AI & Soc (1992) 6: 78–93

AI & SOCIETY

Open Forum

Emergent Functionality Among Intelligent Systems: Cooperation Within and Without Minds

Cristiano Castelfranchi and Rosaria Conte
Istituto di Psicologia-CNR, Rome, Italy

Abstract. In this paper, the current AI view that emergent functionalities apply only to the study of subcognitive agents is questioned; a hypercognitive view of autonomous agents as proposed in some AI subareas is also rejected. As an alternative view, a unified theory of social interaction is proposed which allows for the consideration of both cognitive and extracognitive social relations. A notion of functional effect is proposed, and the application of a formal model of cooperation is illustrated. Functional cooperation shows the role of extracognitive phenomena in the interaction of intelligent agents, thus representing a typical example of emergent functionality.

Keywords: Cooperation; DAI; Emergent functionality

"Emergent Functionality" is not Incompatible with Cognitive Agents

Current literature on "emergent functionalities" is usually about systems that we call "sub-cognitive". By subcognitive, we mean either "reactive systems" (Agre and Chapman, 1987; Agre, 1989; Brooks, 1989), that do not calculate the utility of their actions, nor plan, *stricto sensu*, to obtain what they realize; or, "sub-symbolic agents", acting on a neural network base.

In our opinion, the association between emergent functionality and subcognitive systems is restrictive: even the actions of cognitive agents give rise to unpredicted effects, which sometimes prove to be functional. This is a truism if one consider routines and other reactive actions that all systems share to some extent. Less obvious is the fact that planned actions in the true sense may produce outcomes far beyond any agent's prediction and understanding.

Indeed, it seems misleading to propose two alternative models of "intelligence" (Steels, 1990), one pointing to emergent functional properties of dynamic systems, where problem solving and social behaviour are wholly extramental, and the other pointing to a cognitive agent with far-reaching predictions and fully rational calculations and decision-making. The latter shows a *hypercognitive view of autonomous agents*, a view which favours, but is also derived from, the theoretical opposition set out above.

In this paper, we question both this opposition and the hypercognitive view of intelligent autonomous systems. We argue for a unified view, where functional effects are allowed to emerge from cognitive agents in interaction.

Hypercognitive Agents

The hypercognitive view dominates especially within logic-based approaches to cognitive modeling. We characterise it as follows:

Omniscience: All consequences logically implied by any agent's beliefs are also believed by that agent (there are several attempts at mitigating omniscience, in consideration of both cognitive plausibility and computational tractability (for a well-known example, cf. Fagin and Halpern, 1985).

Introspection: Agents meta-believe everything they believe and want.

Mental transparence: Agents share terminological knowledge and almost all factual knowledge. They interact under condition of mutual knowledge (any agent knows what the others believe and want, and knows that others know that s/he knows).

Lack of an evolutionary perspective: Agents are so fully aware of the conditions under which they interact, that the evolutionary steps of social actions are substantially ignored.

Social subjectivism: Social relationships are investigated only in as much as they are mentally represented, that is, only starting from what are considered as "social" goals and beliefs (agent believes and wants what others believe and want).

Emphasis on communication: As a consequence of previous features, social action is only conceived of in terms of communication, aimed at modifying mutual beliefs and goals (Galliers, 1988; Werner, 1989; Cohen and Levesque, in press).

Society is out There, and not Only in the Mind

For a general and explanatory theory of social action, we believe it is necessary to study extracognitive social relations, which allow to predict and explain social interactions of agents, and provide premises and evolutionary steps of cognitive social relationships. More specifically, the following aspects need be considered:

a) *precognitive bases of social interaction*, such as the relations of dependence, interest, power, communality and sharing of interests and/or goals (see, for a preliminary work in this direction, Castelfranchi, 1990; Conte & Castelfranchi, 1991);
b) *emergent* (extracognitive) *functionalities* of actions intended and planned by cognitive agents.

In this paper, we focus on the second aspect, and in particular on "functional cooperation". Obviously, not all emergent properties, or systemic effects, are cooperative and useful for the agents involved.

