ELSEVIER

Journal of Cognitive Systems Research 2 (2001) 5–38

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

The theory of social functions: challenges for computational social science and multi-agent learning

Action editor: Ron Sun

Cristiano Castelfranchi*

Department of Communication Sciences, University of Siena, Siena, Italy

Received 1 October 2000; accepted 30 October 2000

Abstract

A basic claim of this paper is that the foundational theoretical problem of the social sciences — the possibility of unconscious, unplanned forms of cooperation and intelligence among intentional agents (the very hard issue of the 'invisible hand', of the 'spontaneous social order' but also of 'social functions') — will eventually be clarified thanks to the contribution of AI (and, in particular, of cognitive Agent modelling, learning, and MAS) and its entering the social simulation domain. After introducing Multi-Agent-Based Social Simulation and its trends, the limits of the very popular notion of 'emergence' are discussed, Smith's and Hayek's view of 'spontaneous social order' are critically introduced, and serious contradictions in the theory of 'social functions' among intentional agents are pointed out. The problem is how to reconcile the 'external' teleology that orients the agent's behaviour with the 'internal' teleology governing it. In order to account for *the functional character of intentional action*, we need a somewhat sophisticated model of intention, and a different view of layered cognitive architectures combining explicit beliefs and goals with association and conditioning. On such a basis we sketch a model of unknown functions impinging on intentional actions through a high level form of (MA) reinforcement learning. This model accounts for both eu-functions and dys-functions, autonomous and heteronomous functions. It is argued that, in order to reproduce some behaviour, its effects should not necessarily be 'good', i.e. useful for the goal of the agent or of some higher macro-system. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Social functions; Emergence; Intention; Cognitive architecture; Learning; Agents; Social simulation

1. Introduction

The social paradigm is rapidly growing within AI because of the situated and interactive perspective (Bobrow, 1991) and of Agent-oriented computing and Multi-Agent Systems (MAS) (Gasser, 1991; Hunhs & Singh, 1998). Such a paradigm will strongly contribute — mainly thanks to Agent-Based Social Simulation — to the birth of the 'Computational Social Sciences' (Carley, 2000; Müller, Malsch & Schulz-Schaeffer, 1998; Castelfranchi, 1998d). Social sciences will contribute to the design and understanding of artificial societies, cyber-or-

*Tel.: +39-068-609-0518; fax: +39-068-24737.
E-mail address: castelfranc@unisi.it (C. Castelfranchi).

1389-0417/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.

PII: S1389-0417(01)00013-4

ganisations and computer-mediated interaction, while the sciences of the artificial will transform the social sciences, providing experimental platforms, operational and formal conceptualisations, and new models of social phenomena. A significant interdisciplinary fertilisation is expected like that which, in the 1960s and 1970s, gave birth to Cognitive Science.

The basic claims of this paper are as follows:

- The main contribution of AI (and, in particular, of cognitive-Agent modelling and MAS) entering the social simulation domain will be an impressive *advance in the theory of the micro–macro link*. In particular, the foundational theoretical problem of the social sciences — the possibility of unconscious, unplanned emergent forms of cooperation, organisation and intelligence among intentional, planning agents (the 'vexata quaestio' of the 'invisible hand', of the 'spontaneous social order' but also of 'social functions') — will eventually be clarified.
- A very serious problem for the theory (and architecture) of cognitive agents is how *to reconcile the 'external' teleology of behaviour with the 'internal' teleology governing it*; how to reconcile intentionality, deliberation, and planning with playing social functions and contributing to the social order.
- To solve these foundational and architectural problems, *complex models of learning are needed*, where learning does not operate within a 'reactive' architecture made of simple rules, classifiers, associations, and stereotypic behaviours, but operates upon high level anticipatory cognitive representations (beliefs, goals) which govern intentional action. A theory of the *relationships between individual intentional behaviour, reinforcement learning, and the feedback of collective emerging effects* is needed.

I will present a critical characterisation of the problem of self-organising social phenomena and functions among intentional agents, discussing both unsatisfactory accounts in social theory and in MAS, and the hard theoretical problems to be solved.

I will also try to sketch a possible line of reconciliation between emergence and cognition, by building a notion of *behavioural function of intentional action*. To do this, I have to build on the unintended social effects of the agents' behaviours, and on some sort of reinforcement learning dealing with beliefs and intentions.

