem we face when entering a social interaction is simply to figure out "what game we're playing" and to know the rules, and the proper behavioral sequences. Social order and social structures of everyday life, emerge, stabilize, and work, thanks to our natural suspension of doubts, uncertainty, and worries. Our by-default assumption is that what is going to happen will be normal, as usual, and like it looks (Garfinkel's notion of "perceived normality"). This is how social interaction works: not every time created *de novo*, *ex nihilo*, and based on an unmanageable uncertainty; but based on learned (and prescribed) scripts and roles (Castelfranchi 2012b).

However, it is important to look at the mental side of these coordination processes and tools. They are not just multi-agent sequences of actions, or merely behavioral structures.[3] In this regard, Garfinkel's behavioristic (anti-psychological) claim is unduly restrictive:

(a) First of all, scripts, roles, and "games," are internalized in our mind, and become *mental* tools, for thinking: an intellectual artifact for recognition, anticipation, decision, and reasoning tasks.
(b) Moreover, and more importantly, they include the mental states of the social actors (roles); they are grounded on those mental states.

The ascription/recognition of intentions, beliefs, motives, etc. is not only fundamental for understanding (and dealing with) single, non-routine social actions, and non-conventionalized interactions; it is also indispensable for suitably playing the recurrent games of everyday activities.

[1] "Social" is a very broad notion. Especially "social mind" is an ambiguous expression that is frequently mixed up with "collective mind" (Castelfranchi 1997). Here, I will use the term in two senses: (a) to refer to "individual social actions" (regulated by an "individual social mind") towards other people as people (helping, exchanging, seducing, asking, competing, beating, etc.); (b) to refer to the individual mind as culturally and socially (by interacting with the others) shaped; filled with values, knowledge, scripts, etc. Of course, all this goes in the collective direction, just because we shape the minds of the members of our society with shared contents. The claim I defend in this paper is even stronger: not only the individual mind is socially shaped but it is constructed as an institutional coordination artifact for the individual and collective advantage. Granted this, I do not intend to question that there are also basic mental contents, concepts, and mechanisms which are not culturally constructed, but endogenous or due to individual experience and learning in the "natural" environment.

[2] Exactly like individual actions in our action repertoire, which are specialized and stored recipes and plan structures; tools to be used exactly like a hammer or a stair.

[3] The view that "scripts" organize behavioral sequences is the standard interpretation accepted in the social sciences and the one often given for granted also in cognitive science. However, see Apperly (2010) for an interoperation of scripts which is more similar to the one developed in this paper.

Indeed, without those mental contents (and their reciprocal ascription) there would be no game: the game could not be played, or its nature would be changed. The postulated minds of the role-players are part of the rules.

Our *minds-in-the-game too are cultural artifacts*; specialized for those social uses and ends, and carefully built within us.

Minds are "institutionalized" in (at least) two senses:

(a) they are shaped in order to use institutional artifacts (calendar, money, marriage, signature, etc.) in their internal activity (reasoning, deciding, evaluating, planning, and so on).
(b) they are themselves socially postulated and created institutional artifacts.

The second claim is the central objective and challenge of this paper: minds can be properly considered as social institutions. The idea is that a specific mind-set, the compound of beliefs, goals, and intentions that are ascribed to the interacting partners, is not only necessary to make sense of another agent's action but is often defining of a specific social encounter (see the gift-giving example § 6). Does it really matter whether the agent in gift-giving "really" has a benevolent intention or not? We don't think so. What matters is the staging of the gift-representation including the assumed mental attitudes.

However, we have to proceed step by step.

Coordination is for something

Though correct, the view that behavioral interactive patterns are coordination artifacts can be misleading, since it hides their multiple *goal nature*. Coordination is a means, and behavioral interactive patterns should be seen as "achievement" artifacts: they serve to both the individual and the cooperative achievement of some goal. They are multi-agent plans,[4] not just sequences; that is, tools evolved and specialized for particular effects/functions. Their result is not the mere performance of a given sequence and the coordination among the actors' behaviors (like in a dance!); the result is some specific outcome, related to the motivations of the participants. The game is not just a "game," an end in itself.

These artifacts are tools for a collective work. Consider the use of a big saw for two workers. Each worker cannot use it alone; its use requires a coordination of the workers' actions, but it also serves to achieve a specific result and requires specific skills and intentions in the users.

Similarly, scripts are immaterial tools—just made of mental representations—but they work exactly in the same way. They specify tasks and shares in the plan, and presuppose (and build) specific skills and abilities. Take for instance the "game" played among a surgeon, his assistant, the nurse, the anesthetist (and the patient) in the operating room. These are the true "cognitive technologies" much older than

[4] These plans are not necessarily (fully) represented into the minds of the involved agents; not necessarily understood and intended by them; but sometimes just emerging, self-organizing, or merely conceived and enforced by an organization or an authority, where the agents are mere executors (Castelfranchi and Conte 1991).

computer technologies; or better, these are computational technologies too, although running in our culturally modified (programmed) brains.

In sum, scripts and similar frames do not just assemble behaviors, but also minds (both beliefs and goals) and also that *part of mind* consisting of know-how (i.e., skills and recipes), of sense of competence, of self-trust and trust in the other's expertise, etc. All these are needed for cooperatively performing a task. Beyond coordination artifacts, scripts are *work-division and specialization artifacts*, *which shape and specialize our mind and body.*

Scripts of minds

Mentalizing the scripts that organize our behavior doesn't imply that they are fully conscious, subject to deliberation acts and decisions to conform to the norm. Once we have learned them, we follow them (as well as several kinds of norms) just by routine and in a rather automatic, executive, way. Nevertheless, a reductive behavioral version of scripts and roles is not credible: they necessarily specify mental assumptions to be activated step by step, as part of the procedure. Just like overt behaviors, mental processes, and representations can be automatic and routine based (Bargh et al. 2001). How do then scripts and roles specify the mental attitudes of the agents? Let's give some examples.

After we accept to sit at a table of a restaurant the waiter—without asking or saying nothing—may just hand us the menu: isn't the waiter necessarily assuming that we might be *willing* to read it (goal), that we *believe* that we will find the food available at the restaurant, and that we will *choose* and *order* (goals) something? Or is the waiter just performing a prescribed behavior in the script sequence and expecting our next move? If this were the case, why might he also add "Would you *like to know* what's the special food of the day? It is not on the menu, but I can tell you" or "Sorry, but we are out of ravioli." Why apologizing if not in relation to a possible *goal*/*desire* of us? And how could we just refuse the menu and saying "No thank you, I already *know* what I *want*."

To smoothly proceed in this scenario, we can and must ascribe even second- and third-order beliefs and goals to the actors: the waiter's belief that the client believes something, or his belief that the client wants something, etc.

