Modeling Social Action for AI Agents

Cristiano Castelfranchi
Istituto di Psicologia del CNR - Unit of "AI, Cognitive Modelling & Interaction"
v. Marx 15 - 00137 Roma - ITALY
cris@pscs2.irmkant.rm.cnr.it

0 Premise

AI is a science, not merely technology, engineering. It cannot find an identity (*ubi consistam*) 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 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, we have to construct socially intelligent systems to understand it, and we have to build social entities to have intelligent systems. If we want that the computer is not "just a glorified pencil" [Popper, BBC interview), that it is not a simple *tool* but a *collaborator* [Grosz, 1995], 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 intelligences 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 responded to and went out of its crisis. It is one of the way 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, that in the '60s and '70s gave rise to the Cognitive Science, and now will strongly impact on the social sciences.

This stream is part of the new AI of the '90s 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, with hybrid architectures, interfering --either co-operatively or competitively-- with other systems. The new password is *interaction* [Bobrow, 1991]: 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 modeling social intelligence and behavior. In my talk I will just attempt a principled systematization. On the one side, I will illustrate what I believe to be the basic ontological categories for social *action*, *structure*, and *mind*; letting, first, sociality (social action, social structure) emerge bottom-up 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. In this paper I will focus on the bottom-up perspective. On the other side, I will propose some critical reflections on current approaches and future directions. Doing this I will in particular stress five points.

• Social vs. collective

"Social action" is frequently used -in AI, in philosophy- as opposed to individual action, thus as the action not of an individual but of a group, of a team. It is intended as a form of collective activity, possibly coordinated and orchestrated, then tending to joint action. My claim is that *we should not confuse or identify social action/intelligence with the collective one.*

Many of the theories about joint or group action 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.*

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 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 modelling mental states* in agents and to have representations (both beliefs and goals) about the mind of the other agents.

I will stress what non-cognitive agents cannot do at the social level.

• Social action vs. communication

The notion of social action (that is foundational for the notion of Agents) cannot be reduced to communication or modelled on the basis of communication. *Agents are not "agents" in virtue of the fact that they communicate; they are*

not "social" because they communicate (they communicate because they are social). They are social because they act in a common world and because they interfere with, depend on each other, and influence each other.

- **Social action & communication vs. cooperation**

Social interaction (included communication) is not the joint construction and execution of a M-A plan, of a shared script, necessarily based on mutual beliefs. It is not necessarily a cooperative activity [Castelfranchi, 1992]. 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 changing their mind. Both are frequently aimed at blocking, damaging, or aggressing against the others, or at competing with them.

- **Reconciling "Emergence" and "Cognition"**

Emergence and cognition are not incompatible with one another, are not two alternative approaches to intelligence and cooperation, two competitive paradigms.

On the one side, Cognition has to be conceived as a level of emergence (from objective to subjective; from implicit to explicit). On the other side, emergent unaware, functional social phenomena (ex. emergent cooperation, and swarm intelligence) should not be modeled only among sub-cognitive agents [Steels, 1990; Mataric, 1992], but also among intelligent agents. In fact, for a theory of cooperation and society among intelligent agents *mind is not enough*[Conte and Castelfranchi, 1996]. I will stress the limits of deliberative and contracting agents as for complex social behavior: cognition cannot dominate and exhaust social complexity [Hayek, 1967].

I will present *a basic ontology of social action* by examining its most important forms, with special attention to pro-social forms, in particular *Goal Delegation* and *Goal Adoption* that are the basic ingredients of social commitments and contracts, and then of exchange, cooperation, group action, and organization. We need such an analytical account of social action not only for the sake of a good scientific conceptual apparatus (and I wan't believe that from confuse notions and theories good applications can follow). I will give some justification of this analysis in term of its theoretical and practical usefulness for AI systems, arguing against some current biases typical of AI social models.

I will argue 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 mind of the other (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.

Sociality step by step

1 Interference and dependence (1° step)

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

A "common world" means that there is *interference* between actions and goals of the agents: the effects of the action of one agent are relevant for the goals of the other: i.e. they either favour, allow the achievement or maintenance of some goals of the other's (*positive interference*), or threat some of them (*negative interference*) [Haddadi, and Sundermeyer, 1993; Castelfranchi, 1991: Piaget, 1977].

