ELSEVIER

Artificial Intelligence 103 (1998) 157–182

Artificial Intelligence

Modelling social action for AI agents

Cristiano Castelfranchi [1]

National Research Council, Institute of Psychology, Unit of "AI, Cognitive Modelling & Interaction", v. Marx 15, 00137 Roma, Italy

Abstract

In the new AI of the 90s an important stream is artificial social intelligence. In this work basic ontological categories for social *action*, *structure*, and *mind* are introduced. Sociality (social action, social structure) is let emerge from the action and intelligence of individual agents in a common world. Also some aspects of the way-down—how emergent collective phenomena shape the individual mind—are examined. First, *interference* and *dependence* are defined, and then different kinds of *coordination* (reactive versus anticipatory; unilateral versus bilateral; selfish versus collaborative) are characterised. "Weak social action", based on beliefs about the mind of the other agents, and "strong social action", based on goals about others' minds and their actions, are distinguished. Special attention is paid to *Goal Delegation* and *Goal Adoption* that are considered as the basic ingredients of social commitment and contract, and then of exchange, cooperation, group action, and organisation. Different levels of delegation and then of *autonomy* of the delegated agent are described; and different levels of goal-adoption are shown to characterise true collaboration. Social goals in the minds of the group members are argued to be the real glue of joint activity, and the notion of *social commitment*, as different from individual and from collective commitment, is underlined. The necessity for modelling social *objective structures* and constraints is emphasised and the "shared mind" view of groups and organisations is criticised. The spontaneous and unaware emergence of a *dependence structure* is explained, as well as its feedback on the participants' minds and behaviours. Critical observations are presented on current confusions such as that between "social" and "collective" action, or between communication and social action.

The main claims of the paper are the following: (a) The real foundation of all sociality (cooperation, competition, groups, organisation, etc.) is *the individual social action and mind*. One cannot reduce or connect action at the collective level to action at the individual level unless one passes through the social character of the individual action. (b) Important levels of coordination and cooperation necessarily require *minds* and *cognitive agents* (beliefs, desires, intentions, etc.). (c) However, cognition, communication and agreement are not enough for modelling and implementing cooperation: emergent pre-cognitive structures and constraints should be formalised, and emergent forms of cooperation are needed also among planning and deliberative agents. (d) We are going

[1] Email: cris@pscs2.irmkant.rm.cnr.it.

0004-3702/98/$ – see front matter
PII: S0004-3702(98)00056-3

towards a *synthetic paradigm* in AI and Cognitive Science, reconciling situatedness and plans, reactivity and mental representations, cognition, emergence and self-organisation.

Keywords: Social agent; Social structures; Dependence; BDI; Emergence; Cooperation

Premise

AI is a science, not merely technology or engineering. It cannot find an identity in a technology, or set of technologies, and we know that such an identification is quite dangerous. AI is the science of possible forms of intelligence, both individual and collective. To rephrase Doyle's [21] claim, AI is *the discipline aimed at understanding intelligent beings by constructing intelligent systems.*

Since intelligence is mainly a social phenomenon and is due to the necessity of social life, there is the need to construct socially intelligent systems to understand it, and we have to build social entities to have intelligent systems. If we want a computer to be not "just a glorified pencil" (Popper, BBC interview), not a simple *tool* but a *collaborator* [28], an assistant, we need to model social intelligence in the computer. If we want to embed intelligent functions in both the virtual and physical environment (*ubiquitous computing*) in order to support human action, these distributed intelligent entities *must be social* to understand and help the users, and to coordinate, compete and collaborate with each other.

In fact Social Intelligence is one of the ways AI reacted to and got out of its crisis. It is one of the ways it is "back to the future", trying to recover all the original challenges of the discipline, its strong scientific identity, its cultural role and influence. In the 1960s and 1970s this gave rise to Cognitive Science, now it will strongly impact on the social sciences.

The Social Intelligence stream is a part of the new AI of the 1990s where systems and models are conceived for reasoning and acting in open unpredictable worlds, with limited and uncertain knowledge, in real time, with bounded (both cognitive and material) resources, interfering—either co-operatively or competitively—with other systems. The new password is *interaction* [3]: interaction with an evolving environment; among several, distributed and heterogeneous artificial systems in a network; with human users; among humans through computers.

Important work has been done in AI (in several domains from DAI to HCI, from Agents to logic for action, knowledge, and speech acts) for modelling social intelligence and behavior. Here I will just attempt a principled systematization of these achievements.

On the one hand, I will illustrate what I believe to be the basic ontological categories for social *action*, *structure*, and *mind*. I will let, first, sociality (social action, social structure) emerge from the action and intelligence of individual agents in a common world, and, second, examine some aspects of the way-down: how emergent collective phenomena shape the individual mind. I will mainly focus on the bottom-up perspective. On the other hand, I will develop the following critical reflections on current approaches and futures directions.

Social versus collective. "Social action" is frequently used—both in AI and philosophy —as the opposite of individual action, that is as the action of a group or a team rather than of an individual. It is intended to be a form of collective activity, possibly coordinated and orchestrated, thus leading to joint action. However, *we should not confuse or identify social action/intelligence with the collective one.*

Many of the theories about joint or group action (for example, [34,51,52]) try to build it up on the basis of individual action: by reducing for example joint intention to individual non-social intentions, joint plan to individual plans, group commitment (to a given joint intention and plan) to individual commitments to individual tasks. This is just a simplistic shortcut. In this attempt the intermediate level between individual and collective action is bypassed. The real foundation of all sociality (cooperation, competition, groups, organization, etc.) is missed: i.e., *the individual social action and mind* [11].

One cannot reduce or connect action at the collective level to action at the individual level unless one passes through the social character of the individual action. Collective agency presupposes individual social agents: the individual social mind is the necessary precondition for society (among cognitive agents). Thus we need a definition and a theory of *individual* social action and its forms.

The intentional stance: *mind reading*. Individual action is either social or non social depending on its purposive effects and on the mind of the agent. The notion of social action cannot be a behavioral notion—just based on an external description. *We need to model mental states* in agents and to have representations (both beliefs and goals) of the minds of other agents.

Social action versus communication. The notion of social action (that is foundational for the notion of Agent) cannot be reduced to communication or modelled on the basis of communication. *Agents are not "agents" by virtue of the fact that they communicate; they cannot be called "social" because they communicate* but the other way around: they communicate because they are social. They are social because they act in a common world and because they interfere with, depend on, and influence each other.

Social action and communication versus cooperation. Social interaction and communication are mainly based on some exercise of power, on either unilateral or bilateral attempts to influence the behavior of the other agents by changing their minds. Both interaction and communication are frequently aimed at blocking, damaging, or aggressing against the others, or at competing with them. Social interaction (including communication) is not the joint construction and execution of a multi-agent plan, of a shared script, necessarily based on mutual beliefs. It is not necessarily a cooperative activity [7].

Reconciling "Emergence" and "Cognition". Emergence and cognition are not incompatible with one another, neither they are two alternative approaches to intelligence and cooperation.

On the one hand, Cognition has to be conceived as a level of emergence (from sub-symbolic to symbolic; from objective to subjective; from implicit to explicit). On the other side, emergent and unaware functional social phenomena (for example, emergent cooperation, and swarm intelligence) should not be modelled only among sub-cognitive agents (see Section 1) [37,50], but also among intelligent agents. In fact, for a theory of

cooperation and society among intelligent agents *mind is not enough* [19]. I will stress the limits of deliberative and contracting agents as for complex social behavior: cognition cannot dominate and exhaust social complexity [31].

I will present *a basic ontology of social action* by examining its most important forms, with special focus on the pro-social ones, in particular *Goal Delegation* and *Goal Adoption*. They are the basic ingredients of social commitment and contract, and then of exchange, cooperation, group action, and organization. We need such an analytical account of social action to provide a good conceptual apparatus for social theory. I will give some justification of this analysis also in terms of its theoretical and practical usefulness for AI systems, arguing against some current biases typical of AI social models.

I will try to show why we need *mind-reading* and cognitive agents (and therefore why we have to characterize cognitive levels of coordination and social action); why we need goals about the other's mind (in interaction and in collaboration), or social commitment to the other; why cognition, communication and agreement are not enough for modelling and implementing cooperation; why emergent pre-cognitive structures and constraints should be formalized, and why emergent cooperation is needed also among planning and deliberative agents.

1. Sociality step by step. The goal-oriented character of agents and actions

Sociality presupposes agents and goals. At a very basic level, an agent is any entity *able to act*, i.e., to produce some causal effect and some change in its environment. Of course this broad notion (including even natural forces and events) is not enough for sociality. We need a more complex level of agenthood. An agent receives and exploits relevant information from and about the world [25]. In which sense this is "information" for the agent? why is it "relevant"? Our agent bases its action on it, i.e., on its perception of the world. In such a way its behavior or reaction is adapted to the environment. In other terms, the agent's behavior is aimed at producing some result: thus we are talking of a *goal-oriented* action and of a goal-oriented agent [18,38].