There is a very central path for social life: *the automatic mind-ascription* from actions, material tools,[5] roles, scripts, social games, social rules, and norms (Castelfranchi 2012b; see also Apperly 2010). Indeed, when one adopts a "role," one automatically adopts the required goals, tools, acceptances, of that role, necessary for "playing that game"; while the others automatically ascribe such a mental endowment to X, and take it as a *common ground*. This is precisely one of the advantages of creating institutional (in the broad sense) *roles*: in order to endow them with a publicly known and prescribed mind, in terms of goals (motivations, values, mission, norms) and assumptions and competence (knowledge). Those mental assumptions are part of the "rules of the game": if you want to correctly play that game

[5] For example, if I see you taking a scissor, I assume that you intend to cut something; if I see you taking a hammer I assume that you intend to use it, to knock something; *the goal/function of the tools is* "*read*" *as the intention of the user.*

you have to *assume* so and so, have such and such a *goal*. In this sense, your mind is strictly part of the game.

Without the assumption of those mental attitudes the relevant games are no longer there: they change their nature (see below the "gift game," § 6).

Let us be a bit more precise about the goals ascribed to the actor, since mind or cognition does not include only the epistemic mental attitudes.

Some goals are expected (and in a sense "prescribed") but also assumed to be there spontaneously; ascribed to the subject as her own goals, recognized and presupposed (presumed), like the Client's desire to eat something when we see him entering a restaurant; other goals, like using real money to pay, are "prescribed" in a stronger sense: they have to be "adopted" from outside, they are duties (obligations or prohibitions) deriving from (and implementing) social (and legal) *norms* and conventions (Conte and Castelfranchi 1995).

When adopting a role, the actor "fills in" the "form" also with her own goals; in fact, she searches for and plays a given game and interprets a given script because it is useful for achieving her own goals too; scripts are *tools* to be appropriately chosen and used. But she also has to accept additional goals coming from the rules of the game/script. The ascription of the presupposed intention is not only ascription to the other of "desires" or "needs," but also of perceived "duties" based on *norms*—integral part of the social scripts—impinging on each role. X assumes that Y "has to order," "has to pay," "has to ask for payment," etc., and that he knows that. That is why (reason) he will formulate that intention.

The prescriptive (deontic) nature of scripts (and of their expectations on the others; Tummolini and Castelfranchi 2006; Tummolini et al. 2013) is very important: scripts and roles and rules are not just for reducing uncertainty, facilitating the interpretation of the other behavior (reading) and the prediction but for shaping, causing, regulating it, by becoming an internal/mental device, a proximal causal mechanism. This is what goes beyond Garfinkel's view.

Not only "scripts" and "games" and "roles": concepts, motives, emotions

One shouldn't reductively conceive as *cognitive coordination artifacts*[6] only scripts, roles, etc., which *explicitly* represent and organize multi-agent interactions. In fact, all our frames, schemes, and culturally constructed concepts are cognitive coordination artifacts, because they are shared and they work *as long as* they are shared. No doubt, schemes, and concepts are tools for individually thinking (and then acting); however, they are also cultural because:

(a) they are a way of conceptualizing the world that resulted to be useful and effective in a multi-agent practice, and therefore it is imitated, propagated, and handed down; that scheme captures relevant discriminations, effective in a specific context;

[6] Cognitive coordination artifacts coordinate our behaviors thanks to mental representations that appropriately control the conducts of the involved actors. However, not all the coordination artifacts are merely "cognitive": consider for instance guard-rails (which physically prevent us from straying, but also rely on the fact that we understand their function and "decide" accordingly); in the same way, think of corridors, revolving doors, etc.

(b) they allow a shared interpretation/construction of the common world, whose sharing is known and presupposed; a fundamental common ground for understanding each other, for predicting each other, and thus for coordinating our conducts[7];
(c) they are communicable (and built also through communication), and they allow that fundamental coordination device that is language (and vice versa);
(d) they allow the (individual and collective) "membership" and "identity" building, as well as the sharing of those schemes, rules, and games, of given "values," norms, which are "ours"; or better "we" consist in them and self-recognize in them.

In fact, this view of cognitive coordination artifacts doesn't only apply to the conceptual frames, schemas, and similar tools; it also holds for cultural *motivational* constructs, like social values or norms. Our motivations are culturally shaped too, and this allows social cooperation (also for the public good) or coordinated conflicts, or the division and specialization of labor.

For coordinating with each other, we do not only need shared categories (ontologies) and beliefs; clearly we need common or complementary (but mutually understood) motives and instrumental goals. It is goals (not beliefs) what directly regulate our actions.

Even emotions are culturally shaped and specialized; and there are rules about their appropriateness, meaning learning to discriminate between close emotions (like envy vs. jealousy, emulation, or admiration (Castelfranchi and Miceli 2009; Miceli and Castelfranchi 2007), learning when to appropriately feel one or the other, how and when to express it, which are the appropriate emotional responses; some sort of *emotional scripts and games*.

Thus, those emotional mental schemes are not only for interpreting events and reacting to them, but also for coordinating (affective based) social interaction in a culturally approved way.

Not only "scripts" and "games" and "roles": institutions and institutional outcomes

The mental attitudes of the involved agents are constitutive parts of any institution with its defining "artificial," "conventional," effects: those effects and the objects endowed with that "function" *are* the institution.

Consider for example the institution of the "signature" with its conventional outcomes. If the signature is effective, this means that, say, a contract is operative in the conduct of the involved subjects; it will have its effects on/through the behavior of the agents. The partners of the contract, the bank, the court (in case of insolvency), the employees, etc. will act *accordingly*, coherently with what happened; and while doing so they make it really happen (Tummolini and Castelfranchi 2006).

However, to be more precise the involved agents act on the ground of what they *believe* and—on such a basis—of what they *intend*. The first and main effect of the "signature" is obviously on the mind of the involved agents; they are supposed and requested to act *in agreement* with what they *know*.

[7] That's why we negotiate the meaning of words.

Thus the first "conventional" effect of a "signature" is that everybody (concerned) *knows* that the document has been "signed" and knows what that means; that X (the signer) was willing to sign and knows that the others (will) know that he signed; X believes that the other will be *willing* to act accordingly; Y knows that X believes so; X and Y believe that they will consider each other as "contract partners" or "spouses," and they know the conventional meaning of that; they know the *obligations* and *rights* following from that signature and intend to comply with them; etc. The institution of the "signature" *is* such complex distributed change of minds, functional to producing the appropriate behaviors. Not only those minds are among the conventional and definitional effects of the signature, but they *are* a "convention" (Lewis 1969; Tummolini et al. 2013). They are not simply consequences of an institution they *are* an institution.

Not only ascribed beliefs, goals, mental attitudes are "institutional" objects, also the scripts, roles, social games discussed above are "institutions," with conventional ground and conventional effects.[8]

Ascribed and prescribed minds

Social coordination and the games we play aren't just based on specific minds (specific goals, assumptions, etc.), but they allow the prediction and ascription of minds, and they are tools *for* that too. Those who receive a gift (see § 6.2), if they perceive it as a "gift," are ascribing (in an rather automatic way, and without a complex inferential reasoning) specific motivations to the giver (who is in fact expecting such ascription/recognition); like the waiter who by default ascribes certain desires and duties to the client. More than this: those mental states/contents are not only ascribed, they are *prescribed*.