In a **Dependence** relation not only *y* can favour *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.

1.1 An emergent structure and its feedback into the mind

The structure of interference and interdependence among a population of agents is an *emerging* and *objective* one, *independent of the agents' awareness and decision*, but it constrains the agents' actions determining their success and efficacy. However, this pre-cognitive structure can "cognitively emerge": i.e. part of these constraints can become known: the agents have beliefs about their dependence and power relations.

Either through blind learning (reinforcement) or through this "understanding" (cognitive emergence) the objective emergent structure of interdependencies feedback into the agents' mind: it will change them. Some goals or plans will be drop as impossible, others will be activated or pursued as possible [Sichman, 1995]. Moreover, new goals and intention will rise: social goals. The goal of exploiting or waiting for an action of the other; some goal of blocking or aggressing the other, or of letting or helping it to do something; the goal of influencing the other to do or not to do something (ex. request); the goal of changing dependence relations. These new goals are strictly a consequences of dependence.

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

1.2 Basic moves

Let me first discover sociality from *x*'s (the agent subject to interference) point of view. From its selfinterested perspective, in *interference* and *dependence* an agent *x* has two alternatives:

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.

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

Table 1

Column A represents "mere coordination" (*negative* and *positive*); column B "influencing"; raw 2 "delegation". In both cases (A & B) we possibly have "social action" by *x*, but of a very different nature. And we have "social action" (SA) only at some specific conditions.

2 From non-social action to weak social action: beliefs about the other's mind (2° step)

A SA is *an action that takes into account 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 is needed [Dennett, 1981].

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). Agent's action doesn't change its nature depending on the objective nature of the obstacle. If *x* acts towards another agent *y* as it were just a physical object her action *is not a SA*. Whether it is a social action or not depends on how *x subjectively* considers *y* in her plan. Consider the same situation but with some more pro-active than reactive attitude by *x*: *x* foresees that *y* will cross her road on the basis of her beliefs about *y*'s goals; like in traffic, when we slow down or change our way because we understand the intention of the driver preceding us 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 later).

So, *an action related to another agent is not necessarily social*. Also the opposite is true. A merely practical action, not involving other agents, may be or become social. 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 (figure 1). Nothing social in this.

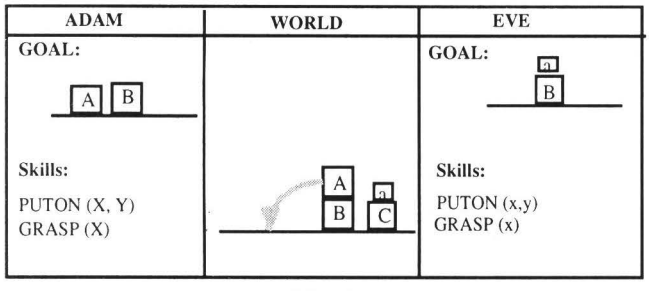

Fig. 1

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: EVE is *dependent on* ADAM, since if ADAM performs the needed action EVE will achieve her goal [Castelfranchi *et al.*, 1992]. Now, suppose that ADAM has no personal goals and knowing about EVE's goals and abilities decides to help EVE, and 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 -although performed just on blocks- because the end-goal of this action, its motive, is to let EVE achieve her goal, and is based on beliefs about EVE's goals.

I call "weak SA" that based just on *social beliefs*: beliefs about other agents' minds or actions (like in the car examples, and like in mere coordination, see later); and "strong SA" that also directed by *social goals*. [1]

The true basis of any level of SA among cognitive agents is *mind-reading* [Baron-Cohen, 1995]: 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 behavior. In other words, the other's behavior becomes a "sign" for the agent: a sign of the other's mind. This understanding, this behavioral and implicit communication is, before strict communication (special message sending), the true basis of reciprocal coordination and collaboration [Rich and Sidner, 1997]. Differently from current machines, we 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 we let the other do the same.