Levels of Social Action

The effects of action may be:

a) *accidental*: let **x**'s action be (**PUTON bl b2**). If, say, **y**'s want is (**FREE b2**), **x** will step on **y**'s toes. In such a case, action is not really social.
b) *finalistic*: the social effect is represented as a goal (internal goal) in **x**'s mind (he knows and wants to produce it). Action is social.
c) *functional*: the effect of action is not finalistic, but is not accidental either: the action is undertaken precisely because it produces that effect (external goal). The action is then *functional* to the effect produced.

About Functions

The functional category is essential to the understanding and modeling of social interaction. A well-known example of the use of functional categories is the theory of evolution, where phenomena of study are traced back to natural selection. Think of the grass bug, which is camouflaged against the background: this has the functional effect that the animal is hidden from predators. It could hardly be said that, when its colour changes, the bug has the goal to avoid predation.

After decades of indiscriminate utilisation, the notion of function has undergone a great deal of criticisms and manipulation both in biological and social sciences, and now it has come up once again in many different fields (systems theory, AI, sociology, etc.).

We define a *function* of a trait or behaviour as a selective effect; more explicitly:

Let **x** be an entity which is instantiated in a *sequence of distinct repetitions* (**x1**; **x2**; . . . ; **xn**). A sequence of repetitions is defined as a set of occurrences, or instances, of the same entity all *genetically linked* to one another, that is, linked in such a way that each is produced by the preceding, if any, and produces the following occurrence in the sequence, if any, thanks to a *mechanism of reproduction* whatsoever;

Let also **Bx** be a set of *behaviours* and *characters* of **x**, and

some items in **Bx** produce *effects* (states of the world) *unintended* by, and unknown to, **x**;

We will say that any item in **Bx** that produces unintended effect is *functional*, if that effect acts through a *causal feedback loop* on the mechanism of reproduction, favouring **x**'s reproduction, and as a consequence that of the item itself. The effect is no longer a simple one among others, but is a function of the behaviour or character in question.

In our opinion, all basic social concepts (cooperation, competition, aggression, communication, etc.) can be used with regard to cognitive as well as subcognitive agents (for instance, in animal communication and interaction), provided that the finalistic effect (effect represented in the agent's mind) be replaced with a functional one. Only if this is done will a general theory of social interaction be developed.

Take the biological world; Let *communication* be defined as that behaviour which produces an effect (be it finalistic or functional) that another organism comes to have a new belief. Suppose that a lion is roaring to a second lion and, as a consequence, a frightened gazelle runs away: the lion's roaring is not communicative towards the gazelle since it is purely accidental. In fact, however vital the message proves for the gazelle, the lion had no goal to communicate with the gazelle. Nor can we say that the gazelle's survival has selected the lion's roar. On the other hand, the effect produced on the second lion is a communicative one: signals of different sorts are selected by the information transmitted and the reactions consequently produced on other individuals of the same species.

Many examples of functional *adoption* are found in the biological world, especially among insects, e.g. solider ants which "sacrifice" their lives to save their larvae. Their actions are not planned for this purpose, but are selected by the advantages that their genes obtain – in terms of "inclusive fitness".

A general theory of social interaction (human, animal, and artificial) should and could be worked out, such that it includes both *intentional* and *functional* actions and relations, internal and external social goals. To be fully adequate, in sum, such a theory should account for cultural and social functions (like those studied by anthropology and sociology). Indeed, functional cooperation (see next section) among cognitive agents is due to the selective pressure of social systems rather than to the role of biological functions.

From Accidental to Functional Cooperation

In the following we will address cooperation as an example of social interaction evolving from precognitive, accidental cooperation (Conte, Miceli and Castelfranchi, in press), to extracognitive, functional cooperation.

Unifying Subjective and Objective Cooperation

The hypercognitive approach in AI precludes a unified theory of cooperation. If society is exclusively placed in the mind, it is not possible to unify intentional and functional cooperation.

Indeed, functional cooperation is not only relevant to comparison with other species, but is also of extreme importance within human societies. Besides, cooperation in complex differentiated systems (human–computer and human–human via computer), is largely functional: it is a cooperation among social roles rather than one decided and negotiated by agents.

There are several models of functional cooperation in DAI: from market-like models (Malone, Fikes and Howard, 1988), to models of negotiation (Rosenschein and Genesereth, 1988) sometimes enriched by an intermediary (see, Martial, 1990), and from blackboard architectures (Fennell and Lesser, 1977) to Computer Supported Cooperative Work (Greif, 1988). Many of these approaches do not attribute to the agents involved any mental representation of the joint plan, the common goals, the other's beliefs, and mutual dependence.