Systems oriented towards some goals (although without any explicit internal representation of those goals) can exhibit social behavior. An "agent" can be helped or damaged, favoured or threatened, it can compete or cooperate. These notions can meaningfully apply only to systems endowed with some form of goal.

Among goal-oriented systems I will consider in particular *goal-directed* system. In these systems not only action is based on perception, but the latter is also the perception of the action's effects and results, and the agent regulates and controls its action on such a basis. *The agent is endowed with goals,* i.e., internal anticipatory and regulatory representations of action results. To be more precise: *Actions* are teleonomic or goal-oriented behavior. We admit goal-oriented behaviors that are not goal-directed (for example, in many animals, or in functional tools), i.e., behaviors that are not motivated, monitored and guided by an internal (mental) representation of the effects. [2]

[2] For an attempt to theoretically unify mental and non-mental notions of goal, action, sociality, etc., see [18, Chapter 8; 5.3].

A **goal** is a mental representation of a world state or process that is candidate for: [3]

- *controlling* and *guiding* action by means of repeated tests of the action's expected or actual results against the representation itself;
- determining the action search and selection;
- qualifying its success or failure.

This notion of goal-directed behaviour is based on the very operational notion of goal and "purposive behaviour" proposed by Rosenblueth and Wiener [45], and developed, in psychology, by Miller, Galanter and Pribram [39]. This very clear definition of the purposive character of action unfortunately is currently quite disregarded in AI, agent theory, and action logics.

Action and social action (SA) is possible also at the reactive level, among sub-cognitive agents (like bees). [4] By "sub-cognitive" agents I mean agents whose behavior is not regulated by an internal explicit representation of its purpose and by explicit beliefs. Sub-cognitive agents are for example simple neural-net agents, or mere reactive agents.

We will analyse here only *goal-directed action* that requires *cognitive agents*, i.e., agents whose actions are internally regulated by goals and whose goals, decisions, and plans are based on beliefs. Both goals and beliefs are cognitive representations that can be internally generated, manipulated, and subject to inferences and reasoning.

Since a goal-directed agent may have more than one goal active in the same situation, it must have some form of choice/decision. It also has an action repertoire (skills), some recipes, and some resources. It has limited abilities and resources; thus it is able to achieve only some of its goals.

2. Interference and dependence (first step)

Sociality obviously presupposes two or more agents in a common, shared world.

A "Common world" implies that there is *interference* among the actions and goals of the agents: the effects of the action of one agent are relevant for the goals of another: i.e., they either favour the achievement or maintenance of some goals of the other's (*positive*

[3] Notice that we use "goal" as the general family term for all motivational representations: from desires to intentions, from objectives to motives, from needs to ambitions, etc. In fact, "desire" is not a good general term since it cannot comprehend duties, obligations, needs, and other types of goal (see Section 7.1).

[4] A definition of SA, communication, adoption, aggression, etc. is possible also for non-cognitive agents. However, also at this level, those notions must be goal-based. Also at a sub-cognitive level, a SA is *a goal-oriented behavior that deals with another entity as an agent, i.e., as an active, autonomous, goal-oriented entity.*

Consider for example animal communication. One cannot consider as "communication" any meaningful signal arriving to an agent, for example the light of a fire-fly intercepted by a predator. For sure neither the prey is sending this message to the predator, nor the fire-fly's message has been selected by evolution *for* informing the predator about the position of the prey. The "function" of that signal—what it is for—is to inform the male fire-fly about the position of the female. So, first, there may be "meaning" and "sign" without communication; second, there is communication when the signal is purposively or at least finalistically sent to the addressee. In this case the "goal" is the biological "function" of the behavior.

Thus, a theory of *merely* goal-oriented (not "goal-directed") systems and of *implicit goals* and *functions* is needed [18, Chapter 8].

Although there are SA, communication, adoption, and aggression also among non-cognitive agents, however, there are levels of sociality that cannot be attained reactively (see later).

Table 1

		A To adapt	B To induce
1	Negative interference	To modify one's plan to avoid the obstacle	To induce the other to abandon his threatening goal
2	Positive interference	To modify one's plan by inserting y's action to exploit it	To induce the other to pursue the goal one needs

interference), or threat some of them (*negative interference*) [6,30,40]. Dependence is a special and strong case of interference.

In a Dependence relation not only agent y can favour agent x's goal, but x is not able to achieve her own goal (because she lacks a necessary resource or any useful action) while y controls the needed resource or is able to do the required action.

2.1. *Basic moves*

Let us first discover sociality from the point of view of the agent subject to interference. From her self-interested perspective, in *interference* and *dependence* agent x has two alternatives [11]:

(A) **to adapt her behavior** (goals, plans) to y's behavior, in order to exploit y's action or to avoid y's negative interference;

(B) to attempt **to change y's behavior** (goals, plans) by inducing him to do what she needs or to abandon the dangerous behavior.

Column A in Table 1 represents "mere coordination" (either *negative* or *positive*); column B "influencing"; in row 2 we find "delegation". In both cases A and B we possibly have "social action" by x, but of a very different nature. And we have "social action" only under some specific conditions.

3. From non-social action to weak social action: beliefs about the other's mind (second step)

Any action is in fact inter-action, since its environment is never a passive entity: the action is aimed at producing effects *on* the environment and is controlled by the feedback *from* the environment. More than this: there is always some "delegation" to the environment of part of the causal process determining the intended effect, some reliance on the "activity" of the environment and its causal forces (see our very general definition of weak delegation, Section 5.1). So actions are in fact interactions between the agents and the environment. However, this does not imply that any action should be a "social" one. The environment is—as just said—a causal "agent" involved in our plan/action, but this (inter)action is not social, because the environment is not a goal-oriented agent. For example, we can exploit the sun, but we really cannot "help" it.

Of course, if a primitive or superstitious man considers nature and objects as animate beings, from his subjective point of view he is performing social actions (and collaboration) when he is helped by the "spirits" of the objects.

Exploiting biological nature starts to be a "social" behavior at the weakest level, because the plants, ferments, viruses, etc. we exploit or try to avoid (preventing their activity) are in fact goal-oriented systems and we treat them as such. While we are not collaborating with sun and rain, since they do not have "ends", plants, on the contrary, are in some sense "unintentionally" collaborating with us since they have the "goal" of producing fruits, etc., and we insert not only their effects but their "goals" in our plans, collaborating with them. Agriculture is in fact some sort of "collaboration" between man and nature.

A SA is *an action that deals with another entity as an agent, i.e., as an active, autonomous, goal-oriented entity.*

For *cognitive agents*, a SA is *an action that deals with another cognitive agent considered as a cognitive agent, whose behavior is regulated by beliefs and goals.*

In SA the agent takes an Intentional Stance towards the other agents: i.e., a representation of the other agent's mind in intentional terms [20].

Consider a person (or a robot) running in a corridor and suddenly changing direction or stopping because of a moving obstacle which crosses its path. Such a moving obstacle might be either a door (opened by the wind) or another person (or robot). The agent's action does not change its nature depending on the objective nature of the obstacle. If x acts towards another agent as if it were just a physical object, her action *is not a SA*. Its being a social action or not depends on how x *subjectively* considers the other entity in her plan. Consider the same situation but with some more pro-active (rather than reactive) attitude by x: x foresees that y will cross her way on the basis of her beliefs about y's goals, as it happens in traffic, when we slow down or change our way because we understand a driver's intention just on the basis of his behavior (without any special signal). This action of x starts to be "social", since it is based on x's belief about y's mind and action (not just behavior). This is in fact a true example of social "coordination" (see Section 4).

So, *an action related to another agent is not necessarily social* [11]. Also the opposite is true. A merely practical action, not directly involving other agents, may be or become social. The practical action of closing a door is social when we close the door to avoid that some agent enters or looks inside our room; the *same* action performed to block wind or rain or noise is not social. *Not behavioral differences but goals distinguish social action from non social action.*

Consider an agent ADAM in a block world, just doing his practical actions on blocks. His goal is "blocks A and B on the table". Thus he grasps A and puts it on the table (Fig. 1). There is no social aspect present in this action.

Now suppose that another agent, EVE, enters this world. EVE has the goal "small block a on block B" but she is not able to grasp big blocks, so she cannot achieve her goal. ADAM is able to grasp big blocks: so EVE is *dependent on* ADAM, since if ADAM performs the needed action EVE will achieve her goal [8]. Now, suppose that ADAM, knowing about EVE's goals and abilities, decides to help EVE. He grasps A and puts it on the table so that EVE finally can perform the action of putting a on B and achieve her goal. ADAM's action *is exactly the same action on blocks* performed when he was alone, but now it is a SA: ADAM is *helping* EVE. It is a SA based on beliefs about EVE's goals.