The script you are interpreting, the role you are enacting, the game you have to play prescribe you the mind that you *should* have; and I'm *entitled* (not only cognitively prepared) to expect and assume as certain (without further inquiry or negotiation[9]) that, since you are playing that game, you are wearing that mind.

Moreover, as we have claimed, you also have to possess specific know-how, abilities, skills needed for playing with me that game. I "presuppose" all that; it is on that that I rely on and trust: this is your trustworthiness (Castelfranchi and Falcone 2010).

Let's suppose that a guy goes to a newsagent without knowing the game of "buying/selling newspapers"; for example, he doesn't go there as a "client" (with the related goals, beliefs, means) but he intends to take a newspaper without paying anything (like he has seen other people doing at the exit of the metro-station); or consider a guy who brings back the newspaper and asks back for the money; or a guy who asks for a particular journal but claims the free book of another newspaper; or a guy that asks to the seller to sell the newspapers by weight. In all such cases, the guy

[8] In this view an institution is a collective construct, a coordination artifact, based on conventions, mutual expectations, and coordinated and relied practices, with artificial/conventional "count as" actions, objects, effects.

[9] That's why there is a fundamental reduction of the "negotiation costs".

is ignoring the rules of the game of "buying newspapers," which—on the contrary—the newsagent automatically ascribes to him. Is that guy wearing the right role, with its mental constituents? Does he reason and act following the appropriate scripts? Has the (strictly cultural) coordination artifact that organizes their interaction been internalized?

Not only the expected behavior, but—first of all—the mechanism which controls the behavior is a cultural product, an institutional artifact.

As we have said, to have those mental states/contents is a social "convention" or better is an essential part of the convention; not less expected and implicitly prescribed than the expected behaviors. And our behaviors will be read as expressions and signals of those mental states; knowing that, we can "communicate" them (we confirm that we have them or make clear which game we intend to play or are playing). Actually this is not just a problem of mind-reading, of understanding which are the mental attitudes in the other's mind (and in our own mind); the problem is to "*assume*" (and adopt) them: that is, we have to act consistently, we have to regulate our behavior on that base, "as if" they were there.[10]

X depends on Y as for the success of their interaction (coordination). He is relying on Y's conduct and mind, and risks/bets on that. Thus, not only X *believes*, forecasts that Y believes/wants/does something but X *wants* so; and Y would wrong X if he were not to conform to such an expectation.

This normativity is also internalized: we prescribe those mental ingredients—motivations, beliefs, or assumptions, moods—to ourselves; and we might be disappointed and blame ourselves if we do not feel, think, or desire what would be "appropriate" to feel or think in that role in those circumstances

Mind as social artifacts: worn minds

In sum, we cooperate in a *mise en scène* (Goffman 1959) and we have also to wear the appropriate minds, not only the masks and costumes of the comedy (script): the "persona" becomes a person.

Thus, by simply looking at your mask or costume I "know" your mind; I'm entitled and even requested to "assume" that you "assume" P & Q and that you "want" R. And on such a basis I'm supposed to inter-act with you, and to play my role/character.

It is false that—in order to have assumptions about the other's mental attitudes—we need either to mirror her/him, or to simulate what we (would) have in mind in similar circumstances, or to abduct by reasoning the possible intentions and beliefs. The other's mind (and not just behavior) is *inscribed and prescribed* in our scripts, roles, games, norms, tools, etc. (Castelfranchi 2012b).

This is what we learn since the time of symbolic play: when we start to pretend that we have certain goals and beliefs (as a doctor or as a king) and that the other has certain goals and beliefs (as a patient or as a princess).

[10] This is our definition of "acceptance" (Tummolini and Castelfranchi 2006).

"Pretending" is this: it is *always* pretending a mind. When I pretend that this piece of wood is a knife I'm in fact pretending that I (and you) *believe* that it's a knife; and that's why I *use* it as a knife and it "*counts as*" a knife in our interaction.

From functions to goals

Playing our roles and social games "mentally" doesn't mean that we have a complete representation and understanding of them, and that we intend all their functions; and even less that we jointly intentionally build our social structures and scripts: they self-organize, emerge, but with a (necessary though partial) mediation of our mental representations and processes. We just play them (frequently just in a passive, unaware way); like actors, which neither necessarily "understand" the message of the play nor pursue the ends of the author.

Our social "functions" (the functions of our roles, games, scripts) are not necessarily our "intentions": like in tool-use, *the goal of the tool and action is not necessarily the mental goal of the agent*, even when the behavior is not a mere routine.

Social functions can impinge upon and be played through personal intentional actions without being intended (Castelfranchi 2000, 2001). Sometimes we do not realize that the agent—while doing/saying something—is implementing a social function. We can see his goal as a "personal" one; for instance, a father who reproaches his son for making a mess of his room may do so for personal reasons just for educating the future good citizen, or for both motives.

Conversely, sometimes we consider Y's action in an impersonal perspective; especially when Y is playing an institutional and formal "role": not the role of "father," but the role of "policeman" or "professor." In those cases, we believe in a partial internalization and awareness of the goal-structure specific to that role (goal "adoption"; Conte and Castelfranchi 1995).

The agent explicitly represents and pursues part of his role plan, of his "mission." Or at least—when the agent's behavior is highly routinized—we tend to ascribe him beliefs and goals that actually are just "implicit," potential, in the background, not currently explicit and available. Notice that we can even collaborate with or oppose to Y and its role/action because we share or reject precisely some goal belonging to his social role/function (e.g., a soldier and a policeman); and we interact with Y "as if" he were a conscious supporter of that goal (whereas he might be just an executor of sub-goals and implementing practices). Policemen or soldiers may be insulted or attacked by demonstrators because they are supposed to do something unfair or antidemocratic as if they were personally willing to do so and not just obeying to someone else's orders, sometimes without understanding the political or legal aspects of what they are doing.

What matters in all these cases is that we regulate our behavior on the basis of those "pursued" goals (more or less cognitively represented or even conscious). Institutions exploit our minds and our goal-directed control of behavior while shaping part of our motives (other higher-level goals we pursue are "alienated" and viewed as "powers" of the institution).

In sum, "*pursuing*" *goals is not necessarily a* "*mental*" *operation*; and thus also "ascribing" a goal to Y is not necessarily a real "mental" ascription. It may be "as if" Y were intending the goals he is pursuing or trying to achieve (Castelfranchi 2012b).

"As if" minds and how they "count as"

However, a theory of social games capable of explaining how mental states are automatically ascribed is not enough. We also need to make a more radical view: people involved in social interaction "*assume*" *and deal with*[11] *the goals implied by the others*' *behavior* "*as if*" *they were mentally represented*.

Human (socialized) minds are "as if" minds, "as if" artifacts. In two senses:

(a) On the one side, they must be able to have the mental construct of an "as if" object and effect, of merely conventional (institutional) entities which "count as" (Searle 1995); and to coherently behave on such (shared) conventional basis, thus making the social effects of those "fictions" true.

(b) On the other side, minds are themselves "as if" constructs, to be assumed and used as such. They are "institutional" artifacts, and we have to ascribe them "as if" they were in a given prescribed and expected way. What matters is not necessarily what is in a mind, but what we have to assume; and we have to behave "as if" that was the mental content. The "real" subjective content doesn't matter very much; what matters is the "conventional" content.