Communication, Agents, and Social Action

It is common sense in AI that "social agents" are equal to "communicating agents". According to many students communication is a necessary feature of agency (in the AI sense) [Jennings and Wooldridge, 1995; Genesereth and Ketchpel, 1994; Russell and Norvig, 1995]. 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: cooperative or aggressive [Castelfranchi, 1992]). Communication is also a type of SA aimed at giving beliefs to the addressee. This is a true and typical Social Goal, since the intended result is about a mental state of another agent. Notice that this typical SA does not necessarily involve any "sharing"; in fact, contrary to common sense,

[1] A definition of SA, communication, adoption, aggression, etc. is possible also for non-cognitive agents. However those notion must be goal-based. Thus, a theory of goal-oriented (not "goal-directed") systems and of implicit goals is needed [Conte e Castelfranchi, 1995, cap.10]. However, there are levels of sociality that cannot be attained reactively (see later).

communication is not necessarily truthful, and *x* can either believe or not believe what she is communicating to *y*: also lies are communication.

Communication in fact 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 is not based on communication (since it does not necessarily require agreement). Of course, strict bilateral cooperation is based on *agreement* and requires some form of communication.

To conclude, my claim is that SA *is not grounded on Communication*.

3 Principles of coordination

In simple coordination (column A) *x* is just coordinating her behavior with the perceived or predicted behavior of *y*, ignoring the possibility to change it; like in our first example of car avoidance, *x* changes her own plan (sub-goal) and elaborates a new goal which is based on her beliefs about *y*'s goal (weak SA). One might call "coordination" quite all forms of social interaction (including negotiation, cooperation, conflict, etc..) [Malone and Crownston, 1994], while I prefer to restrict the use to this simpler form, in which there is merely coordination without influencing or communication.

3.1 Reactive vs. anticipatory: coordination among cognitive agents

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

- **reactive coordination**, is based on the direct perception of an obstacle or opportunity and on a reaction to it;
- **proactive** or **anticipatory coordination**, lies on the anticipation either based on learning or on inferences (prediction) 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: prediction gives more time. In a sense a completely successful avoidance coordination cannot really be done without some anticipation. When the obstacle/damage is directly perceived it is - at least partially - "too late"; either the risk is higher or there is already some loss.

Anticipatory coordination with very complex and long term effects, needs some theory or model: i.e. some cognitive intelligence. Anticipatory coordination with cognitive goal-directed agents cannot be based just on learning or inferences about trajectories or the frequencies of action sequences. Under this respect, since agent combine their basic actions in several long and creative sequences, the prediction (and then the anticipatory coordination) must be based on *mind-reading*: on the understanding of the goals and the plan of the other [Bratman, 1990]. Conflicts or opportunities are detected comparing their 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" (also partially) its behavior in a M-A world without some anticipatory coordination. There is a co-evolutionary coupling between planning in a M-A world and 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 cannot do.* They could at most have some -learned, built in, or inherited- reaction to some short-term behavioral fixed sequence.

3.2 Positive and negative coordination; unilateral, bilateral, and mutual

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

In **unilateral coordination** only *x* is coordinating her own activity relative to *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 teither agent does not understand the new coordinated plan of the other there will be some trouble. The bilateral coordination is **mutual** when both the agents are aware of their coordination intentions and they try to arrive at some (implicit) agreement. Mutual coordination necessarily requires some collaborative coordination.

3.3 Selfish vs. collaborative coordination

All the previous ones (Table 1 column A) are the basic forms of the ***ego-centred*** or ***selfish coordination***: *x* tries to achieve her own goal 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 ex. *x* might continue to modify, adapt her own behavior but in order to avoid negative interference in the other's action or to create positive interferences. This is ***Collaborative Coordination***: *x* is adapting her behavior trying to favour *y*'s actions [Piaget, 1977]. However the Collaborative coordination is 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 later). The collaborative coordination is the basis of Grosz and Kraus' "intention that" [Grosz and Kraus, 1996].

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

4 Relying on (Delegating) - (3° step)

There are basic forms of SA that are the ingredients of help, exchange, cooperation, and then of partnership, groups and

team work. We will 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 achieves her goal relying on the other's action) -I will 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) - I will 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 y. So, she is constructing a Multi-Agent plan and *y* has a share in this plan: *y*'s delegated *task* is either a state-goal or an action-goal [Castelfranchi and Falcone ,1997].

If EVE is aware of ADAM's action, she is *delegating* ADAM a task useful for her:

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

These conditions define EVE's "**trust**" in ADAM.