We may call "weak SA" the one based just on *social beliefs*: beliefs about other agents' minds or actions; and "strong SA" that which is also directed by *social goals*.

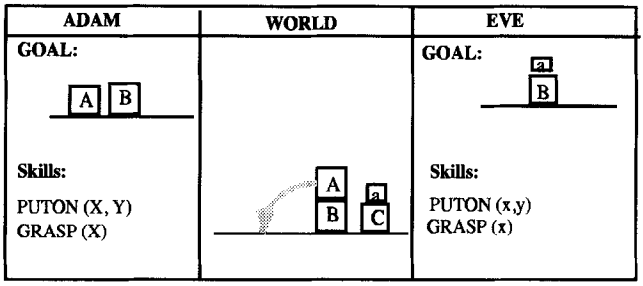

Fig. 1.

The true basis of any level of SA among cognitive agents is *mind-reading* [1]: the representation of the mind of the other agent.

Notice that beliefs about the other's mind are not only the result of communication about mental states (emotions; language), or of stereotypical ascription, but also of "interpretation" of the other's behavior. Strictly speaking, the other's behavior becomes a "sign" of his own mind. This *understanding*, as well as behavioral and implicit communication is, before explicit communication (special message sending), the true basis of reciprocal coordination and collaboration [44]. In contrast to current machines, human beings do not coordinate with each other by continuously sending special messages (like in the first CSCW systems): we monitor the other's behavior or its results, and let the other do the same.

3.1. Communication, agents, and social action

It is common agreement in AI that "social agents" are equivalent to "communicating agents". According to many authors communication is a necessary feature of agency (in the AI sense) [27,47,56]. Moreover, the advantages of communication are systematically mixed up with the advantages of coordination or of cooperation.

Communication is just an instrument for SA (of any kind: either cooperative or aggressive [7]). Communication is also a kind of SA aimed at giving beliefs to the addressee. This is a true and typical Social Goal, since the intended result concerns a mental state of another agent. [5]

Thus *communication is not a necessary component of social action and interaction.* To kill somebody is for sure a SA (although not very sociable!) but it neither is nor requires communication. Also pro-social actions do not necessarily require communication. As we saw in EVE's example, unilateral help does not need communication (since it does not necessarily require awareness or agreement). [6]

[5] Notice that this typical SA does not necessarily involve any "sharing" or putting in common; in fact, contrary to common sense, communication is not necessarily truthful, and x can either believe or not believe what she is communicating to y: also lies are communication.

[6] A more subtle analysis of the example might reveal that in this case of help, Eve should be aware of the new state of the world. Thus, Adam not only has the goal that "B is clear" but also the goal that "Eve knows that B is clear". Although Adam is not sending a specific message to Eve (an act specialised to communicate) he has

However, strict bilateral cooperation is based on *agreement* and requires some form of communication.

In AI—and in general in computer science—we tend to equate (inter)action to communication for quite a basic reason. In everyday life, or in manufacturing, or in market, agents exchange both messages and material resources (their products and services), and there is not only information transfer (for coordination and for collaboration) but also material transfer. On the contrary, with computers and networks the exchanged resources or the provided services are just data and information. That is why these two very different kinds of transfer among agents are mixed up as "communication". We should clearly maintain the distinction between the practical (inter)action—although in this case it is about data and information—and the specific sending of messages for preparing, coordinating and controlling these practical interactions [54].

To conclude, neither agenthood nor sociality are grounded on communication, although, of course, communication is very important for social interaction [11].

4. Principles of coordination

In simple coordination (Table 1, column A) x has just coordinated her behavior with y's perceived or predicted behavior, ignoring the possibility to change it; like in our traffic example, x changes her own plan (sub-goal) and produces a new goal which is based on her beliefs about y's goal (weak SA). While any forms of social interaction (including negotiation, cooperation, conflict, etc.) might be called "coordination" [36] since it requires some coordination or is useful for coordination, I prefer to restrict the use to this simpler form, in which there is *merely* coordination without influencing or communication.

4.1. Reactive versus anticipatory: coordination among cognitive agents

There are two types of mere coordination, depending on the kind of detection of the interference:

- **reactive coordination** which is based on the direct perception of an obstacle or opportunity and on a reaction to it;
- **proactive** or **anticipatory coordination** which lies in the anticipation (based either on learning or on inferences) of possible interference or opportunities.

The advantages of anticipatory coordination are clear: it can prevent damages or losses of resources; moreover a good coordination might require time to adapt the action to the new situation and prediction gives more time. In a sense a completely successful avoidance coordination cannot really be done without some form of anticipation. When the obstacle/damage is directly perceived, it is—at least partially—"too late", because either the risk increases or there is already some loss.

Some form of anticipation is possible also in mere reactive systems, when they learn to react to some forerunner of the relevant event (like in Pavlov's conditioned reflexes).

---

the goal of letting Eve believe something. This is a forerunner of communication. So his practical action acquires also some feature of implicit and behavioural communication.

However, this association must be very regular, based on fixed sequences, and on short intervals between the signal and the event.

Anticipatory coordination with complex and long term effects needs some theory or model of the world: i.e., some cognitive intelligence. Anticipatory coordination with cognitive goal-directed agents cannot be based just on learning or inferences about trajectories or frequencies of action sequences. Since agents combine their basic actions in several long and creative sequences, the prediction (and then the anticipatory coordination) must be based on *mind-reading*, that is on the understanding of the goals and the plan of the other [5]. Conflicts or opportunities are detected by the agent by comparing its own goals and plans with the goals/plans ascribed to the other. Of course, in social agents, stereotypes, scripts, habits, roles, rules, and personalities help this anticipation and understanding.

No agent could really "plan" its behavior in a Multi-Agent world without some anticipatory coordination. There is a co-evolutionary coupling between planning in a MA world and the mind-reading ability.

To anticipate a conflict is clearly much better that discovering it by crash. Avoiding damages is better than recovering from them. *This is something reactive agents can do in a very limited way*: as we said they could at most have some—learned, built in, or inherited—reaction to some short-term behavioral fixed sequence.

4.2. Positive and negative coordination; unilateral, bilateral, and mutual

Avoidance coordination or **negative coordination** is due to negative interference and aimed at avoiding a damage or an "obstacle". In **exploitation coordination** or **positive coordination**, x changes her plan (assigning at least a part of it to another agent: delegation) in order to profit by a favourable (social) circumstance.

In **unilateral coordination** only x is coordinating her own activity with y's activity; but it is possible that y is doing the same. In this case the coordination is **bilateral**. The two coordination intentions and actions may be independent of each other. If either agent does not understand the new coordinated plan there can be some trouble. The bilateral coordination is **mutual** when both the agents are aware of their coordination intentions and try to arrive at some agreement. Mutual coordination necessarily requires some collaborative coordination.

4.3. Selfish versus collaborative coordination

All the previous ones (cf. Table 1, column A) are the basic forms of the ***ego-centred*** or ***selfish coordination***: x tries to achieve her own goal while dealing with y's presence and action in the same world, adapting her behavior to the other's behavior. However other forms of coordination are possible: for example x might modify her own behavior in order to avoid negative interference with the other's action or to create positive interferences. This is ***Collaborative Coordination***: x is adapting her behavior in order to favour y's actions [40].

Collaborative coordination is already a form of strong SA. In fact, it is not only based on beliefs relative to the other mind, but is guided by a Social Goal: the goal that the other achieves his goal. It necessarily implies some form of either passive or active help (Goal-

Adoption—see Section 7). Collaborative coordination is the basis of Grosz and Kraus' social "intention that" [29].

Box A2 in Table 1 represents a very important form of Coordination because it is also the simplest, elementary form of Delegation or Reliance.

5. Relying on (Delegating)—(third step)

We will now examine those elementary forms of SA that are the basic ingredients of help, exchange, cooperation, and then of partnership, groups and team work. Let us see them at their "statu nascenti", starting from the mere unilateral case.

On the one side, there is the mental state and the role of the future "client" (who relies on another agent's action to achieve her goal). Let us call this *Delegation* or *Reliance*. On the other side, there is the mental state and role of the future "contractor" (who decides to do something useful for another agent, adopting a goal of hers). Let us call this *Goal Adoption*.

In **Delegation** x needs or likes an action of y and includes it in her own plan: she relies on y. She plans to achieve p through the activity of y. So, she is constructing a MA plan and y has a share in this plan: y's delegated *task* is either a state-goal or an action-goal [14].