Let's focus on the second aspect in order to explain in which sense the ascribed and feigned content of the player's mind can be actually "implicit," not really "in his mind."

"Implicit"

"Implicit," referred to mental representations, is an ambiguous term, with several uses or meanings that are often conflated. In order to avoid misunderstandings, below we try to disentangle some different senses in which a mental content can be said to be "implicit."[12]

1. Active but "unconscious" representations

   As we have suggested, beliefs, and goals are not necessarily "conscious."[13] On the one side, *ascribing a belief or a goal to another is a belief*; and it might be unconscious. On the other side, the ascribed belief may actually be unconscious in the other's mind; he really has the ascribed belief, and it is active and taken into account, but the subject is not conscious of it and of its role. Thus, given that "unconscious" already clearly qualifies this kind of representations, the term "implicit" should be avoided.

2. "Implicit" as non-active

   Beliefs (and goals) that are not activated remain in the background of our cognitive processes and might be primed in case of surprise or in order to face novel problems or to answer specific questions (Lorini and Castelfranchi 2007).

[11] This is our notion of "acceptance": not only assuming but acting accordingly.

[12] We neglect here to discuss some important uses of "implicit" knowledge: on the one hand "implicit" as "tacit" knowledge; that is, knowledge that is just procedural or sensory-motor, and thus not fully verbalized and difficult to communicate by words (Polanyi 1958); on the other hand, knowledge that we are not aware of: knowledge that we don't know to have lacking appropriate meta-representation.

[13] "The unconscious is not identifiably less flexible, complex, controlling, deliberative, or action oriented than its counterpart" (Bargh and Morsella 2008), and this holds also for complex social interaction.

For example, whenever we automatically stop at a red light, we "know" that we are prohibited from crossing; however, although following the norm, we are not considering it explicitly, not acting (consciously or unconsciously) on such beliefs. They are inactive in our memory, we just follow a routine. Only in case of problems we will activate those beliefs; for example if somebody encourages us to cross (because the traffic light is broken or we hear an ambulance hooter behind us; etc.) we activate the relevant belief and consider the prohibition for a higher-level deliberation: should we violate the norm or not?

3. "Implicit" as just "potential"

Another kind of "implicit" mental attitudes is however less considered: the potential ones.

By "*implicit*" *belief* one may also mean a belief that is not present in any "database" (short-term or working memory, or long-term one) but is only *potentially* known by the subject because it can be derived from other beliefs one actually endorses.

What does pragmatically mean that Mary "knows" that p? It might mean—for example—that she is able to correctly respond to a specific question about p. For example, if I ask Mary "Is Athens the capital of Greece?" she will correctly answer "Yes, sure!". She "knows" that. However she is equally able to answer other strange questions about Athens and Greece: "Is Athens the capital of Egypt?" "No!" "Is Rome the capital of Greece?" "No!" Therefore, she "knows" that too, that is, she "knows" that "Rome is not the capital of Greece" and "Athens is not the capital of Egypt"; and so on… Shall we say that Mary's memory include all the non-capitals of each country? Not at all. While the former information is in some memory file, written in her brain and waiting to be retrieved, the second piece of knowledge is not written in her brain at all; it will be there only while computing the answer. However, this is for sure a kind of knowledge, something that Mary "knows." What Mary has in her memory is that "Athens is the capital of Greece," "Countries have only one capital" and "A capital is capital of just one country"; from these premises she can derive all the countries Athens is not the capital. She "knows" all the non-capitals of the world because she is able to compute or derive them, if needed.

The general cognitive principles underlying this form of knowledge are the following:

(a) Differently from the classic assumption of formal logic (i.e. epistemic closure) *we do not explicitly know whatever logically follows from what we know*; we do not derive all the logical implications of what we know; for obvious reasons of bounded cognitive resources and of irrelevance.

(b) *We derive inferences only if and when needed* for specific tasks and uses, like question-answering, problem-solving, predictions, etc.

This kind of unwritten, *generative* knowledge is not only "propositional." It can also be of the sensory-motor kind. When I walk down a staircase I "know" that it will support me, I rely on that, and I "assume" so since it is an implicit presupposition of my act of walking.

The same also holds in social interaction and mind coordination. For

instance, I know that my colleague has a salary, that he owns a house where he lives even if I never had the occasion to see that, to think or to discuss about that. If you are a seller and I have to pay, when I give you the money, I assume that you "know" that that is "money," and that this belief is active in your mind and used for recognition of money; but I also believe that you know that this money has an exchange value, that one can circulate it and that we also "know" that after having given you the money "I have paid" but also that "I have less money." Actually these beliefs can remain just implicit; I can use the "payment" game and script just automatically and routinely; as a behavioral sequence; but they are there: if needed, they will be derived or activated, and everything works "as if" they actually were explicitly there (Castelfranchi 2012b).

Moreover, we not only have "implicit as potential" epistemic attitudes (knowledge, beliefs) but also "implicit as potential" goals. For example when we use a hammer, besides (implicitly) assuming that our hammer will resist the collision impact (that its head and the handle won't break), we also (implicitly) *want* so. This goal is not formulated, or represented in any way (unless we have previously had a surprising negative experience). Indeed, there is an open chain of possible goals: e.g., that the hammer's head doesn't slip away, that the handle is inflexible, non-slippery, and graspable, that the head is flat on the impact side, etc. etc. *All the necessary conditions for action execution and its success are* "*potential*" *beliefs and goals while we perform that action*.

4. Simulated beliefs

As we have said, one may act "as if" he/she has certain beliefs and goals, without necessarily formulating such beliefs, goals, or expectations in his/her mind. They can remain merely "potential" in the subject's mind.

This doesn't necessarily imply that one has a different explicit belief or goal; however, even this is possible. One may *personally* have beliefs that contrast with the assumptions implied by one's own behavior. Suppose X is a policeman who is personally persuaded that Y is innocent, that Y would never be able to commit a crime like the one he is accused of. However, according to the current circumstantial evidence, he is compelled to treat Y as a criminal. X shouldn't even manifest his doubts (except as a contribution to the investigation in the right context). Officially, he believes (or better "accepts") that Y is guilty.

In the same fashion (but with a less strong feeling of "obligation"), if X is a waiter in a restaurant and there is a client that X believes to be unable to pay the bill, he cannot refuse to serve him: he has to act as if he assumes that the client will pay. He cannot ask the client to show the money in advance (unless he chooses to violate the normal script and its social norms). Perhaps, he could dismiss the client for reasons of alcohol, hygiene, or indecency but in such case it would be the client who violates the "rules of the game" of a "restaurant."

In sum, the ascribed mental attitudes might be unconscious or just "implicit": either, not-activated at all at that moment, that is not really included in the agent's mental process, or just potential, or even simulated, that is only officially endowed. However, for our social games such differences don't matter; these cognitive-based

behavioral interactions are made *less costly from the mental-computational point of view*. The agent's "implicit" beliefs are part of the game, and part of a *socially effective mind*: the others do not spend time in ascertaining whether those beliefs are already explicit and activated in the agent's mind or they are just implicit or silent.