There are three basic kinds of Delegation or Reliance (let me expand raw 2 of Table 1) :

4.1 From non-social to social delegation

Unilateral Reliance (weak delegation)

In Unilateral Delegation there is no bilateral awareness of the delegation, no agreement: *y* is not aware of the fact that *x* is exploiting her action. One can even "delegate" some task to an object or *tool*, relying on it for some support and result [Luck and D'Inverno, 1995; Conte e Castelfranchi, 1995, cap.10].

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; 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, inducing the desired *y*'s behavior to exploit it. Depending on the reactive or deliberative character of *x*, the induction is just based on some stimulus or is based on beliefs and complex types of influence.

As an example of unilateral Delegation by induction consider now a fisherman: differently from the hunter example, 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 in order to fully understand this important and more social form of Delegation (based on social goals) we need a good notion of Social Goal-Adoption (see later) and we have to characterise not only the mind of the delegating agent but also that of the delegated one, in a "contract".

Even more important for a theory of collaborative agents are the *levels* of delegation.

4.2 Plan-based levels of delegation

Given a goal and a plan (sub-goals) to achieve it, *x* can delegate goals/actions (tasks) at different level of abstraction and specification [Falcone and Castelfranchi, 1997]. We can distinguish between several levels, but the most important are the following ones:

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

The object of delegation can be minimally specified (***open delegation***), completely specified (***close delegation***) or specified at any intermediate level. We wish to stress that *open delegation* is not only due to *x*'s preference, practical ignorance or limited ability. Of course, when *x* is delegating a task to *y*, she is always *depending on y* for that task: she needs *y*'s action for some of her goals (either domain goals or more general ones, like saving time, effort, resources and so on). However, open delegation is also due to *x*'s ignorance about the world and its dynamics: *fully specifying a task is often impossible or not convenient,* because some local and updated knowledge is needed in order for that part of the plan to be successfully performed. Open delegation is one of the bases for the *flexibility* of distributed and MA plans.

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

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* [Grosz and Kraus, 1996]. 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.

5 *Strong SA*: goals about the other's action/goal (4° 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 has also the goal that y has the goal (more precisely intends) to do that action*. I call this "**cognitive delegation**": delegation to an intentional agent. This goal of *x* is the motive for *influencing y* [Porn, 1989; Castefranchi, 1991], but it does not necessarily lead to inducing or influencing *y*. The world may by itself realise our goals. In fact, 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 considered as a cognitive agent: *a social goal is a goal about other agents' minds or actions* (like in EVE's example). 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.

For a cognitive autonomous agent to have a new goal, he ought to acquire some new *belief* [Castelfranchi, 1995]. Therefore, cognitive influencing consists of providing the addressee with information that is pretended to be relevant for some of his goals, and this is done in order to ensure that the recipient has a new goal.

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 something? Of course, normally -but not necessarily- by communicating.

However, communication can only inform the other about our goals and beliefs about its action: *why should he care about our goals and expectations?* He is not necessarily a benevolent agent, 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, are others). However the most important basis of our power is the fact that probably 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 [Castelfranchi, 1991].

Incentive engineering, manipulating the other's utility function, is not the only way we have to change the mind (behavior) of the other agent. In fact in a cognitive agent pursuing or abandoning a goal does not depends only on preferences and on beliefs about utility. To pursue or abandon his intention, *y* should have a host of beliefs, that are neither reducible nor related 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 of incentive engineering, but of modifying the beliefs which "support" goals and intentions and provide reasons for behavior. Beliefs about incentives represent only a sub-case.

6 *Strong SA*: Social Goal Adoption (5° step)

Let me now look at SA from *y*'s (the contractor, 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) frequently enough its role in cooperation in not understood.

Either agents are just presupposed to have the same goal [ex. Werner, 1988], or the adoption of the goal from the other partners is not explicitly accounted for [Tuomela and Miller, 1988; Levesque *et al.* 1990; Tuomela, 1993]; or the reasons for adopting the others' goal and take part in the collective activity are not explored.