If EVE is aware of ADAM's action, she is *delegating* ADAM a task useful for her. The following conditions characterize EVE's **trust** in ADAM [15]:

- she believes that ADAM can and will do a given action;
- she has the goal that ADAM does it (since she has the goal that it will be done),
- she relies on it (she abstains from doing it, from delegating to others, and coordinates her own action with the predicted action of ADAM).

5.1. From non-social to social delegation

Unilateral weak delegation. In Unilateral Delegation there is neither bilateral awareness of the delegation, nor agreement: y is not aware of the fact that x is exploiting his action. One can even "delegate" some task to an object or *tool*, relying on it for some support and result [18, Chapter 8], [35]. In the weakest and passive form of unilateral delegation x is just exploiting the autonomous behavior of y; she does not cause or elicit it.

As an example of weak and passive, but already social delegation (which is the simplest form of social delegation) consider a hunter who is ready to shoot an arrow at a flying bird. In his plan the hunter includes an action of the bird: to continue to fly in the same direction (which is a goal-oriented behavior); in fact, this is why he is not pointing at the bird but at where the bird will be in a second. He is delegating to the bird an action in his plan; and the bird is unconsciously and unintentionally "collaborating" with the hunter's plan.

Delegation by induction. In this stronger form of delegation x is herself eliciting or inducing y's behavior in order to exploit it. Depending on the reactive or deliberative character of y, the induction is either based on some stimulus or on beliefs and complex types of influence.

As an example of unilateral Delegation by induction consider a fisherman: unlike the hunter, the fisherman elicits by himself—with the bait—the fish's action (snapping) that is

part of his plan. He delegates this action to the fish (he does not personally attach the fish to the hook) but he also induces this reactive behavior.

Delegation by acceptance (strong delegation). This Delegation is based on y's awareness of x's intention to exploit his action; normally it is based on y's adopting x's goal (Social Goal-Adoption), possibly after some negotiation (request, offer, etc.) concluded by some agreement and social commitment. EVE asks ADAM to do what she needs and ADAM accepts to adopt EVE's goal (for any reason: love, reciprocation, common interest, etc.). Thus to fully understand this important and more social form of Delegation (based on social goals) we need a notion of Social Goal-Adoption (see Section 7); we have to characterise not only the mind of the delegating agent but also that of the delegated one.

5.2. *Plan-based levels of delegation*

Now we introduce the notion of *levels* of delegation, which become crucial for a theory of collaborative agents.

Given a goal and a plan (sub-goals) to achieve it, x can delegate goals/actions (tasks) at different levels of abstraction and specification [14]. We can distinguish between several levels, but the most important ones are the following:

- *pure* ***executive delegation*** *versus* ***open delegation***;
- ***domain task delegation*** *versus* ***planning and control task delegation*** *(meta-actions)*

The object of delegation can be minimally specified (***open delegation***), completely specified (***close*** or ***executive delegation***) or specified at any intermediate level.

Open delegation necessarily implies the delegation of some meta-action (planning, decision, etc.); it exploits intelligence, information, and expertise of the delegated agent. Only *cognitive delegation* (the delegation of a *goal*, an abstract action or plan that need to be autonomously specified—Section 7) can be "open": thus, it is *something that non-cognitive agents, and in particular non goal-directed agents, cannot do.*

Necessity for and advantages of open delegation. It is worth stressing that *open delegation* is not only due to x's preferences, practical ignorance or limited ability. It can also be due to x's ignorance about the world and its dynamics: *fully specifying a task is often impossible or inconvenient*, because some local and updated knowledge is needed for that part of the plan to be successfully performed. Open delegation ensures the *flexibility* of distributed and MA plans.

The distributed character of the MA plans derives from the *open delegation*. In fact, x can delegate to y either an entire plan or some part of it (*partial delegation*). The combination of the *partial delegation* (where y might ignore the other parts of the plan) and of the *open delegation* (where x might ignore the sub-plan chosen and developed by y) creates the possibility that x and y (or y and z, both delegated by x) collaborate in a plan that they do not share and that *nobody* entirely knows: that is a *distributed plan* [26,29,30]. However, for each part of the plan there will be at least one agent that knows it. This is also the basis for Orchestrated cooperation (a boss deciding about a general plan), but it is not enough for the emergence of functional and unaware cooperation among planning agents [12].

5.3. *Motivation for delegation*

Why should an agent delegate some action to another, trust it and bet on it?

Delegation is due to *dependence*: x, in order to achieve some goal that she is not able to achieve by herself—be it a concrete domain action or result, or a goal like saving time, effort, resources—delegates the action to another agent both able and willing to do it.

Agent x either lacks some know how, or ability, or resource, or right and permission, and is depending on the other agent for them.

Of course, x can delegate actions that she is able to do by herself; she just *prefers* to let the others perform them on her behalf. This is Jennings' notion of "weak dependence" [32]. However, if x prefers exploiting the action of the other agent for her goal that p, this means that this choice is better to her, i.e., there is some additional goal she achieves by delegating (for example, saving effort, time, resources; or having a proper realization of the goal, etc.). Relative to this more global or complete goal which includes p, x strictly depends on the other. So the dependence relative to global intended results of delegation is the general basis of delegation.

The effect of delegation and of the complementary adoption, and in general of pro-social interaction, is to augment and multiply the power of an individual of achieving its goals, by exploiting the powers of other agents.

6. *Strong* SA: goals about the other's action/goal (fourth Step)

In Delegation x *has the goal that* y *does a given action* (that she needs and includes in her plan). If y is a cognitive agent, x *also has the goal that* y *has the goal* (more precisely *intends*) *to do that action*. Let us call this "**cognitive delegation**", that is delegation to an intentional agent. This goal of x is the motive for *influencing* y [6,41], but it does not necessarily lead to influencing y. In fact, our goals may be realized by the independent evolution of the environment (including events and other agents' actions). Thus, it might be that x has nothing to do because y independently intends to do the needed action.

Strong social action is characterized by social goals. A social goal is defined as a goal that is *directed toward* another agent, i.e., whose intended results include another agent as a cognitive agent: *a social goal is a goal about other agents' minds or actions*. Examples of typical social goals (strong SAs) are: changing the other mind, communication, hostility (blocking the other goal), cognitive Delegation, adoption (favouring the other's goal).

We not only have *beliefs* about others' beliefs or goals (weak social action) but also *goals* about the mind of the other: EVE wants that ADAM believes something; EVE wants that ADAM wants something. We cannot understand social interaction or collaboration or organisations without these social goals. Personal intentions of doing one's own tasks, plus beliefs (although mutual) about others' intentions (as used in the great majority of current AI models of collaboration) are not enough (see Section 8).

For a cognitive autonomous agent to have a new goal, he ought to acquire some new *belief* [9]. Therefore, cognitive influencing consists in providing the addressee with information that is supposed to be relevant for some of her/his goals, and this is done in order to ensure that the recipient has a new goal.

6.1. *Influencing, power and incentive engineering*

The basic problem of social life among cognitive agents lies beyond mere coordination: *how to change the mind of the other agent? how to induce the other to believe and even to want something* (Table 1, column B)? How to obtain that y does or does not do something? Of course, normally—but not necessarily—by communicating.

However, communication can only inform the other about our goals and beliefs about his action: but *why should he care about our goals and expectations*? He is not necessarily a benevolent agent or an obedient slave. Thus, in order to induce him to do or not to do something we need power over him, power of influencing him. His benevolence towards us is just one of the possible basis of our power of influencing him (authority, sympathy, etc. can be others). However, the most important basis of our power is the fact that also our actions are potentially interfering with his goals: we might either damage or favour him: he is depending on us for some of his goals. We can exploit this (his *dependence*, our *reward or incentive power*) to change his mind and induce him to do or not to do something [6].

Incentive engineering, i.e., manipulating the other's utility function (his outcomes or rewards due to goal achievement or frustration), is not the only way we may have to change the mind (behavior) of the other agent. In fact, in a cognitive agent, pursuing or abandoning a goal does not depend only on preferences and on beliefs about utility. To pursue or abandon his intention, y should have a host of beliefs, that are not reducible to his outcomes. For example, to do p y should believe that "p is possible", that "he is able to do p", that "p's preconditions hold", that "necessary resources are allowed", etc. It is sufficient that x modifies one of these beliefs in order to induce y to drop his intention and then restore some other goal which was left aside but could now be pursued.

The general law of influencing cognitive agents' behavior does not consist in incentive engineering, but in modifying the beliefs which "support" goals and intentions and provide reasons for behavior. Beliefs about incentives represent only a sub-case.