Signaling

One of the reasons why these "as if" minds work and are effective, is that our behavior (in a given context) *signals those beliefs or goals even when they are just implicit*. One doesn't really "have them in mind" (she acts by a routine or by a partial explicit representation of the presupposed beliefs and goals); like when we ascribe to Y, the driver before us, the intention to warn us because we see his turn signal, while he is just automatically and unconsciously carrying out the usual motor routine for turning. Sometimes Y forgets or doesn't realize to have the turn signals on and doesn't change his direction, and we do think than he has changed his mind or that has just forgotten, that is, that he hasn't the intention to turn: we *revise* our (implicit) belief about Y's intention.

All this doesn't matter: we can and must ascribe such mental attitudes to the other, and interact with him "as if" he had them. It works.[14]

The mind postulated beyond our brain

An interesting consequence of our "potential" representations (knowledge and goals) is the following one: *our mind is not just in our brain*. In other words, our mind is "exuberant" similarly to what has been argued by proponents of the Extended Mind thesis (Clark and Chalmers 1998; Clark 2008) but also in a different sense:

1. What our mind "knows" and "wants/wishes" (and the causal effectiveness of its beliefs and goals) goes beyond what is materially and directly present in it (whether represented in a declarative or in a procedural form, in a propositional or sensory-motor one).
2. Our "mind," as a cooperative social construct and coordination artifact, goes beyond what is neurally represented and processed. People deal with me by ascribing me mental contents that I do not actually have; thus in fact, "to all intents and purposes," I pragmatically "know/want" beyond what I have in mind; or better: *what I have in my mind goes beyond what I have in my brain*.

At least "potentially" and socially (institutionally), we know much more than what is stored in our internal or external (working or long term) memory. Part of our mind is "immaterial," or better it is materialized externally: in the collective assumptions and practices. This seems another way of its being "environmentally embedded" (Clark and Chalmers 1998).

This view is in part close to the philosophy we can find in Gallagher and Crisafi' extension of the "extended mind" view (Gallagher and Crisafi 2009), that they apply

[14] At worst, he will be obliged to activate or derive those beliefs and goals that we assume are in his mind and that we react to.

to legal systems, museums, etc., including the "externalization" but also some "internalization" of cognitive processes/tools.[15] "We use these institutions instrumentally to do further cognitive work, for example, to solve problems or to control behavior..... That part of the cognitive process that in [other] case[s] involves cognitive schemas that run on [our] brain, in [these] case[s] is *replaced by cognitive schemas that are processed according to the rules* of a legal institution." However, exactly the same reasoning and description should already be applied to "scripts," roles, etc. before complex and formal "institutions."[16] Scripts exactly are *solutions* for complex problems (in this case multi-agent coordination and planning problems) that somebody has elaborated (by individual and group reasoning, experimenting, learning, sharing mental states, tacit negotiation and agreement), and that have been *stored as a common good*: a heritage of cognitive work already done. They give us a semifinished product to be adjusted to the specific circumstances—without the need for finding again a solution but just exploiting/applying the previous knowledge and cognitive work—with an impressive reduction not only of uncertainty but of cognitive work and of negotiation costs.[17]

Effectiveness of those postulated attitudes

Even if some "goals" and "knowledge" of mine are only potential, they can still have effects in the world, "as if" they were materially there, since the others (implicitly or explicitly) ascribe them to me. They interact with me "as if" I had them, that is, they "practically" assume that I have them by acting according to such assumptions.

Suppose that I interact with Mary knowing that she "knows" that Athens is not the capital of Italy, or that 2×2 is not equal to 5, or that she is not my daughter, or that we are not in New York City etc. When I say to her (while talking about museums) "In Salamanca, there is a fantastic museum of Art Nouveau & Deco," I assume that she assumes that we are not in Salamanca; is this knowledge really in her mind, or is there just the knowledge that "we are in Roma"?

I address my client knowing that he does not "want" to pay more than the actual price of the ticket; or that he wants that the medical doctor visiting him be a real doctor; and so on.

The intentional stance

Are such ascribed beliefs, goals, etc. actually there, in the other's brain/mind? Do they really exist? Or are they just external constructions and descriptions?

What matters is that it is "as if" they were there! *What matters is the others*' "*intentional stance*" towards the individual, which makes them act in a consistent manner on such a basis (Dennett 1987). Sometimes, the "intentional stance" too

[15] To me this is Vigotsky's idea that social tools, artifacts for interaction and collective work (like language but non only language) are internalized as tools of mental activity.

[16] We are also using a more basic Searlean notion of "institution" and "institutional"; not only laws, courts, museums, etc. but also a "signature", an "arrest/detention". Our thesis is more basic and radical.

[17] The disadvantages of this cognitive and exchange economy is that these "solutions" are rather conservative; in a sense, are *pre*-judices.

becomes an institutional convention (fundamental—for example—for the notion of "guilt"; see below).

Even if one is a realist about mental states,[18] it can be conceded that often we pretend and collectively stage also what is true; represented, depicted, played doesn't mean false!

Indeed, sometimes mental contents/attitudes are there "immaterially," just as a social convention, but they (the assumption that they are in fact there) do all the work we need.

Things that do not "exist," if assumed and regulating material actions are effective, like (the idea of) god.[19] Causally ascribed effects are the only evidence that is needed; actually, the effects of our own actions. We just ascribe to those unobservable entities some of our own alienated powers.

The "institutionalization" of mentally defined acts

Let us give few examples of social acts strictly *defined on the basis of specific mental attitudes*, and how this presupposition becomes "conventional."

The needed mind for an altruistic act

A given act is altruistic only with reference to the underlying intentions. "Altruistic" is a *subjective* notion; it just depends on the mental representations (in particular the motivational ones) ascribed to the agent and underlying his act; it is not—if applied to cognitive agents—an objective and behavioral notion. It is not sufficient that X's behavior be (not accidentally but functionally and regularly) beneficial to Y and costly for X. It is even not enough that this behavior be *intentionally* in favor of or beneficial to Y; psychological altruism is a matter of final motives, of the ends of the act.

In our view (Lorini et al. 2005), it is impossible to solve the problem of the existence or not of "true" altruistic actions and people without making clear two issues:

(a) Being an "autonomous" agent, endogenously motivated and regulated by one's own goals (like in purposive systems) is not the same as being "selfish." What common sense means by "selfish" or "egoist," is not that one is driven by "one's own" internal motives and choices; and "altruist" does not mean that one is

[18] Indeed I am a realist about mental attitudes and believe in the distinction between lies and sincerity, and in the possibility that what one believes can be false or wrong. It is obvious for instance that the two notions do not coincide. We might cross them and have: (1) lies that are true, since the speaker doesn't believes so; (2) lies which are really false (the speaker's belief that they are false is true); (3) false things which are sincere (mistakes, incompetence, no lies); (4) things which are sincere and true.