Theoretical unification between models of objective cooperation and models of cooperative agents fully aware of highly rational deals (see also Levesque, Cohen and Nunes, 1990) is possible only if one works out a theory of plans and goals so general as to embrace both functional and intentional actions.

Plans Inside and Outside Minds

In the following, we will examine four types/levels of cooperation that must be premised to any understanding of functional cooperation.

Accidental Cooperation

We have defined (Conte et al., in press) accidental cooperation as:

$$\text{D1. (A-COOP x y p)} \equiv \text{def } \exists\text{act-x } \exists\text{act-y (M-DEP x y p)} \wedge \text{(DONE act-x)} \wedge \text{(DONE act-y)}$$

Elsewhere, we have defined (**M-DEP x y p**) as agents depending on each other with regard to two actions act-x and act-y such that (**CANDO x act-x**) and (**CANDO y act-y**), and also that ((**DONE act-x**) $\wedge$ (**DONE act-y**)) $\supset$ (**EVENTUALLY** ((**OBTAIN x p**) $\wedge$ (**OBTAIN y p**)). In words, x and y are accidentally cooperating with each other when: a) they have a *common goal* (**CG**) (defined as an identical goal p with regard to which agents depend on each other); and b) *each* of the two does the act with regard to which the other's dependence occurs. Consider a variation of Power's (1984) example of what he calls "accidental coordination": two delinquents independently arrive at an art gallery with the goal of destroying a particular picture. One, who is intercepted by a guard, diverts his attention. In doing so, she enables the other to succeed in tearing the picture.

This is not true "cooperation", although we name it as such it is a useful milestone for a theory of cooperation and an evolutionary forerunner of

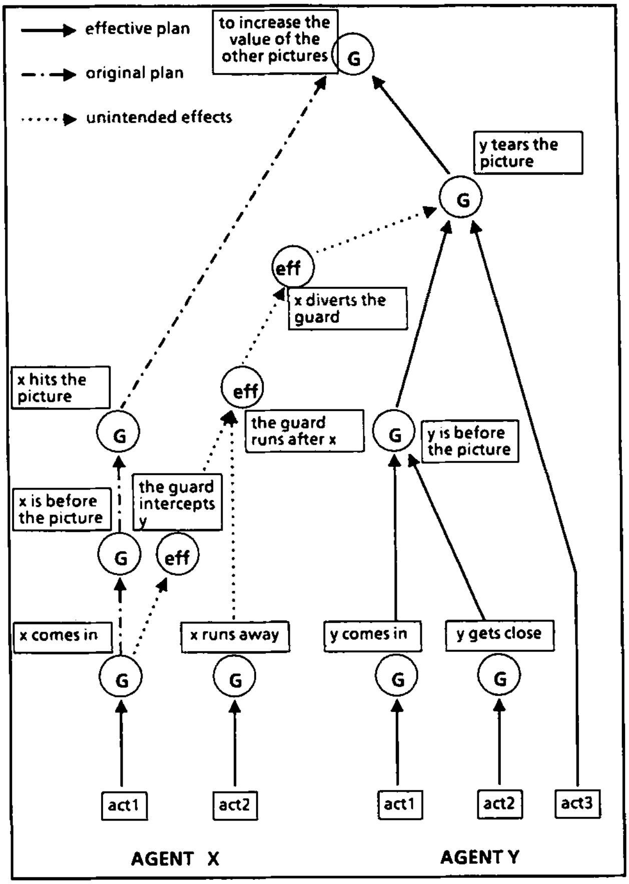

Fig. 1.

functional cooperation. Indeed, no goal or function is involved. Both actions have favouring effects, which are neither wanted or believed by the agents involved (cf. Fig. 1).

Intentional Cooperation

Elsewhere (Conte, Miceli and Castelfranchi, in press), we examined a series of weak and unilateral forms of cooperation. However, we defined full cooperation as:

$$\text{D2. (M-COOP x y p)} \equiv \text{def } \exists\text{act-x } \exists\text{act-y (KMK x y ((M-DEP xy p)} \\ \Lambda \text{ (C-GOAL x y ((DONE act-x) } \Lambda \text{ (DONE act-y)))))}$$

In this definition, not only do agents have mutual knowledge about mutual dependence but also about their having a further **CG**: due to the action-precondition rule, each agent has the goal that both actions be done, and then the goal to do what him/herself can; due to mutual knowledge of dependence, each agent wants to have the other agent in turn doing what s/he can to obtain the **CG**.