7. *Strong SA*: Social Goal Adoption (fifth step)

Let us now consider SA from y's (the contractor or the helper) perspective. Social Goal-Adoption (shortly *G-Adoption*) deserves a more detailed treatment, since:

(a) it is the true essence of all pro-social behavior, and has several different forms and motivations;
(b) its role in cooperation is often not well understood. In fact, either agents are just presupposed to have the same goal (see, for example, [53]), or the adoption of the goal from the other partners is not explicitly accounted for [34,51,52]; or the reasons for adopting the other's goal and take part in the collective activity are not explored.

In **G-Adoption** y's mind is changing: *he comes to have a new goal or at least to have new reasons for an already existing goal*. The reason for this (new) goal is the fact that another agent x wants to achieve this goal: y knows this and decides to make/let her achieve it. So, y comes to have the same goal as x, *because* he knows that it is x's goal.

However, the previous characterisation is too broad: this social attitude and action should not be mixed up with simple *imitation*, which might be covered by that definition. In G-

Adoption y has the goal that p (wants p to be true) in order for x to achieve it. In other words, *y is adopting a goal of x's when y wants x to obtain it as long as y believes that x wants to achieve that goal* [18].

7.1. Goal-adhesion or compliance

Among the various forms of G-Adoption, *G-Adhesion* or *Compliance* has a special relevance, especially for modelling agreement, contract and team work. That occurs when the G-Adoption responds to another's request (implicit or explicit). It is the opposite of spontaneous forms of G-Adoption. Not only x has a goal p and y adopts this goal, but x herself has the goal that y does something for p, and the goal of letting y know about her expectation (request). Thus in G-Adhesion y adopts x's goal that he adopts her goal, i.e., he complies with x's expectations.

In order to satisfy x, not only y must achieve p (like in spontaneous and weak G-Adoption) and let x know that p, but he must also let x know that he performed the expected and delegated action and produced p.

G-Adhesion is the strongest form of G-Adoption. Agreement is based on adhesion; strong delegation is request for adhesion. In negotiations, speech acts, norms, etc. that are all based on the communication by x of her intention that the other does something, or better adopts her goal (for example, obeys), G-Adhesion is what really matters.

7.2. Social agent's architecture and multiple goal-sources

Through social *goal-adoption* we obtain a very important result as for the architecture of a social agent:

Goals (and then intentions) are not all originated from Desires or Wishes, they do not derive all from internal motives. A social agent is able to "receive" goals from outside: from other agents, from the group, as requests, needs, commands, norms.

If the agent is really autonomous it will decide (on the basis of its own motives) whether to adopt or not the incoming goal [9].

In architectural terms this means that there is not a unique origin of potential intentions [43] or candidate goals [2]. There are several origins or sources of goals: bodily needs; goals activated by beliefs; goals elicited by emotions; goals generated by practical reasoning and planning; and goals adopted, i.e., introjected from outside. All these goals have to converge at a given level in the same goal processing, in order to become *intentions* and be pursued through some action.

7.3. Motivation for G-adoption

Adoption does not coincide with *benevolence* [46]. A relation of benevolence, indeed, is a form of generalised adoption and is related to the motivation for G-Adoption.

Benevolence is a ***terminal*** (non instrumental) form of G-Adoption (pity, altruism, love, friendship). But *Goal-adoption can be also **instrumental** to the achievement of selfish goals*. For example, feeding chickens (satisfying their need for food) is a means

for eventually eating them. Instrumental G-Adoption also occurs in *social exchange* (reciprocal conditional G-Adoption).

Another motive-based type of G-Adoption (that might be considered also a sub-type of the instrumental one) is ***cooperative*** G-Adoption: y adopts x's goal because he is co-interested in (some of) x's intended results: they have a common goal. Collaborative coordination (Section 4.3) is just one example of it.

The distinction between these three forms of G-Adoption is very important, since their different motivational bases allow important predictions on y's "cooperative" behavior. For example, if y is a rational agent, in social exchange he should try to cheat, not reciprocating x's adoption. On the contrary, in cooperative adoption y normally is not interested in free riding since he has the same goal as x and they are *mutually dependent* on each other as for this goal p (Section 9.2): both x's action and y's action are necessary for p, so y's damaging x would damage himself. Analogously, while in terminal and in cooperative adoption it might be rational in many cases to inform x about difficulties, obstacles, or defections [32, 34], in *exchange*, and especially in forced, coercive G-Adoption, this is not the case at all.

Current AI models of collaboration, group, and organizations are not able to distinguish between these motive-based forms of Goal Adoption, while those distinctions will become practically quite important in MA collaboration and negotiation in the Web (self-interested agents; iterated interactions; deception; etc.).

7.4. Levels of collaboration

Analogously to delegation, several dimensions of adoption can be characterized [14]. In particular, the following levels of adoption of a delegated task can be considered:

- ***Literal help***: x adopts exactly what was delegated by y (elementary or complex action, etc.).
- ***Overhelp***: x goes beyond what was delegated by y, without changing y's plan.
- ***Critical help***: x satisfies the relevant results of the requested plan/action, but modifies it.
- ***Overcritical help***: x realizes an Overhelp by, at the same time, modifying or changing the plan/action.
- ***Hyper-critical help***: x adopts goals or interests of y that y himself did not consider; by doing so, x does not perform the action/plan, nor satisfies the results that were delegated.

On such a basis one can characterize the *level of collaboration* of the adopting agent.

An agent that helps another by doing just what is literally requested, is not a very collaborative agent. He has no initiative, he does not care for our interests, does not use his knowledge and intelligence to correct our plans and requests that might be incomplete, wrong or self-defeating.

A truly helpful agent should care for our goals and interests, and go beyond our delegation and request [14,16]. But, *only cognitive agents can non-accidentally help beyond delegation*, recognizing our current and contextual needs.

Of course, there is danger also in taking the initiative of helping us beyond our request. Troubles may be either due to misunderstandings and wrong ascriptions, or to conflicts and paternalism.

8. Social goals as the glue of joint action

Although clearly distinct from each other, *social* action/goal and *joint* action/goal are not two independent phenomena. In order to have a theory of joint action or of group and organization, a theory of social goals and actions is needed. In fact *social goals in the minds of the group members are the real glue of joint activity.*

I cannot examine here the very complex structure of a team activity, or a collaboration, and the social mind of the involved agents; or the mind of the group assumed as a complex agent. There are very advanced and valid formal characterisations of this [22,29,34,42,52, 55]. I would simply like to stress how social action and goals, as previously characterised, play a crucial role in joint action.

No group activity, no joint plan, no true collaboration can be established without:

(a) the goal of x (member or group) about the intention of y of doing a given action/task a (delegation);
(b) x's "intention that" [29] y is able and has the opportunity to do a; and in general the "collaborative coordination" of x relative to y's task. This is derived from the delegation and from the necessary coordination among actions in any plan; [7]
(c) the *social commitment* of y to x as for a, which is a form of goal-adoption or better adhesion.

Both *Goal-Adoption* in collaboration and groups, and the *goal about the intention of the other* (influencing) are either ignored or just implicitly presupposed in the above mentioned approaches. They mainly rely on the agents' beliefs about others' intentions; i.e., a weak form of social action and mind. The same is true for the notion of cooperation in Game Theory [13].

As for the social commitment, it has been frequently confused with the individual (non social) commitment of the agent to his task.

8.1. Social commitment

Social Commitment results from the merging of a strong delegation and the corresponding strong adoption: *reciprocal social commitments constitute the most important structure of groups and organizations.*

There is a pre-social level of commitment: the **Internal** or **individual** Commitment [10]. It refers to *a relation between an agent and an action.* The agent has decided to do

[7] To be more precise, the "intention that"—as defined by Grosz and Kraus [29]—has two distinct origins. On the one hand, it comes from G-Adoption; on the other hand, from Goal Delegation.

When y adopts a goal of x in a passive form, i.e., just letting x pursue her goal, he has the goal that x achieves p, and then he has also the goal that x is able to pursue p. This can lead to Collaborative Coordination, where y actively tries to favour x's pursuit.

In G-Delegation, since x relies on y's action for the achievement of her own goal, she will have both the goal that y performs that action, and the goals that he intends and is able to do it. Thus, for sure, she will also use some Collaborative Coordination towards y.

Notice that this attitude is simply the social version of the usual intention, present in any (non-social) plan, of not hindering with some action performance the execution of other actions in the same plan. Since in collaboration (and in general in Delegation) the plan is a multi-agent one, the intention of coordinating and not hindering becomes a social one.

something, the agent is determined to execute a given action (at the scheduled time), and the goal (intention) is a persistent one: for example, the intention will be abandoned only if and when the agent believes that the goal has been reached, or that it is impossible to achieve it, or that it is no longer motivated.