In **G-Adoption** *x is changing her mind: she 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 *y* wants to achieve this goal: *x* knows this and decides to make/let him achieve it. *x* comes to have the same goal of *y*, because she knows that is *y*'s goal but not as in simple imitation: here *x* has the goal that p (wants p to be true) in order for *y* to achieve it. In other words, *x is adopting a goal of y's when x wants y to obtain it as long as x believes that y wants to achieve that goal* [Conte and Castelfranchi, 1995].

Among the various forms of G-Adoption, especially for modelling agreement, contract and team work, *G-Adhesion* or *Compliance* has a special relevance. That occurs when the G-Adoption is due to the other's request (implicit or explicit), to his goal that *x* does a given action, or better to his goal that *x* adopts a given goal. It is the opposite of spontaneous forms of G-Adoption. So in *Adhesion x* adopts *y*'s goal that she adopts, she complies with *y*'s expectations.

G-Adhesion is the strongest form of G-Adoption. Agreement is based on adhesion; strong delegation is request for adhesion. In negotiation, 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 ex. obeys) G-Adhesion is what really matters.

6.1 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 born all as 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 [Castelfranchi, 1995].

In architectural terms this means that there is not an unique origin of potential intentions [Rao and Georgeff, 1991] or candidate goals [Bell and Huang, 1997]. There are several goal origins or sources (bodily needs ; goals activated by beliefs; goals elicited by emotions; goals generated by practical reasoning and planning; and goals adopted: introjected from outside). All these goals have to converge at a given level in the same path, in the same goal processing, to become *intentions* and be pursued through some action.

6.2 Motivation for G-Adoption

Adoption does not coincide with *benevolence* [Rosenschein and Genesereth, 1985]. A relation of benevolence, indeed, is a form of generalised adoption. This has to do with the motivation for G-Adoption.

Benevolence is a ***terminal*** (non instrumental) form of G-Adoption (pity, altruism, love, friendship). *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: *x* adopts *y*'s goal since she is co-interested in (some of) *y*'s intended results: they have a common goal. Collaborative coordination (3.3) is just one example of it.

The distinction between these three forms of G-Adoption is very important, since their different motivational basis (why *x* adopts) allows important predictions on *x*' s "cooperative" behavior. For example, if *x* is a rational agent, in social exchange she should try to cheat, not reciprocating *y*'s adoption. On the contrary, in cooperative adoption *x* normally is not interested in free riding since she have the same goal as *y* and they are *mutually dependent* on each other as for this goal p: both *x*'s action and *y*'s action are necessary for p, so *x*'s damaging *y* would damage herself. Analogously, while in terminal and in cooperative adoption it might be rational in many cases to inform *y* about difficulties, obstacles, or defections [Levesque *et al.*, 1990; Jennings, 1993], 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.).

6.3 Levels of collaboration

In analogy with delegation, several dimensions of adoption can be characterized [Falcone and Castelfranchi, 1997]. 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 just doing what is literally requested to do, is not a very collaborative agent. She has no initiative, does not care of our interests, does not use her knowledge and intelligence to correct our plans and requests that might be incomplete, wrong or self-defeating.

A truly helpful agent should care of our goals and interests going beyond our delelegation and request [Chu-Carroll and Carberry, 1994]. But, *only cognitive agents can non-accidentally help beyond delegation*, recognizing our current needs case by case.

Of course, there are dangers also when the agent takes the initiative of helping us beyond our request. Troubles either due to misunderstandingfs and wrong ascriptions, or to conflicts and paternalism.

7 Social Goals as the glue of Joint Action: Social-Commitment

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 here examine 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 [Tuomela and Miller, 1987; Levesque *et al.* , 1990; Rao *et al.*, 1992; Grosz and Krauss, 1996; Wooldridge and Jennings, 1994]. I would like just to stress how social action and goals, as previously characterised, play a crucial role in it.

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" [Grosz and Kraus, 1996] 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.
c) the *social commitment* of y to x as for a, which is a form of goal-adoption or better of adhesion.

Normally, 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 those accounts. They mainly rely on the agents' beliefs about the intentions of the others; i.e. a weak form of social action and mind. The same is true for the notion of cooperation in Game Theory. As for the social commitment it has been frequently confused with the individual (non social) commitment of the agent to his task.