2. MAS, agent-based social simulation and their promises

Computer simulation of behavioural and social phenomena is a successful and rapidly growing interdisciplinary area (Conte & Gilbert, 1995; Troitzsch, 1997). Suffice to mention the renewed interest of sociologists and economists, testified by several workshops in the international conferences of sociology, economics, and game theory, several 'social' papers in the new area of Artificial Life (ALife), papers in the *Journal of Mathematical Sociology*, books on simulating organisations (Masuch, 1995; Prietula, Carley & Gasser, 1998), the series of SimSoc ('Social Simulation') and MABS ('Multi-Agent based social simulation') workshops, the ICSSS conference, and the electronic journal JASSS.

This impressive development is also due to the influence of AI, and in particular of Autonomous Agents and Multi-Agent Systems (MAS). Both sociologically (and cognitively) inspired approaches (see, for example, Edmonds & Dautenhahn, 2000) and reactive — biologically inspired — approaches (Agre, 1989; Ferber, 1995; Drogoul, Corbara & Lalande, 1995) give their methodological contribution. However, in this paper we will explore what could be the most relevant contribution of AI and, in particular, of Multi-Agent Systems (MAS) to social simulation (SS).

2.1. The agent-based paradigm

A new paradigm is emerging in 'SS for the social sciences'.

Evolutionary, ecological, system-theoretic, game theoretic approaches meet with approaches based on reactive and subsymbolic architectures for robotics and Artificial Life, or organisation and communication modelling, or complex Multi-Agent Systems. Beyond their clear epistemological, theoretical and methodological differences, there is something in common among all these approaches and experiments, something that characterises this new enthusiasm for social simulation: their ‘agent-based’ character. In fact, we use computers not just to calculate some complicated system of equations, to run some mathematical model of complex phenomena, but we base this analysis of global or diachronic unpredictable results on some (however primitive) ‘model’ of some ‘agent’ acting locally, with local information and local interactions (Forrest, 1990).

This direction is very promising, especially if it will exploit the contribution of AI with its impressive explosion of *agent theories and models*. Promising for what? I will answer this question not from the point of view of AI and technology, but from the point of view of the behavioural and social sciences. However, before doing this, it is worth briefly explaining why this is also an unavoidable direction for simulating social phenomena.

If one aims at simulating and explaining complex phenomena in human collective behaviour, what can be done just with systems of equations is limited; there are strong limitations also in simulating those behaviours with very simple, rule-based or neural agents. If you have to simulate the behaviour of insects or fishes, ‘swarm intelligence’ models are enough, and also if you have to simulate very general and abstract ecological or evolutionary properties such as aggregation, separation, communication, etc. But if you need a more sophisticated simulation of specific laws and phenomena of human collective behaviour (like normative behaviour, organisation dynamics, coalition formation, etc.) beyond a first coarse account for it, you need a more complex model of the agent, and precisely *a more complex theory of action and models of mind* (Castelfranchi, 1997a,b, 1998a; Conte, 2000; Chattoe, 1998). In principle, the ‘equilibrium’ you can produce (and simulate) depends on your agent model, on your *Homo* (*oeconomicus, sociologicus,...*). Consider, for example, the conclusions drawn in a quite interesting area of social simulation: the ‘population ecology of organisations’ (for example, Lomi & Larsen, 1995). Discussing various limits of these experiments (relative to the specific rules for local interaction, or to the deterministic nature of the models) the authors conclude that many of those limits might be mildened using cellular individual ‘agents’ *endowed with some more intelligence, proactivity and prevision, and memory*; and/or with some ‘strategic’ decision ability. Of course, this modelling strategy requires ‘‘*the development of an explicit theory of action at the organisational level*’’, the lack of which is currently considered as one of the major obstacles to the development of ecological theories of organisations (Hedstrom, 1992).