A "social" commitment is not an individual Commitment shared by several agents. **Social Commitment** is a relational concept: *the Commitment of one agent to another* [10, 49]. More precisely, S-Commitment is a four-argument relation, where x is the committed agent; a is the action (task) x is committed to do; y is the other agent to whom x is committed; z is a third possible agent before whom x is committed (the witness, the guarantor, the authority, the group, etc.).

Social commitment is also different from **Collective** or Group Commitment [22]. The latter is *the Internal Commitment of a Collective agent* or group to a collective action. In other terms, a set of agents is Internally Committed to a certain intention/plan and there is mutual knowledge about that. *The collective commitment requires social commitments* of the members to the others members and to the group.

Not only social commitment combines acceptance-based Delegation and acceptance-based Adoption, but *when x is S-Committed to y, then y can (is entitled to)*: *control* if x does what she "promised"; *exact/require* that she does it; *complain/protest* with x if she doesn't do a; (in some cases) *make good his losses* (pledges, compensations, retaliations). Thus, Social Commitment *creates rights and duties* between x and y [10].

Although very relevant, the social commitment structure is not the only important structure constraining the organizational activity and society.

9. Social structures and organization

There is an implicit agreement about organizations in recent computational studies. Either in DAI theories of organization [4,26], or in formal theories of collective activity, team or group work, joint intention, and "social agents" [34], or in CSCW approaches to cooperation [54], organization is in fact accounted for by means of the crucial notion of "commitment". However, this approach is quite unsatisfactory, for a number of reasons:

(a) as already observed, the current definitions of commitment are insufficient to really account for stable group formation and activity: the theory of "social" commitment as a necessary premise for the theory of collective or group commitment, is in progress, and the normative aspects of commitment are not well developed;
(b) agents seem to be completely free (also in Organizations) to negotiate and establish any sort of commitment with any partner, without any constraint of dependence and power relations, of norms and procedures, of pre-established plans and coalitions.

Current views of Organization dominant in computer science (DAI, CSCW) risk to be *too "subjective"* and *too based on communication*. They risk to neglect the *objective basis* of social interaction (dependence and power relations) and its normative components.

Both *the "shared mind" view of groups*, team work, and coordination, just based on agents' beliefs and intentions, and *the "conversational" view of Organization* [54], find no structural objective bases, no external limits and constraints for the individual initiative: the "structure" of the group or organization is *just* the structure of interpersonal communication

and agreement, and the structure of the joint plan. The agents are aware of the social structure their are involved in: in fact, they create it by their contractual activity, and social organization lies only in their joint mental representations (*social constructivism*) [4,26]. There is also a remarkable lack of attention to the individual motivations to participate in groups and organizations: agents are supposed to be benevolent and willing to cooperate with each other.

9.1. *Structures and constraints*

Coordination in a group or organization is not only guaranteed by a shared mind (joint intentions, agreed plans, shared beliefs), reciprocal benevolence, and communication; there are several *structures* in any MA system:

- the interdependence and power structure;
- the acquaintance structure emerging from the union of all the personal acquaintances of each agent [24,30];
- the communication structure (the global net of direct or indirect communication channels and opportunities);
- the commitment structure, emerging from all the Delegation–Adoption relationships and from partnerships or coalitions among the agents;
- the structure determined by pre-established rules and norms about actions and interactions.

Each structure affects both the possibility and the success of the agents actions, and constrains (when known) their decisions, goals and plans. The agents are not so free to commit themselves as they like: their are conditioned by their dependence and power relations, their knowledge, their possible communication, their roles and commitments, social rules and norms.

Some of those structures are deliberately constructed by he agents (at least partially); others are emerging in an objective way.

Let us focus in particular on one structure: the network of interdependencies. Not only because it is the more basic one (see Section 2), but also because it is emerging before and beyond any social action, contract, and decision of the involved agents.

9.2. *An emergent objective structure: the dependence network*

Although all the paper has been a bottom-up construction of sociality from its elementary basis, I would like to give at least a clear example of the spontaneous emergence of some global phenomenon or structure, and of the "way-down": how such an emergent global phenomenon or structure feedbacks to change the mind (and the behavior) of the involved agents.

The structure of interference and interdependence among a population of agents is an *emergent* and *objective* one, *independent of the agents' awareness and decisions*, but it constrains the agents' actions by determining their success and efficacy.

Given a bunch of agents in a common world, and given their goals and their *different* and *limited* abilities and resources, they *are* in fact interdependent on each other: a dependence structure emerges.

There are several typical dependence patterns. In [8] the *OR-Dependence*, a disjunctive composition of dependence relations, and the *AND-dependence*, a conjunction of dependence relations, were distinguished. To give a flavor of those distinctions let me just detail the case of a two-way dependence between agents (*bilateral dependence*). There are two possible kinds of bilateral dependence:

Mutual dependence occurs when x and y depend on each other for realising a common goal p, which can be achieved by means of a plan including at least two different acts such that x depends on y's doing ay, and y depends on x's doing ax:

Cooperation is a function of mutual dependence: in cooperation, in the strict sense, agents depends on one another to achieve one and the same goal [53]; they are co-interested in the convergent result of the common activity.

Reciprocal dependence occurs when x and y depend on each other for realising different goals, that is, when x depends on y for realising x's goal that p x, while y depends on x for realising y's goal that p y, with p$x \neq$ py.

Reciprocal dependence is to *social exchange* what mutual dependence is to cooperation.

The Dependence network *determines* and *predicts* partnerships and coalitions formation, competition, cooperation, exchange, functional structure in organizations, rational and effective communication, and negotiation power.

9.3. The "way back": How an emergent structure feedbacks into the agents' mind

This pre-cognitive structure can "cognitively emerge": i.e., part of these constraints can become known. The agents, in fact, may have beliefs about their dependence and power relations.

Either through this "understanding" (***cognitive emergence***) or through blind learning (based for example on reinforcement), the objective emergent structure of interdependencies feedbacks into the agents minds, and changes them. Some goals or plans will be abandoned as impossible, others will be activated or pursued [48]. Moreover, new goals and intentions will rise, especially social goals. The goal of exploiting or waiting for some action of the other; the goal of blocking or aggressing against another, or helping her action; the goal of influencing another to do or not to do something; the goal of changing dependence relations. These new goals are direct consequences of dependence relationships.

Without the emergence of this self-organising (undecided and non-contractual) structure, social goals would never evolve or be derived.

10. Some concluding remarks and challenges

After this complex examination of the individual social action, its bases, its relations with some social structures, and with collective behavior, let me try to extract from these considerations some relevant points in form of conclusions, and some prospective claims.

I will first summarise my arguments about the necessity for *minds* and *cognitive agents* (beliefs, desires, intentions, etc.); second, the necessity for an articulated notion of "social" and for a theory of individual social action; and third, the complementary claim that *minds*

are not enough. Finally, I will consider in this perspective the *contribution of AI to* what I would like to call *the new synthesis*.

10.1. Why we need cognitive agents

Why do we need (and when) cognitive agents, i.e., agents who found their decisions, intentions and actions on their beliefs? Why do we need mental states in agents? The answer is in the difference pointed out by Grosz [28] between a mere tool and a collaborator. Let me stress the four main reasons argued in the paper.

Anticipatory coordination. The ability to understand and anticipate the interference of other events with our actions and goals is highly adaptive. This is important both for negative interference and the avoidance of damages, and for positive interference and the exploitation of opportunities. In a MA world the main problem is the predictability of those special events that are the behaviors of other agents. In particular, anticipatory coordination seems strongly based on the "intentional stance" (mind reading), on the ascription of intentions, and on the possibility to change these intentions (desires, goals) in order to change/obtain a given behavior by the others.

Open delegation. In several cases we need local and decentralized knowledge and decision in our agents. The agent delegated to take care of a given task has to choose among different possible recipes; or it has to build a new plan, or to adapt abstract or previous plans to new situations; it has to find additional (local and updated) information; it has to solve a problem (not just to execute a function, an action, or implement a recipe); sometimes it has to exploit its "expertise". In all these cases the agent takes care of our interests or goals autonomously and "remotely" i.e., far from and without our monitoring and intervention (control).

This requires an "open delegation": the delegation "to bring it about that p"; so, the agent is supposed to use its knowledge, intelligence, ability, and to have some autonomy of decision.

So, if we delegate our goal to an agent without eliciting (or requiring) a specified behavior, the agent has to reason, choose, and plan *for* this goal; therefore, we need explicitly represented mental attitudes.

Goal-Adoption and "deep" help. Since the knowledge of the client (i.e., the delegating agent or user) about the domain and about the helping agents is limited (both incomplete and incorrect) it might happen that the delegated task (the specific request, or the elicited behavior) is not so useful for the client, for a number of reasons: the expected behavior would be useful but it cannot be executed, or it is useless, or self-defeating, or dangerous for other goals, or there is a better solution for the client's needs; and perhaps the helping agent is able to provide a better help with its knowledge and ability, going beyond the "literally" delegated task. To be really helpful this kind of agent must be able to recognise and reason about the goals and plans of the client, and to have/generate different solutions (like in open delegation).