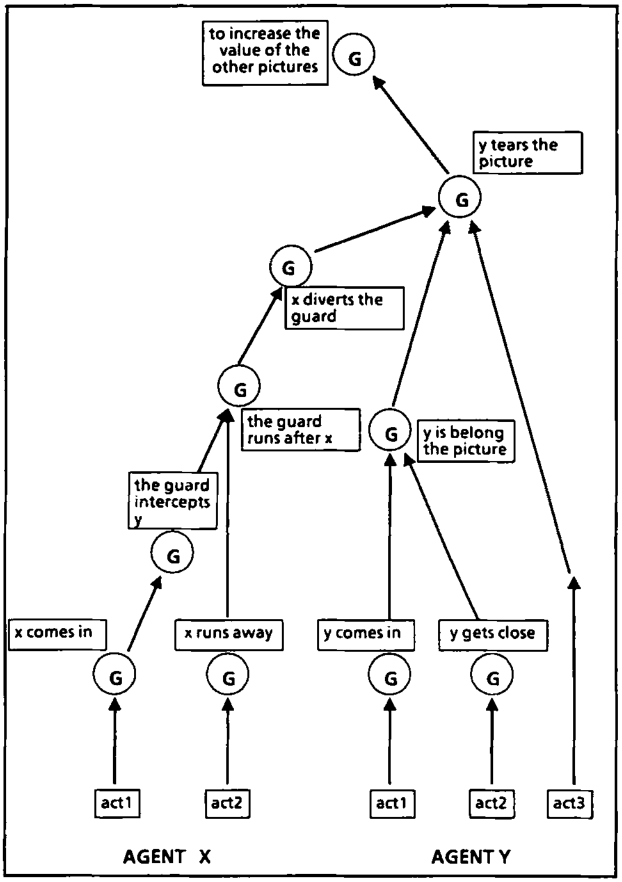

Fig. 2.

In Fig. 2, a cooperative *cognitive plan* – a plan being defined as a set of actions which converge on at least one and the same goal – is shown, as it is represented in both agents' minds.

Out-Designed Cooperation

Suppose, now, that some of the mental attitudes attributed to x and y in Fig. 2, which allow them to fully cooperate, are no longer represented in their minds, but in some third agent's, who acts like a *chairman* or *manager*.

A chairman knows what CG is realised by x and y, and also knows that agents are mutually dependent. She is the one who plans the goal's achievement in view of, say, increasing the value of some other pictures of the same author that she owns. It is even possible that agents ignore both each other's (as in accidental cooperation) and the chairman's existence. The cooperative plan outlined in Fig. 2 is this time in a single mind: the chairman's (see Fig. 3).

In agents' minds, there is usually a "social exchange" plan, that is, adoption of a chairman's goal in view of some return benefit. Cooperation here is not