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 [Cohen & Levesque 1990]. It refers to *a relation between an agent and an action.* The agent has decided to do 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* [Singh, 1992; Castelfranchi, 1996]. 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.

Social commitment is also different from **Collective** or Group Commitment [Dunin-Keplicz and Verbrugge, 1996] . 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). So Social Commitment *creates rights and duties* among x and y [Castelfranchi, 1996].

Although so relevant (and although it introduces some normative aspects) the social commitment structure is not the only important structure constraining the organizational activity and society.

8 Social structures and organization

There is an implicit agreement about organizations in recent computational studies. Either in DAI theories of organization [Bond, 1989; Gasser 1991], or in formal theories of collective activity, team or group work, joint intention, and "social agents" [ex. Levesque *et al.*, 1990], or in CSCW approaches to cooperation [Winograd, 1987], organization is in fact accounted for by means of the crucial notion of "commitment". However, this account 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: there is no theory of "social" commitment as a necessary premise for a theory of collective or group commitment, and normative aspects of commitment are ignored;

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 cooperations.

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 group*, team work, and coordination, just based on agents' beliefs and intentions, and *the "conversational" view of Organization* [Winograd, 1987], 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*) [Bond, 1989; [Gasser, 1991]. There is also a conspicuous lack of attention to the individual motivations to participate in groups and organizations: agents are supposed to be benevolent and willing to coooperate with each other.

Coordination in a group or organization is not guaranteed only by a shared mind (joint intentions, agreed plans, shared beliefs), reciprocal benevolence, and communication; there are several *structures* in any M-A system: the interdependence and power structure; the acquaintance structure emerging from the union of all the personal acquaintances of each agent [Ferber 1995; Haddadi and Sundermeyer, 1993]; the communication structure (the global net of direct or indirect communication channels and opportunities); the commitment structure, emerging from all the Delegation-Adoption relationship and from partnership or coalitions formation among the agents; the structure determined by pre-established rules and norms about actions and interactions. Each structure determines 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 of committing themselves as they like: their are conditioned by their dependence and power, by

their knowledge, by their possible communication, by their roles and commitments, by social rules and norms.

9 Some concluding remarks and challenges

Why are agents social? Because they interfere with and depend on each other. Thus, to multuply their powers (their possibility to achieve goals); to exploit actions, abilities, and resources (included knowledge and intelligence) of the others.

Why should AI agents be social? To really assist and help the users, and to coordinate, compete and collaborate with each other.

Why do we need cognitive, intelligent, autonomous agents acting on our behalf? In order to do *Open delegation* exploiting local knowledge and adaptation, personal expertise and intelligence, and in order to receive *Over* and *Critical Help* case by case: the deepest form of cooperation; that a reactive (although learning) agent cannot provide.

Which are the basic ingredient of cooperation, exchange, organization? *Goal delegation* and *Goal-Adoption*. How to obtain Adoption from an autonomous agent? By influencing and power. Why should it waste its own resources for another agent? Always for its own motives (autonomy) but of several kinds: benevolence, advantages, common goal, norms, etc. One should't mix up "self-interested" (rational) with "selfish".

Why modeling individual social action and mind is necessary for modelling collective behavior and organization? Because 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 ignore the goals of coordination and influencing, the commitments, the obligations and rights relating one to another. Without this, the collaboration among artificial agents will be unreliable, fragile and incomplete.

Why do we need emergent functional cooperation also among intelligent planning agents? Emergence does not pertain only to reactive agents. Mind cannot understand, predict, and dominate all the global and compound effects at the collective level. Some of these effects are positive and self-organising. Mind is not enough: not all cooperation is based on knowledge, mutual beliefs, reasoning and constructed social structure and agreements.

What kind/notion of Emergence do we need?
An emergence 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 [Mataric, 1992]? (like stars "cooperate" to the emergence of our beautiful constellations) are not enough. 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 mechanism 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 [Elster, 1982]. This is the most challenging problem of reconciliation between cognition and emergence: unaware social functions impinging on intentional actions.

Acknowledgement

I wish to thank Amedeo Cesta, Rosaria Conte, Rino Falcone, Maria Miceli of the IP-CNR group, since I'm just summarising an approach that was collectively developed. Thanks also to the MAAMAW, ATAL and ICMAS communities were it was possible to explore around in AI social theory and systems, receiving both encouragement and insightful feedbacks.