In fact, it is a wrong move and an illusion of separating and contrasting ‘emergent intelligence’ (or emergent cooperation) and ‘mental intelligence’ (or deliberate cooperation). In MAS we risk having this opposition (for example, Steels, 1990; Mataric, 1992): on the one side, reactive agents with collective unconscious problem solving (*emergent functionalities*), and on the other, cognitive agents that should base all their cooperation on mutual knowledge, joint intentions, negotiation, awareness and deliberation of their cooperative mechanisms (for example, Levesque, Cohen & Nunes, 1990; Grosz, 1996; Tuomela, 2000). But this opposition is, both theoretically and practically, quite ridiculous. However intelligent and knowledge-based, agents cannot have complete and certain knowledge, cannot be aware of and responsible for all the potential, long-term and combined consequences of their actions; thus, *emerging functionalities, unconscious cooperation, collective unaware intelligence must also exist among cognitive agents*! What else is, for example, Adam Smith’s ‘invisible hand’ that organises social cooperation in the technical division of labour, or in the market? Also in practical MA systems we should have levels of cooperation that are not intended and negotiated by the participants; we should model unintentional mechanisms of social coordination in human organisations and societies (Castelfranchi & Conte, 1992; Conte & Castelfranchi, 1995).

2.2. Multi-agent-based SS and the micro–macro link

If the trend is really characterised by: (a) an *agent-based simulation*, and — at least in part — (b) *the specification of the agent’s mind* with explicit and implemented models, the consequences will be quite revolutionary. In the end, we will have theoretical and experimental instruments to deal with the old *vexata quaestio* of the **micro–macro link** (Alexander, Giesen, Muench & Smelser, 1987), which is really foundational for all the social sciences.

In fact, by merging conceptual and technical tools from 'social' AI and from other SS approaches, one will show at the same time what cognitive and motivational processes determine which actions and interactions, and what combined, unintended, collective effects are determined by such interactions and in turn determine specific mental changes (awareness, learning, change of preferences, etc.) that, either reproducing or changing local, individual actions, make global phenomena evolve or stabilise. Simulations not only of real (human or animal) social phenomena, but also of *artificial and possible social behaviours* (which is one of the main contribution of DAI to SS; Gilbert & Conte, 1995), normative models and clear ontologies will help disentangle very complex and multilayered systems and notions.

Only such a 'mind-based' social simulation will allow us to 'observe' at the same time the minds of the individual agents (beliefs, desires, decisions), their learning processes, and the emerging collective action and equilibrium (perhaps even some *collective mind*) which *co-evolve*, determining each other. Only SS will allow social sciences to understand, both experimentally and formally, how deliberate actions and interactions produce unconscious social structures and phenomena (*way up*), and how social interaction, social structures and collective phenomena shape and influence (beyond explicit understanding) the individual mind (*way down*) and then the individual action that reproduces them (Conte & Castelfranchi, 1995). The term 'dialectic' (which is the only one adequate to 'explain' this kind of circular, inter-level, co-evolutionary relation) is just a philosophical label waiting for a substantial model of the co-dynamics of mind, action, and society.

Of course, this potential role is also a challenge for SS, quite an ambitious and hard one. It is also possible that SS will avoid such a challenge, limiting itself to the simulation of specific phenomena or to specific disciplinary theories or applications.

I will try to contribute to this challenge for SS — which is also a challenge of SS to the cognitive and social sciences — by providing some conceptual and theoretical analysis on the problem of the micro–macro link and, in particular, on the notion of 'emergence' and on the notion and model of social 'functions'. In fact, before considering and providing specific models and the topic of SS, a theoretical discussion is needed. This is meant to push (with my limited ability) the discipline toward the awareness of its ambitious theoretical potential, to point out the current status of epistemological and theoretical weakness and confusion in both the cognitive and the social sciences, which can create serious obstacles and cannot be bypassed just by means of good models and experiments.

To start with, it is worth addressing the very popular but unclear notion of 'emergence'.

3. An emergent confusion

The triumphant notion of 'emergence' has a bad conceptual and epistemological status. Its different meanings exemplify the confusion and the need for a discussion. This is particularly important since, in my view, only computer simulation of emerging phenomena (including social ones) can finally provide some clear notions of 'emergence'. My aim is also to stress which notion of emergence is really needed, and how to model it on the basis of selection processes (evolution or learning).

'Emergent' is a commonsense word whose meaning components and connotations play an important role in inducing us to use it in extended, ill-defined ways.