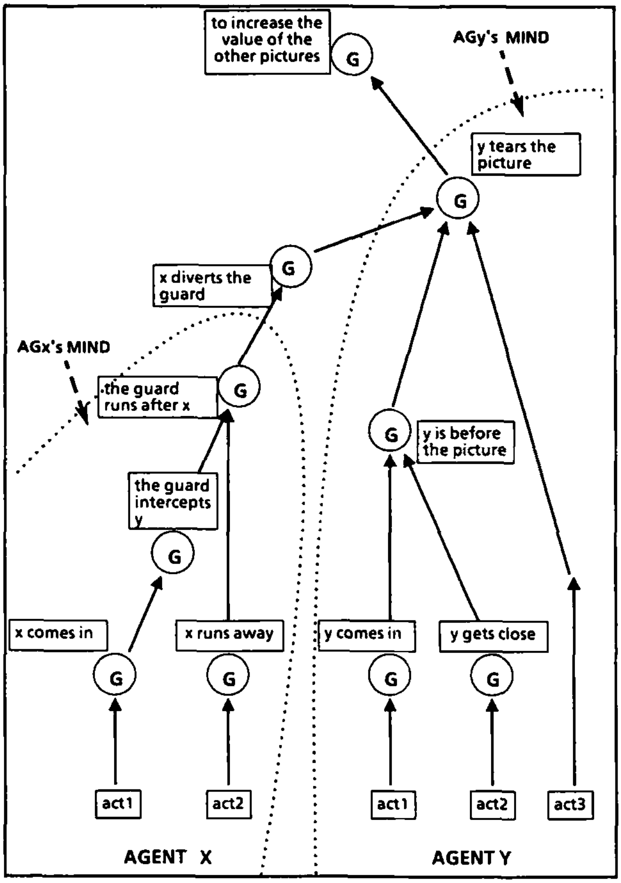

Fig. 3.

intentional, but is not accidental either, for it is wanted by a chairman. If we look at it from the chairman's point of view, it is a cognitive phenomenon, but if we take the agents' perspective it is in a certain sense "functional": to cooperate is a goal external to their minds. It is an effect unintended by agents which selects their behaviours.

This external structure of cooperation is typical of group leadership as well as of many organisations: here a structure of institutional roles is worked out (by the one who designed the organisation) in view of a CG: the institution's task.

Of course, hybrid situations are frequent, where agents know and share some or all of the plan worked out by the chairman. Finally, the chairman might as well be one of the executors of the plan.

Functional Cooperation

Let us now replace the chairman with natural or cultural selection. The CG is now a function, and is not represented as a goal in any mind (although it is "known" by the observer-scientist).

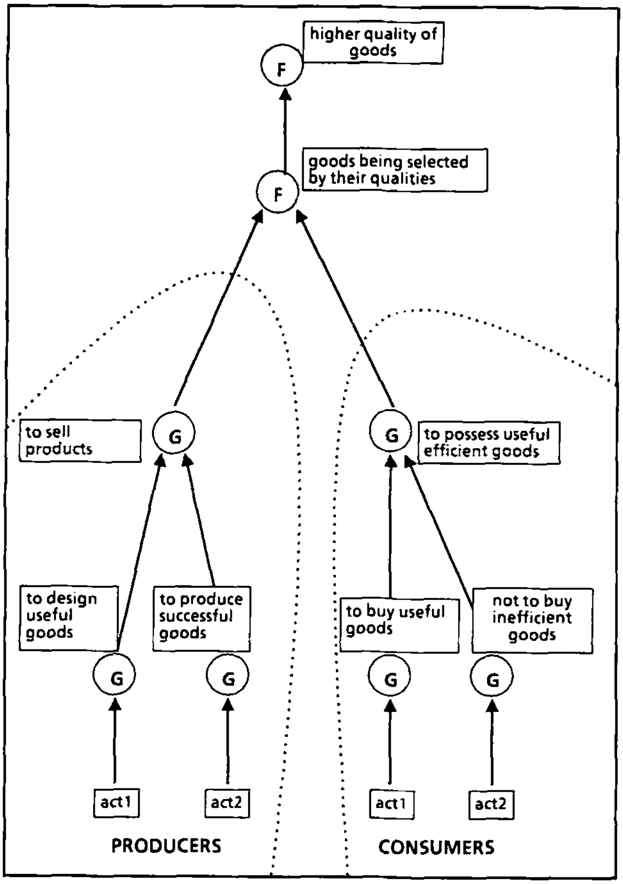
higher quality of goods
F
goods being selected by their qualities
F
to sell products
G
G
to possess useful efficient goods
to design useful goods
to produce successful goods
to buy useful goods
not to buy inefficient goods
G
G
G
G
act1
act2
act1
act2
PRODUCERS
CONSUMERS

Fig. 4.

The CG is reduced to an *objective* relation of mutual dependence. The shared goal implied by dependence (qualitative selection of goods) is a function here, a particular type of unintended effect produced by actions which had been meant to do something else (to sell products, and then to produce useful and successful goods, on the part of x, and to buy useful things, on the part of y). As shown in Fig. 4, only a part of the whole plan is in the agents' minds.

Once more, cooperation is neither intentional or accidental: agents do not produce a common effect by chance. Their actions are functional to, and selected by, that effect.