References

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

[Bell and Huang, 1997] J. Bell and Z. Huang. Dynamic Goal Hierarchies. Practical Reasoning and Rationality 2 - PRR'97. 56-69. Manchester, UK, 1997.

[Bobrow, 1991] D. Bobrow. "Dimensions of Interaction", *AI Magazine*, 12, 3,64-80,1991.

[Bond, 1989] A. H. Bond, Commitments, Some DAI insigths from Symbolic Interactionist Sociology. *AAAI Workshop on DAI.* 239-261. Menlo Park, Calif.: AAAI, Inc. 1989.

[Bratman, 1990] Bratman, M.E. What is Intention? In P.R. Cohen, J. Morgan, and M. E. Pollack (eds.) *Intentions in Communication.* MIT Press.1990

[Castelfranchi, 1991] C. Castelfranchi. Social Power: a missed point in DAI, MA and HCI. In Y. Demazeau & J.P.Mueller (eds), *Decentralized AI.* 49-62. Amsterdam: Elsevier, 1991.

[Castelfranchi, 1992] 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.

[Castelfranchi *et al.*, 1992] C. Castelfranchi., M. Miceli, A. Cesta. Dependence Relations among Autonomous Agents, in Y.Demazeau, E.Werner (Eds), *Decentralized A.I. - 3*, Elsevier (North Holland), 1992.

(Castelfranchi, 1995) C, Castelfranchi, Guaranties for Autonomy in Cognitive Agent Architecture. In [Woolridge and Jennings, 1995]

[Castelfranchi, 1996] Castelfranchi, C., Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'96*, S.Francisco, June 1996, AAAI-MIT Press.

[Castelfranchi and Conte, 1992] C. Castelfranchi and R. Conte. Emergent functionalitiy among intelligent systems: Cooperation within and without minds. *AI & Society, 6,* 78-93, 1992.

[Castelfranchi and Falcone, 1997] C Castelfranchi and R Falcone. Delegation Conflicts. In M. Boman and W. van De Welde (Eds.)*Proceedings of MAAMAW* '97, Springer-Verlag, 1997.

[Chu-Carroll and Carberry, 1994] J. Chu-Carroll and S..S. Carberry. A Plan-Based Model for Response Generation in Collaborative Task-Oriented Dialogues in *Proceeedings of AAAI-94*. 1994.

[Cohen and Levesque, 1990] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. in P R Cohen, J Morgan and M E Pollack (Eds): *Intentions in Communication*. The MIT Press, 1990.

[Conte and Castelfranchi, 1995] Conte,R. and Castelfranchi, C. *Cognitive and Social Action*, UCL Press, London, 1995.

[Conte and Castelfranchi, 1996] R. Conte and C. Castelfranchi. Mind is not enough. Precognitive bases of social interaction. In N. Gilbert (Ed.) *Proceedings of the 1992 Symposium on Simulating Societies*. London, University College of London Press, 1996.

[Dennet, 1981] Dennet, Daniel.C. *Brainstorms*. Harvest Press, N.Y.

[Dunin-Keplicz and Verbrugge, 1996] B. Dunin-Keplicz and R. Verbrugge. Collective Commitments. *ICMAS'96*, Kyoto, Japan

[Elster, 1982] J. Elster. Marxism, functionalism and game-theory: the case for methodological individualism. *Theory and Society* 11, 453-81.

[Falcone and Castelfranchi, 1997] R Falcone and C Castelfranchi. "On behalf of ..": levels of help, levels of delegation and their conflicts, *4th ModelAge Workshop*:"Formal Model of Agents", Certosa di Pontignano (Siena),1997.

[Ferber, 1995] J. Ferber. *Les Systemes Multi-Agents*. InterEditions, iia, Paris, 1995

[Gasser, 1991] L. .Gasser. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence* 47: 107-138.

[Genesereth, and Ketchpel, 1994] M.R. Genesereth and S.P. Ketchpel, S.P. 1994. *Software Agents*. TR, CSD, Stanford University.

[Grosz, 1995] B. Grosz, Collaborative Systems. AI Magazine, summer 1996, 67-85.