[19] See for example Doran (1998). Our minds are built for superstition, for believing beyond material evidence, also because we need mind reading/ascription and on such a basis coordination and cooperation. According to dictionary definition, "superstition" in fact means "a notion maintained despite evidence to the contrary" [or better I would say: "despite the lack of empirical direct evidence"], "a belief or practice resulting from ignorance, fear of the unknown, trust in magic or chance, or a false conception of causation". Even worst: by believing and acting accordingly we make those immaterial/non-existent things effective.

hetero-regulated. What these terms refer to is the nature and origin of the regulating goals; but those goals are always the agent's own goals and preferences. Psychology should be able to grasp and model such a distinction (Castelfranchi 2012a).

(b) As very clearly explained by Seneca,[20] it is crucial to distinguish between expected positive results (the prediction of positive outcomes of my action) and what "motivates" my action. *Not all the expected positive outcomes are motivating for me*; in other words, it is false that I act "in order to" achieve them, just because I predict them; for motivating my action they should be *necessary* and *sufficient* conditions for my decision to act.

Without this distinction and sophisticated modeling of motivations and goal processing, we have just to be satisfied by "pseudo-altruism" (Batson 1991) where the expected (at least internal) positive rewards of one's behavior are unduly identified with one's motivations.

Seneca's solution is very simple and intuitive (although we still do not have the corresponding psychological model!). Aristotle's view is more realistic and sophisticated.

In my interpretation of this famous passage,[21] we are ascribed a motivational *ambiguity* and conflict in social interaction, about ourselves and our view of the other's mind and the other's trustworthiness; thus even if we would like that the interaction be really altruistic, without any "reciprocation" (that we however "expect," like in Seneca's analysis), we oscillate between pure altruism and more self-interested motives.

Precisely for this reason, in my view, because of our underlying ambiguity and ambivalence when we carry out an altruistic act, make a gift, or forgive somebody, those kinds of acts are constructed in a conventional way, are institutionalized, and we "play" them. In such a way our "mind" is clear: predefined, predictable, conventionally assumed (in both senses). We assume that and behave "as if" we had altruistic motivations; sometimes this is true, sometimes it is just pretended but positively intended and defended, and thus it "counts as."

Consider for instance what happens when A addresses an offensive sentence to B: Socially speaking, what is "insulting" is more the meta-intention to offend than the specific content of the insult (usually false and not believed by the speaker!) and the communicated intention of the speech act, which is false. In the same vein, what is kind is your attitude, your intention to be kind, to show your care for me, not what you say, your simulated intention.

[20] "But," says our adversary, "you yourself only practise virtue because you hope to obtain some pleasure from it." In the first place, even though virtue may afford us pleasure, still we do not seek after her on that account: for she does not bestow this, but bestows this to boot, nor is this the end for which she labors, but her labor wins this also, although it be directed to another end. As in a tilled-field, when ploughed for corn, some flowers are found amongst it, and yet, though these posies may charm the eye, all this labor was not spent in order to produce them—the man who sowed the field had another object in view, he gained this over and above it—so pleasure is not the reward or the cause of virtue, but comes in addition to it; nor do we choose virtue because she gives us pleasure, but she gives us pleasure also if we choose her." ("Of a Happy Life," translated by Aubrey Stewart from the Bohn's Classical Library Edition of *L. Annaeus Seneca*, *Minor Dialogs Together with the Dialog "On Clemency"*; George Bell and Sons, London, 1900).

[21] I do not claim any philological accuracy. Here what matters are ideas, claims, not their author or historical origin.

The "count as" can work even when it is notoriously false. The statue of the founder of Harvard University (John Harvard) is not his portrait but the portrait of an obscure student chosen by the sculptor. Does this matter? Only if you naively use the statue to know which were John Harvard's physical features. But as a homage to and memory of John Harvard, as a "monument" with this function, it works perfectly. How many portraits of ancient philosophers are mere imagination, the creation of a shared icon?

Gift

By definition a "gift" in our culture is not *for* exchanging something and conditional to that, it is not a means for exchange, and is not for paying off a debt. The giver, X, has nothing to exact and the receiver, Y, is not in "debt," committed/obliged to give something back; except for acknowledging and appreciating the gift. The only debt Y has is one of "gratitude" (Castelfranchi 2012b). Y is perhaps expected to reciprocate but he is not strictly required to do so, since he has not taken any commitment (there was no promise, no deal or contract); the expectation is due to moral or politeness rules. But, what is important is that, even if Y is expected to reciprocate (and X expects so), X does (is assumed to) *not act* "*in order*" *to receive* the reciprocation; his end/motive is not the possible expected return from Y (see above).

In order to *make* a real "gift"—recognized and treated as such—X is *supposed to* be (and has to act "as if" being) motivated by the good of Y, by a benevolent attitude, not by a calculated selfish advantage: it is a selfless act. We cooperate in the *mise-en-scène* of that special situation, of that *institutional act and object* (the gift), that *necessarily implies assumptions about the* "*motives*" *and* "*beliefs*" of X and Y.

Do we know the "rules" of the gift-game? Including the assumed mental attitudes: altruism and gratitude? Without those assumptions our gift-interaction becomes just an exchange, the payment of a favor, like it might behaviorally and actually be. It remains a gift if we mutually presuppose the appropriate motives and beliefs and reactions.

Forgiving

The act of "forgiving" may imply various kinds of attitudes and behaviors, and these too can be culturally shaped. Let's distinguish at least between two kinds of forgiving, one closer to the Christian culture, the other related to the Jewish tradition.

The former kind is an "intrapsychic" form of forgiving, which may be independent of public declarations, and may occur regardless of Y's repentance (sincere or not) or Y's expiation or punishment. X may either communicate or not to Y that he forgives her: what really matters is that X, in his own mind, makes Y free from any "debt" towards him, and makes himself free from any resentment and expectation and brooding about settling accounts. This cancelation of Y's debt is a free gift, independent of external circumstances (such as X's actual impossibility to retaliate or his fear of Y's or others' reactions). X is in no way required and obliged to forgive. And forgiveness doesn't necessarily imply the restoration of the relationship with the wrongdoer.

A different kind of forgiveness is the social institution and interpersonal act in the Jewish tradition. Here, forgiveness must be signaled and communicated. It is a public act. The wrongdoer's repentance is due, and should be manifested, and X's forgiveness becomes "due" after Y's repentance, and should also be publicly communicated. In this way forgiving becomes a *ritual*, not just a private issue; a public institution with recognized social functions: the *reconciliation* between the involved individuals and their groups. Both Y and X have to conform to the social norm and expectation and have to signal this conformity with their behavior. At that point X's (and even Y's) sincerity doesn't matter so much (Miceli and Castelfranchi 2011): Y has declared her repentance, and her act "counts as" an actual repentance; X has performed a forgiving act, and therefore he is "officially" without resentment and desire of personal revenge. Also, the subsequent reconciliation is official and operative: X and Y and all their community members will act in such a way, coherently with these assumptions, norms, and ascribed minds.