Conclusions

To sum up, we have attempted to show that:

a) *Social interaction* (and in particular, cooperation) *is not necessarily represented in cognitive agents' minds.* To arrive at a general unified theory of social interaction, objective relations among social agents need to be analysed.

b) Some *consequences of social actions* are socially relevant, although *unintended*, and some of them are not accidental.

c) There are "*emergent functionalities*" set up by intentional actions of cognitive agents: *many social relationships are functional*, and not intentional. Social roles, for instance, are the result of functional mechanisms, even though they are played by agents for personal reasons.

d) *Functional cooperation* is a phenomenon of great importance (both in sociology and in AI). A theory of action, where *goals* are allowed to be "*out of the mind*" of agents is needed for this type of cooperation to be integrated with other levels of social action. This form of cooperation, as well as other types of "goals" and "plans" at the same level of analysis, may be seen as "emergent functionalities".

References

Agre, P. E. (1989). *The dynamic structure of everyday life.* Phd Thesis, Department of Electrical Engineering and Computer Science, MIT.

Agre, P. E. and Chapman, D. (1987). PENGI: An implementation of a theory of activity. *Proceedings of the 6th Conference of the American Association of AI.* Kaufmann, San Mateo, CA.

Brooks, R. A. (1989). *A robot that walks. Emergent behaviours from a carefully evolved network.* Artificial Intelligence Lab. MIT, Cambridge, MA.

Castelfranchi, C. (1990). Social power: A point missed in multi-agent, DAI, and HCI. In Demazeau and Mueller (eds.) *Decentralized AI.* Elsevier, North-Holland.

Cohen, P. R. and Levesque, H. J. (in press). Rational interaction as the basis for communication. In Cohen, Morgan and Pollack (eds) *Intentions in Communication.* MIT Press, Cambridge, MA.

Conte, R. and Castelfranchi, C. (1991). *Mind is not enough. Pre-cognitive bases of social interaction.* TR-IP-PSCS 39, Institute of Psychology, Rome.

Conte, R., Miceli, M., and Castelfranchi, C. (in press). Limits and levels of cooperation. Disentangling various types of prosocial interaction. *Proceedings 2nd European Workshop on Modeling Autonomous Agents in a Multi-Agent World.* Elsevier, North-Holland.

Fagin, R. and Halpern, J. Y. (1985). Belief, awareness and limited reasoning. *Proceedings of the 9th International Joint Conference on AI.* Kaufmann, San Mateo, CA.

Fennell, R. D. and Lesser, V. R. (1977). Parallelism in Artificial Intelligence problem solving: A case-study of Hearsay-II. *IEEE Transactions on Computers*, C-26, 98–111.

Galliers, J. R. (1988). A strategic framework for multi-agent cooperative dialogue. *Proceedings of the 8th European Conference on AI.* Pitman, London.

Greif, I. (1988). *Computer-Supported Cooperative Work: A Book of Readings*, Kaufmann, San Mateo, CA.

Levesque, H. J., Cohen, P. R. and Nunes, J. H. T. (1990). On acting together. *Proceedings of the 9th Conference of the American Association of AI.* Kaufmann, San Mateo, CA.

Malone, W. T., Fikes, R. E. and Howard, M. T. (1988). Enterprise: A market-like scheduler for distributed computing environments. In Huberman (ed) *The ecology of computation.* Elsevier, North-Holland.

Martial, von F. (1990) Interactions among autonomous planning agents. In Demazeau and Mueller (eds) *Decentralized AI.* Elsevier, North-Holland.

Power, R. (1984). Mutual intention. *Journal for the Theory of Social Behavior*, 14, 85–101.

Rosenschein, J. S. and Genesereth, M. R. (1988) Deals among rational agents. In Huberman (ed) *The ecology of computation.* Elsevier, North-Holland.

Steels, L. (1990). Cooperation between distributed agents through self-organization. In Demazeau and Mueller (eds) *Decentralized AI.* Elsevier, North-Holland.

Werner, E. (1989). Cooperating agents: A unified theory of communication and social structure. In Gasser and Huhns (eds) Distributed Artificial Intelligence, vol. II. Kaufmann and Pitman, London.

Correspondence and offprint requests to: Cristiano Castelfranchi, Istituto of Psicologia-CNR, V. le Marx 15, 00137 Rome, Italy.