[Grosz and Kraus, 1996] B. Grosz and S. Kraus. Collaborative plans for complex group action, *Artificial Intelligence* 86, pp. 269-357, 1996.

[Haddadi and Sundermeyer, 1993] A. Haddadi and K. Sundermeyer. Knowledge About Other Agents in Heterogeneous Dynamic Domains. ICICIS, Rotterdam,1993, IEEE Press: 64-70.

[Hayek, 1967] F.A. Hayek, *The result of human action but not of human design*. In *Studies in Philosophy, Politics and Economics*, Routledge & Kegan, London, 1967.

[Jennings, 1993] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review* 3, 1993: 223-50.

[Levesque *et al.*, 1990] Levesque H.J., P.R. Cohen, Nunes J.H.T. On acting together. In Proceedings of the 8th National Conference on Artificial Intelligence, 94-100. Kaufmann. 1990

[Luck and d'Inverno, 1995] M. Luck and M. d'Inverno, "A formal freamwork for agency and autonomy". In proceedings of the First International Conference on Multi-Agent Systems, 254-260. AAAI Press/MIT Press, 1995.

[Malone and Crowston, 1994] T.W. Malone and K, Crowston, The interdisciplinary study of coordination. ACN Computing Survey, 26, 1, 1994.

[Mataric, 1992] M. Mataric. Designing Emergent Behavioors: From Local Interactions to Collective Intelligence. In *Simulation of Adaptive Behavior 2*. MIT Press. Cambridge,

[Piaget, 1977] J. Piaget. Etudes sociologiques, Doz, Geneve, 19977 (3)

Pörn, I. 1989. On the Nature of a Social Order. In J.E. Festand et al. (eds.) *Logic, Methodology and Philosophy of Science*, North-Holland: Elsevier; 553-67.

Rao,A. S., Georgeff, M. P., and Sonenberg, E. A. 1992. Social Plans: A preliminary Report. In E. Werner, and Y. Demazeau. eds. *Decentralized A. I. 3*. Amsterdam: Elsevier.

[Rao and Georgeff, 1991] A S Rao and M P Georgeff: Modeling rational agents within a BDI-architecture. In *Principles of Knowledge Representation and Reasoning*, 1991.

[Rich and Sidner, 1997] Ch. Rich and C L Sidner. COLLAGEN: When Agents Collaborate with People. In Proceedings of *Autonomous Agents 97*, Marina Del Rey, Cal., pp. 284-91

[Rosenschein and Genesereth, 1985] J.S. Rosenschein and M.R. Genesereth. Deals Among Rational Agents. In *Proceedings of IJCAI-85*, Los Angeles, CA. AAAI Press, pp. 91-99.

[Russell and Norvig, 1995] S.J. Russell and P. Norvig *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.

[Sichman, 1995] J Sichman, Du Raisonnement Social Chez les Agents. PhD Thesis, Polytechnique - LAFORIA, Grenoble

[Sing, 1991] M.P. Singh, Social and Psychological Commitments in Multiagent Systems. In Preproceedings of "Knowledge and Action at Social & Organizational Levels", Fall Symposium Series, 1991. Menlo Park, Calif.: AAAI, Inc.

[Steels, 1990] L. Steels. Cooperation between distributed agents through self-organization. In Y. Demazeau & J.P. Mueller (eds.) *Decentralized AI* North-Holland, Elsevier, 1990.

[Tuomela, 1993] Tuomela, R. What is Cooperation. *Erkenntnis*, 38, 1993, 87-101

[Tuomela and Miller, 1988] R. Tuomela and K. Miller. "We-Intentions", *Philosophical Studies*, 53, 1988, 115-37.

[Werner, 1988] E. Werner. Social Intentions. In *Proceedings of ECAI-88*, Munich, WG, 719-723. ECCAI.

[Winograd, 1987] T. A. Winograd. Language/Action perspective on the Design of Cooperative Work. In *Human-Computer Interaction* 3, 1: 3-30.

[Wooldridge and Jennings, 1994] M. Wooldridge and N. Jennings. Formalizing the cooperative problem solving process. In *IWDAI-94*, 403-17, 1994.

[Wooldridge and Jennings, 1995] M. Wooldridge and N. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2): 115-52. 1995.