Notice that—considering forgiveness in the previous strictly psychological sense—it is a nonsense to *prescribe* it, it's absurd, like prescribing or swearing love (as it happens in the marriage ceremony). But it is not absurd from an institutional point of view (like for money, where the material substrate doesn't matter any longer). Committing ourselves to love each other forever is impossible but has a perfect symbolic, ritual, official effect, and meaning. It doesn't really depend on the partners' will, but their "as if" behavior depends on their will and commitments ("as if" they would love each other forever) and—if coherent with such a common assumption—it will "count as," it will have the same effect (for society, not psychologically for the spouses).

Responsibility and free will

Finally, "free will" might be a crucial "mental" institution: the institutional postulation and construction of a mind-set.

Actually, socially speaking, it is not so relevant whether free will and actual responsibility exist or not. What is more crucial is our shared representation that "what happens is not strictly necessary, determined, and *it might have been different*." This collective counterfactual assumption is the necessary condition for the attribution of "guilt" and differentiated sanctions.

"*You might have done differently*"[22] is the representation of an "unreality" which is institutionally assumed as a necessary comparative/evaluative parameter of reality.[23]

Notice that we are "guilty" because of something that we didn't think, because it wasn't in our mind: the lack of a given belief makes us guilty; or because a given belief and expected outcome (the other's harm) was not weighted enough to change our decision.

Moreover, a mere fiction of free will and responsibility works. In fact, even admitting that all events are the result of a deterministic universe, this doesn't preclude our learning, anticipation and reasoning, and even decision-making, in

[22] Also because there is "evidence" that different (!) people in similar (!) circumstances behaved differently.
[23] But in our view this also is a natural, inborn, way of reasoning in human beings; necessary also for causal thinking, which implies the counterfactual idea: If A had not happened, B would have not happened.

another sense—that is, our memory, evocation, and anticipation of possible outcomes (including punishments) with their learned weights, and on such a basis an automatic computation and a result, a winning exit. Even if A was not able to decide differently from how he has in fact decided, his decision will take into account the convention of "responsibility," that there is accountability, surveillance, violation, and sanctions. Social rules and conventions will still influence the agent's conduct even if the agent's computations have a deterministic result, with no alternatives. This is the efficacy of the staging of "responsibility"; it is *really* influencing and determining the conduct of the agents. Minds might be deterministic but not unchangeable, predetermined: minds are predetermined at time t', but they can be determined, influenced, at time t, before t'.

It is clear that the crisis of the idea of free will would have (in our culture) very serious consequences on the personal and social moral sense (Holton 2009) and on our (even legal) conception of responsibility. The institution of the "free will" ascription works quite well, and serves for regulating individual "responsible" behavior, and social evaluations and sanctions.

The social and subjective functions played by the idea and convention of "free will" are effective and valid, beyond its empirical pre- or extra-institutional existence.[24]

In sum, the institutional ascription of mental attitudes works, even for unconscious, or implicit or just potential or feigned contents; if even mere superstitions work (play their socio-cultural role) imagine a (partially) true "superstition."

Mental attitudes as illusions

I'm not saying that beliefs, intentions, mental reasons of behavior are "illusions," common sense constructs for interpreting, predicting, and "understanding" and "explaining"[25]; and that only the "intentional stance" is what matters while the ascribed contents of minds are just conventional postulations, not objectively existing.[26] No; I basically have a naturalistic approach to mental attitudes. They are "real," material, to me. Otherwise, I couldn't distinguish between when they are conscious and when they are not, or active vs. inactive, or explicit vs. just potential; or when they are "true" and thus "sincerely" signaled, or simulated.

[24] Sooner or later we will even "discover" that money does not "exist", they are not "material" (like the clothes of the emperor!). Under this perspective, I don't see differences between money and mind. In both cases, we can from external signs just infer, attribute value to them, and their power strongly derives from such ascription and from our coordinately behaving on such a shared assumption.

[25] To me—contrary to the phenomenological view—they are just one and the same thing; human "understanding" is in term of "explaining": Aristotle's position that for humans it's not enough to know "that", we need to know "why".

[26] Consider for example the very clear claims of Scott Churchill's *Reasons*, *Causes*, *and Motives*: *Psychology's Illusive Explanations of Behavior* (1991): "The efforts of psychologists as well as laypersons to identify causes and motives (and thereby explanations) of behavior is examined from an existential-phenomenological perspective. ..... The traditional concepts of 'conditions,' 'causes,' and 'motives' are critiqued and alternative notions such as 'meaning' and 'project' are drawn from the literature of phenomenology as a basis for understanding rather than explaining human behavior. Psychological explanation is presented as a system of discourse that has its own psychological 'motivation'. "

They are implemented in our neural structures and processes and in our externalized interactions and cognitive artifacts. Our beliefs and goals about the others' minds are dynamic structures in our brains, and the same is true for the ascribed beliefs and goals in the others' brains. However, we also have goals and beliefs that have to be adopted and worn while playing our roles, scripts, and games; and many of them remain just unconscious or even inactive in our habit-based behaviors. Some of them are not really "written" at all in our brain (in fact, they "might" be); they are just potential, but presupposed by our practices and ascribed to us, and they work: that is, they have the interactive and collective effects that they should produce and that reproduce their conventional attribution. Just this in a sense are "illusions," but illusions that become effective, have real effects on the world thanks to their assumption and the behavioral consequences, like the gods regulating the life of a given population,[27] or like the value of money, no longer gold, or silver but just paper or just bits in a bank computer. The clothes of the emperor (see note 23) actually exist; if everybody fully behave "as if" they would be there, they actually have the effects that they would have.

Concluding remarks: The internal/external ontology of institutional artifacts

The basic argument of the paper might be synthesized as follows:

1. A sociological view (inspired by Goffman, Garfinkel): social interaction is based on scripts that we give for granted, the assumption that everything will proceed according to the script (the game we play; the perceived normality) is a prerequisite for socially coordinated action.
2. The cognitive view: scripts however are not limited to "ordinary" behaviors but also to "ordinary" minds. The scripted-mind enables a form an automatic mind-ascription (interpretation) that doesn't require a complex form of mind-reading.
3. Combining the views: like in the sociological view, the scripted mind is taken for granted, it is generally assumed to be there, it is crucial for orchestrating social interaction, and it is also "prescribed."
4. The scripted mind that is assumed in social interaction might not correspond to the private minds of the participant. Not all components of the assumed scripted mind-set are necessary explicitly represented because they can be implicit or just simulated.
5. However, this does not impede that social interaction can succeed because in many contexts the mind-set that is collectively assumed to be there (as-if mind) is enough to enable smooth social interaction.

In sum, the central claim is that: *mind is an institution* not only vaguely and as culturally shaped and filled, like any artificial tool designed or prepared for its functions; but in a more radical and strict sense. First, a given mind-set is and has

[27] "Suppose a population of agents with some god that guards promises, pacts, and social taboos, and punishes the violators; for social trust-based relations to work, this credulity, these misbeliefs are enough and work very well. I mean: it is not necessary that such an Authority really exists and sanctions people; what is necessary and sufficient is that people believe and fear this. Sometimes, misbeliefs can be adaptive and can be useful for agents and societies in spite of their being false! (Doran 1998)" (Castelfranchi 2003).

to be conventionally presupposed in specific social games and justifies and gives meaning at that behavors, and acquires special effects.