Open delegation and over/critical help characterise a collaborator vs a simple tool, and presuppose intelligence, capabilities, and autonomy both in deciding (discretion) and in controlling its own actions. That is precisely what we want to exploit of it.

Influencing. The explicit representation of the agents' minds in terms of beliefs, intentions, etc., allows for reasoning about them, and—even more importantly—it allows for the explicit influencing of others, trying to change their behavior (via changing their goals/beliefs) through high level communication: argumentation, promise, negotiation, etc.

In fact, if agents are heterogeneous and represent the interests of different owners-designers-organisations, why should they automatically or reactively be influenced by others' requests and help them (when not busy)? or why should they systematically "refuse"? They should have some decision function (that implicitly or explicitly presupposes some goal/desire/preference). The influencing agent (client) should give them some hints for this decision, in order to change their behavior. Now, even the simplest mechanism for this (incentives/money/market) is based on some contract/exchange. In fact, how to interact with a stranger/heterogeneous agent without some form of "contract"? But, any contract implies "promises" and trust: x has to "believe" that y will provide what it promised (data? money? service? etc.) and y has to give x this "belief".

10.2. Sociality and individuals

Agents are social because they interfere with and depend on each other. Thus, they are social for multiplying their powers (their possibility to achieve goals), and exploiting actions, abilities, and resources (including knowledge and intelligence) of the others.

AI agents must be social in order to really assist and help the users, and to coordinate, compete, and collaborate with each other.

The basic ingredients of cooperation, exchange, and organization are *Goal delegation* and *Goal-Adoption*.

It is possible to obtain Adoption from an autonomous agent by influencing and power, for example by exploiting its dependence. An agent needs some motives to waste its own resources for another agent. An autonomous agent must always act for its own motives. These motives may be of several kinds: benevolence, rewards, common goal, norms, etc. One should not confuse "self-interested" or "self-motivated" (rational) agents with "selfish" agents.

Modelling individual social action and mind is necessary for modelling collective behavior and organization. The individual social mind is the necessary precondition for society (among cognitive agents). In particular, one cannot understand the real glue of a group or team if one ignores the goals of influencing the others, the collaborative coordination, the commitments, the obligations and rights that relate one to the other. Without this, also the collaboration among artificial agents would be unreliable, fragile and incomplete.

10.3. Mind is not enough

In modelling social agents we need "mind", but mind is not enough. We need also **emergence** of complex unplanned social structures, and of non-orchestrated cooperation.

Emergent intelligence and cooperation do not pertain just to reactive agents. Mind cannot understand, predict, and dominate all the global and compound effects of actions at the collective level. Some of these effects are positive, self-reinforcing and self-organising. There are forms of cooperation which are not based on knowledge, mutual beliefs, reasoning and constructed social structure and agreements.

But what kind/notion of emergence do we need to model these forms of social behavior? A notion of emergence which is simply relative to an observer (which sees something interesting or some beautiful effect looking at the screen of a computer running some simulation) or a merely *accidental* cooperation, are not enough for social theory and for artificial social systems. We need an emerging structure *playing some causal role in the system* evolution/dynamics; not merely an epiphenomenon. This is the case of the emergent dependence structure. Possibly we need even more than this: really *self-organizing emergent structures*. Emergent organisations and phenomena should reproduce, maintain, stabilize themselves through some feedback: either through evolutionary/selective mechanisms or through some form of learning. Otherwise we do not have a real emergence of some causal property (a new complexity level of organisation of the domain); but just some subjective and unreliable global interpretation.

This is true also among cognitive/deliberative agents: the emergent phenomena should feedback on them and reproduce themselves without being understood and deliberated [23]. This is the most challenging problem of reconciliation between cognition and emergence: unaware *social functions* impinging on intentional actions.

10.4. Towards a synthetic paradigm

I believe that in the next decades we will not assist to a paradigmatic revolution, as promised (or threatened) by neural nets, dynamic approaches, social constructivisms, etc.: i.e., connectionism eliminating cognitivism and symbolic models; emergentist, dynamic and evolutionary models eliminating reasoning on explicit representations and planning; neuroscience (plus phenomenology) eliminating cognitive processing; situatedness, reactivity, cultural constructivism eliminating general concepts, context independent abstractions, ideal-typical models.

I personally believe that we are going towards the elaboration of synthetic theories and a general *synthetic paradigm*. Neo-cognitivism and the new AI are the beginning of a highly transformative and adaptive reaction to all these radical and fruitful challenges. They are paving the way for this synthetic research, and are starting the job.

This does not mean simply to be tolerant or to let different candidate-paradigms compete. Nor is it sufficient to propose hybrid models and architectures in which reaction and deliberation, neural nets and symbolic manipulation, reasoning and selection, coexist as competing and cooperating layers. The latter is just one (the simplest) solution. Synthetic theories should explain the dynamic and emergent aspects of cognition and symbolic computation; how cognitive processing and individual intelligence emerge from sub-symbolic or sub-cognitive distributed computation, and causally feedbacks in it; how collective phenomena emerge from individual action and intelligence and causally shape back the individual mind. We need a principled theory which is able to reconcile cognition with emergence and with reactivity.

Reconciling "Reactivity" and "Cognition". We should not consider reactivity as alternative to reasoning or to mental states. A reactive agent is not necessarily an agent without mental states and reasoning. Reactivity is not equal to reflexes. Also cognitive and planning agents are and must be reactive (like in several BDI models). They are reactive not only in the sense that they can have some hybrid and compound architecture that includes both deliberated actions and reflexes or other forms of low level reactions (for example, [33]), but because there is some form of *cognitive reactivity*: the agent reacts by changing its mind, its plans, goals, intentions.

Reconciling "Emergence" and "Cognition". Emergence and cognition are not incompatible: they are not two alternative approaches to intelligence and cooperation, two competitive paradigms. They must be reconciled:

- first, considering **cognition itself as a level of emergence**: both as an emergence *from sub-symbolic to symbolic* (symbol grounding, emergent symbolic computation), and as a transition *from objective to subjective* representation (awareness)—like in our example of dependence relations—and from *implicit to explicit knowledge*;
- second, recognizing the necessity for going **beyond cognition**, modelling emergent unaware, functional social phenomena (for example, unaware cooperation, non-orchestrated problem solving) also among cognitive and planning agents. In fact, *mind is not enough for modelling cooperation and society*. We have to explain how collective phenomena emerge from individual action and intelligence, and how a collaborative plan can be only partially represented in the minds of the participants, and some part represented in no mind at all.

AI can significantly contribute to solve the main theoretical problem of all the social sciences [31]: the problem of the micro–macro link, the problem of theoretically reconciling individual decisions and utility with the global, collective phenomena and interests. AI will contribute uniquely to solve this crucial problem, because it is able to formally model and to simulate at the same time the individual minds and behaviors, the emerging collective action, structure or effect, and their feedback to shape minds and reproduce themselves.

Acknowledgments

I wish to thank Amedeo Cesta, Rosaria Conte, Rino Falcone, and Maria Miceli of the IP-CNR group, since I am just summarising an approach that was collectively developed. Thanks also to the MAAMAW, ATAL and ICMAS communities were I received both encouragement and insightful feedbacks. A special thank to Maria Miceli, Barbara Dunin-Keplicz, and Jaime Sichman for their comments.

References

[1] S. Baron-Cohen, Mindblindness. An Essay on Autism and Theory of Mind, MIT Press, Cambridge, MA, 1995.

[2] J. Bell, Z. Huang, Dynamic goal hierarchies, in: Practical Reasoning and Rationality 2—PRR'97, Manchester, UK, 1997, pp. 56–69.

[3] D. Bobrow, Dimensions of interaction, AI Magazine 12 (3) (1991) 64–80.
[4] A.H. Bond, Commitments, Some DAI insigths from symbolic interactionist sociology, in: AAAI Workshop on DAI, AAAI, Menlo Park, CA, 1989, pp. 239–261.
[5] M.E. Bratman, What is Intention? in: P.R. Cohen, J. Morgan, M.E. Pollack (Eds.), Intentions in Communication, MIT Press, Cambridge, MA, 1990.
[6] C. Castelfranchi, Social Power: a missed point in DAI, MA and HCI, in: Y. Demazeau, J.P. Mueller (Eds.), Decentralized AI, Elsevier, Amsterdam, 1991, pp. 49–62.
[7] C. Castelfranchi, No more cooperation please! Controversial points about the social structure of verbal interaction, in: A. Ortony, J. Slack, O. Stock (Eds.), AI and Cognitive Science Perspectives on Communication, Springer, Heidelberg, 1992.
[8] C. Castelfranchi, M. Miceli, A. Cesta, Dependence relations among autonomous agents, in: Y. Demazeau, E. Werner (Eds.), Decentralized A.I., 3, Elsevier (North Holland), Amsterdam, 1992.
[9] C. Castelfranchi, Guaranties for autonomy in cognitive agent architecture, in: M.J. Woolridge, N. R. Jennings (Eds.), Intelligent Agents I, Springer, Berlin, 1995.
[10] C. Castelfranchi, Commitment: from intentions to groups and organizations, in: Proc. ICMAS-96, S. Francisco, AAAI/MIT Press, Cambridge, MA, 1996.
[11] C. Castelfranchi, Individual social action, in: G. Holmstrom-Hintikka, R. Tuomela (Eds.), Contemporary Theory of Action Vol. II, Kluwer, Dordrecht, 1997, pp. 163–192.
[12] C. Castelfranchi, R. Conte, Emergent functionalitiy among intelligent systems: cooperation within and without minds, AI & Society 6 (1992) 78–93.
[13] C. Castelfranchi, R. Conte, Limits of economic and strategic rationality for agents and MA systems, Robotics and Autonomous Systems (Special Issue on Multi-Agent Rationality) 24 (3–4) (1998).
[14] C. Castelfranchi, R. Falcone, Delegation conflicts, in: M. Boman, W. van De Welde (Eds.), Proc. MAAM AW-97, Springer, Berlin, 1997.
[15] C. Castelfranchi, R. Falcone, Principles of trust for MAS: cognitive anatomy, social importance, and quantification, International Conferences on MAS—ICMAS-98, Paris, July 1998, AAAI/MIT Press, Cambridge, MA, 1998.
[16] J. Chu-Carroll, S.S. Carberry, A plan-based model for response generation in collaborative task-oriented dialogues, in: Proc. AAAI-94, Seattle, WA, 1994.
[17] P.R. Cohen, H.J. Levesque, Rational interaction as the basis for communication, in: P.R. Cohen, J. Morgan, M.E. Pollack (Eds.), Intentions in Communication, MIT Press, Cambridge, MA, 1990.
[18] R. Conte, C. Castelfranchi, Cognitive and Social Action, UCL Press, London, 1995.
[19] R. Conte, C. Castelfranchi, Mind is not enough. Precognitive bases of social interaction, in: N. Gilbert (Ed.), Proc. 1992 Symposium on Simulating Societies, University College of London Press, London, 1996.
[20] D.C. Dennet, Brainstorms, Harvest Press, New York, 1981.
[21] J. Doyle, The foundations of psychology: a logico-computational inquiry into the concept of mind, in: R. Cummins, J. Pollock (Eds.), Philosophy and AI, MIT Press, Cambridge, MA, 1991, pp. 39–78.
[22] B. Dunin-Keplicz, R. Verbrugge, Collective commitments, ICMAS-96, Kyoto, Japan, 1996.
[23] J. Elster, Marxism, functionalism and game-theory: the case for methodological individualism, Theory and Society 11 (1982) 453–481.
[24] J. Ferber, Les Systemes Multi-Agents, InterEditions, Paris, 1995.
[25] S. Franklin, A. Graesser, Is it an agent, or just a program: a taxonomy for autonomous agents, in: J.P. Muller, M.J. Wooldridge, N.R. Jennings (Eds.), Intelligent Agents III, Lecture Notes in Artificial Intelligence, Vol. 1193, Springer, Berlin, 1997.
[26] L. Gasser, Social conceptions of knowledge and action: DAI foundations and open systems semantics, Artificial Intelligence 47 (1991) 107–138.
[27] M.R. Genesereth, S.P. Ketchpel, Software Agents, TR, CSD, Stanford University, 1994.
[28] B. Grosz, Collaborative systems, AI Magazine 17 (1996) 67–85.
[29] B. Grosz, S. Kraus, Collaborative plans for complex group action, Artificial Intelligence 86 (1996) 269–357.
[30] A. Haddadi, K. Sundermeyer, Knowledge about Other Agents in Heterogeneous Dynamic Domains, ICICIS, Rotterdam, IEEE Press, 1993, pp. 64–70.
[31] F.A. Hayek, The result of human action but not of human design, in: Studies in Philosophy, Politics and Economics, Routledge & Kegan, London, 1967.

[32] N.R. Jennings, Commitments and conventions: the foundation of coordination in multi-agent systems, The Knowledge Engineering Review 3 (1993) 223–250.
[33] S. Kurihara, S. Aoyagi, R. Onai, Adaptive selection or reactive/deliberate planning for the dynamic environment, in: M. Boman, W. Van de Welde (Eds.), Multi-Agent Rationality—Proceedings of MAAMAW-97, Lecture Notes in Artificial Intelligence, Vol. 1237, Springer, Berlin, 1997, pp. 112–127.
[34] H.J. Levesque, P.R. Cohen, J.H.T. Nunes, On acting together, in: Proc. AAAI-90, Boston, MA, Morgan, Kaufmann, San Mateo, CA, 1990, pp. 94–100.
[35] M. Luck, M. d'Inverno, A formal freamwork for agency and autonomy, in: Proc. First International Conference on Multi-Agent Systems, AAAI Press/MIT Press, Cambridge, MA, 1995, pp. 254–260.
[36] T.W. Malone, K. Crowston, The interdisciplinary study of coordination, ACN Computing Survey 26 (1) (1994).
[37] M. Mataric, Designing emergent behaviors: from local interactions to collective intelligence, in: Simulation of Adaptive Behavior 2, MIT Press, Cambridge, MA, 1992.
[38] D. McFarland, Intentions as goals, open commentary to Dennet, D.C., Intentional systems in cognitive ethology: the "Panglossian paradigm" defended, Behavioural and Brain Sciences 6 (1983) 343–390.
[39] G. Miller, E. Galanter, K.H. Pribram, Plans and the structure of behavior, Holt, Rinehart & Winston, New York, 1960.
[40] J. Piaget, Etudes sociologiques, Doz, Geneve (3) 1977.
[41] I. Pörn, On the nature of a social order, in: J.E. Festand et al. (Eds.), Logic, Methodology and Philosophy of Science, North-Holland, Elsevier, Amsterdam, 1989, pp. 553–567.
[42] A.S. Rao, M.P. Georgeff, E.A. Sonenberg, Social plans: a preliminary report, in: E. Werner, Y. Demazeau, (Eds.), Decentralized A.I., 3, Elsevier, Amsterdam, 1992.
[43] A.S. Rao, M.P. Georgeff, Modeling rational agents within a BDI-architecture, in: Proc. Second Internat. Conf. Principles of Knowledge Representation and Reasoning, Cambridge, MA, 1991.
[44] C. Rich, C.L. Sidner, COLLAGEN: when agents collaborate with people, in: Proc. Autonomous Agents 97, Marina del Rey, CA, 1997, pp. 284–291.
[45] A. Rosenblueth, N. Wiener, Purposeful and non-purposeful behavior, in: W. Buckley (Ed.), Modern Systems Research for the Behavioral Scientist, Aldine, Chicago, 1968.
[46] J.S. Rosenschein, M.R. Genesereth, Deals among rational agents, in: Proc. IJCAI-85, Los Angeles, CA, 1985, pp. 91–99.
[47] S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, Englewood Cliffs, NJ, 1995.
[48] J. Sichman, Du Raisonnement Social Chez les Agents, Ph.D. Thesis, Polytechnique—LAFORIA, Grenoble, 1995.
[49] M.P. Singh, Social and psychological commitments in multiagent systems, in: Proc. "Knowledge and Action at Social & Organizational Levels", Fall Symposium Series, AAAI Inc., Menlo Park, CA, 1991.
[50] L. Steels, Cooperation between distributed agents through self-organization, in: Y. Demazeau, J.P. Mueller (Eds.), Decentralized AI, North-Holland, Elsevier, Amsterdam, 1990.
[51] R. Tuomela, What is Cooperation, Erkenntnis 38 (1993) 87–101.
[52] R. Tuomela, K. Miller, We-intentions, Philosophical Studies 53 (1988) 115–137.
[53] E. Werner, Social intentions, in: Proc. ECAI-88, Munich, 1988, WG, pp. 719–723.
[54] T.A. Winograd, Language/action perspective on the design of cooperative work, Human–Computer Interaction 3 (1) (1987) 3–30.
[55] M. Wooldridge, N. Jennings, Formalizing the cooperative problem solving process, in: IWDAI-94, 1994, pp. 403–417.
[56] M. Wooldridge, N. Jennings, Intelligent agents: theory and practice, The Knowledge Engineering Review 10 (2) (1995) 115–152.