Its conventional and presuppositional nature is so relevant that those representations can even not be there at all or not be true and sincere, but they equally work.

Moreover, even the existence of a "mind" in the other, and of reasons for his/her conduct, goals, intentions, emotions, beliefs, … is a necessary presupposition of "social" interaction (we do not treat people like objects) and the interaction works thanks to such an assumption, which not only shapes that mind but creates it as a conventional entity with conventional effects: an institution.

Let's add that the strict relationship between minds and institutions as been deeply discussed also in Economics (especially in Institutional economics), with important claims. In particular, North and its school gave important contributions. Consider for example this passage: "The relationship between mental models and institutions in a intimate one. *Mental models are the internal representations* that individual cognitive systems create to interpret the environment; *institutions are the external* (*to the mind*) *mechanisms* individuals create to structure and order the environment" (Denzau and North 1994, p. 4).

Our view is more extreme: *institutions are internal*, *inside the mind*; they give structure to our perception and thinking and then to our action; and this structure is a constitutive *part* of the institution. The "constraints" are internal; in other words, *the* "*mental models*" *are* "*institutions*"![28] And the individual cognitive structures are not only "socialized" (that is, learned in interaction and shared by imitation), they are "institutionalized" in a deep sense: they "count as," they work "as if" they were there; they have a conventional nature, ascription, and conventional effects (as we have just seen).

This dual facet and nature of institutions, this dialectics between "internal" and "external" should be made clearer. For example, Veblen's (1994) view of institutions in terms of "*mental habits*" might be reductive, limited as it is to the internal facet of institutions. The latter are also "external" entities (behavior regularities, constraints, norms, organizations, artificial entities like money) they are "coordination artifacts," and are selected also *from and for* these external "effects" (not only for their cognitive advantages).

"Institutions" are not only mentally represented in some way[29]; they are mentally implemented; they work in the minds and through the minds, but also through the external behaviors, and their coordinated effects. They are mental–behavioral, individual–collective, internal–external entities. They cannot be accounted for by a simplistic ontology. However, clearly we have still to explore more explicitly the conditions in which a mind can become an institutional artifact.

Acknowledgments I'm very grateful to Luca Tummolini, Maria Miceli, and all the friends of our "GOAL lab" at ISTC-CNR for our priceless discussions. I would like also to thank the two anonymous reviewers which were very analytic and helpful, obliging me to make more clear several points. I even exploit some of their very synthetic interpretations of my claims.

[28] A cultural artifact with conventional and special effects.

[29] But not really understood, see Castelfranchi (2001) and Conte and Castelfranchi (1995).

References

Apperly, I. (2010) *Mindreaders. The Cognitive Basis of "Theory of Mind"*. East Essex, UK: Psychology Press.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Troetschel, R. (2001). The automated will: unconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81*, 1004–1027.

Bargh, J. A., & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science, 3*(1), 73–79.

Batson, C. D. (1991). *The altruism question. Toward a social-psychological answer*. Hillsdale: Lawrence Erlbaum Associates.

Castelfranchi, C. (1997). Principles of individual social action. In: Tuomela R, Hintikka, G. (eds) *Contemporary action theory*. Kluwer, Norwell, MA

Castelfranchi, C. (2000). Through the agents' minds: cognitive mediators of social action. *Mind and Society*, 1(1):109–140

Castelfranchi, C. (2001). The theory of social functions. *Challenges for multi-agent-based social simulation and multi-agent learning. Journal of Cognitive Systems Research, 2*, 5–38.

Castelfranchi, C. (2003). Formalising the informal?: Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic, 1*(1–2), 47–92.

Castelfranchi, C. (2012a). Goals: The true center of cognition. In F. Paglieri, L. Tummolini, R. Falcone, & M. Miceli (Eds.), *The goals of cognition* (pp. 825–870). London: College Publications.

Castelfranchi, C. (2012b). Ascribing minds. *Cognitive Processing, 13*, 415–425.

Castelfranchi, C., & Conte, R. (1991). Emergent functionalities among intelligent systems: cooperation within and without minds. *AI and Society, 6*(1), 78–87.

Castelfranchi, C., & Miceli, M. (2009). The cognitive-motivational compound of emotional experience. *Emotion Review, 1*, 223–231.

Churchill, S. D. (1991). Reasons, causes, and motives: psychology's illusive explanations of behavior. *Theor Philo Pysch, 11*(1), 24–34.

Clark, A. (2008). *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: UCL Press.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Denzau, A., & North, D. (1994). Shared mental models: Ideologies and institutions. *Kyklos, 47*(1), 3–31.

Doran, J. (1998), Simulating collective misbelief. *Journal of Artificial Societies and Social Simulation*, 1(1), http://jasss.soc.surrey.ac.uk/1/1/3.html.

Gallagher, S., & Crisafi, A. (2009). Mental institutions. *Topoi, 28*, 45–51.

Garfinkel, H. (1963). A conception of, and experiments with, 'trust' as a condition of stable concerted actions. In O. J. Harvey (Ed.), *Motivation and social interaction* (pp. 187–238). New York: The Ronald Press.

Gilbert, D. T. (1998). Ordinary personology. In: Gilbert DT, Fiske ST, Lindzey G (eds.) *The handbook of social psychology*, 4th edn. New York: McGraw Hill

Goffman, E. (1959). *The presentation of self in everyday life*. New York: Anchor Books.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford: Oxford University Press.

Lewis, D. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.

Lorini, E., & Castelfranchi, C. (2007). The cognitive structure of surprise: looking for basic principles. *Topoi: An International Review of Philosophy, 26*(1), 133–149.

Lorini, E., Marzo, F. & Castelfranchi C. (2005). A cognitive model of altruistic mind. In Boicho Kokinov (Ed.) *Advances in cognitive economics*. Sofia: NBU Press, pp. 282–293

Luhmann, N. (1991). *Trust and Power*. Ann Arbor, MI: University Microfilms International.

Miceli, M., & Castelfranchi, C. (2011). Forgiveness: a cognitive-motivational anatomy. *Journal for the Theory of Social Behaviour, 41*, 260–290.

Miceli, M., & Castelfranchi, C. (2007). The envious mind. *Cognition and Emotion, 21*, 449–479.

Polanyi, M. (1958). *Personal knowledge. towards a post critical philosophy*. London: Routledge.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding*. Hillsdale: Erlbaum.

Searle, J. R. (1995). *The construction of social reality*. New York: The Free Press.

Tummolini, L., Andrighetto, G., Castelfranchi, C., & Conte, R. (2013). A convention or (tacit) agreement betwixt us: on reliance and its normative consequences. *Synthese, 190*(4), 585–618.

Tummolini, L., & Castelfranchi, C. (2006). The cognitive and behavioral mediation of institutions: towards an account of institutional actions. *Cognitive Systems Research, 7*(2–3), 307–323.

Veblen, T. (1994). *The collected works of Thorstein Veblen*. London: Routledge.