''The original meaning of the term derives from the idea of rising from a liquid, and by extension it connotes *the appearance of something previously hidden, submerged at a lower level*'' (Memmi & Nguyen-Xuan, 1995). It is a transition related to sight and view: from invisible to *visible*. It ''implies both *revelation* and *a change of level*'' (ibid.). In the 'revelation' connotation, the idea of 'interesting' and 'important' is also implied, and the idea of 'something *unexpected*' (being submerged and then unaware or unperceived or unready). This 'surprising' aspect is, for example, related — in the current technical use — to the 'unpredictability' aspect of both complex and chaotic systems.

An additional important feature comes from the reductionism–anti reductionism debate: *there is some* *property of the whole phenomenon, of the structure, that cannot be predicted of its components or elements, and that is not predictable from their properties*.

Is it possible to disentangle all these associations, connotations and features so as to arrive at some well-defined technical uses? I will just propose some important distinctions: some important 'dimensions' that are currently mixed up or implicit.

3.1. Diachronic emergence

One kind of emergence develops in time (Stephan, 1992): either developmental, or evolutionary, or historical. Let us call it 'diachronic emergence'. To see developmental phenomena as 'emergent' you need the idea that certain components, or ingredients, of forerunners of that phenomena were already present in previous stages of the evolutionary process, but were insufficient (in terms of quantity, type or organisation) to 'form' the emergent phenomena.[1] You also have to consider different periods of time as *layers* or *phases*. So there is some sort of phase transition.

3.2. Synchronic emergence

There are emergence processes that are just relative to different ways of looking at given information from one level to another; they are relative to different levels of description, abstraction, or interaction. Let us call these cases of 'synchronic emergence'. Consider, for example, the relation between the temperature or the pressure of a gas and the movement of its molecules. Or consider the emergence that I will call 'Gestalt'.

3.2.1. Gestalt emergence

'Gestalt emergence' is the emergence of a significant pattern, structure or form from the point of view of a given observer. Gestalt psychology's 'laws' (proximity, similarity, good continuation, closure, etc.) are just principles for the emergence of 'forms' (Gestalts). The most beautiful example is our perception of star 'constellations': we know that stars in our perspective (subjectively) form a group with a special structure, but they have no special relation at all among themselves (objectively), and they are billions of kilometres distant from one another (they 'interact' with each other just on our retina).

Gestalt emergence is particularly far-reaching for a couple of reasons. Unbelievably enough it *does not require 'complex' or 'chaotic' systems*.[2] Although in many domains the idea of emergence is even identified with that of 'complexity', no complexity is needed for the emergence of some 'irreducible' properties: like Gestalts or like 'being a couple', or 'being a herd'. The second reason is that Gestalt emergence is definitely subjective, and merely '*observer relative*', and this might not be true for other forms of emergence.

The significance of this distinction for simulation can be appreciated by considering that, in fact, the emergence observed and magnified in many very complex simulational systems (possibly through some graphic interface) is just some interesting (or desired) aggregation or disposition, some pattern, some Gestalt; nothing more! As Miguel Angel Virasoro said, a bit ironically, in a recent interview (Virasoro, 1996): " 'The Santa Fe' Centre, the temple of Complexity science, has perhaps shown an exaggerated optimism . . . . This was necessary for eliciting a great consensus . . . . However, this is also due to *a methodological mistake, that of studying systems only on computers; this is the best way for self-deception, since with a good graphic interface one can see any kind of behaviour emerge*." While observing coloured moving objects one is tempted to affirm: "Look: life is emerging! Look, a virus!", and so on. I agree with these considerations, and suggest

[1] In this perspective on emergence, not only global or collective effects, but also micro, individual phenomena can 'emerge' as the result of a co-evolutionary process with macro-phenomena. In this sense, for example, the 'individual' as we conceive it in modern cultures is a historical product, not designed but emerging from our cultural, economic and political history.

[2] I adopt the following informal definition of complex and chaotic systems (Virasoro, 1996): a system is 'complex' when the number of interacting objects is great, while the interaction rule among them is relatively simple. Its microscopic description is too long and impractical. A 'chaotic' system can be simple, it can be a system with a few elements, but its macroscopic behavior is complicated.

we take them as an important *caveat* and caution for all computer simulation.

I also claim that what in SS has been called an 'emergent' solution, be it intelligence, cooperation, or whatever, frequently enough is just a structure the observer-designer tried again and again to build up, or that she found interesting for her purposes; but it is merely an observed structure, with *no causal effects on the phenomenon itself, not really self-maintaining and self-reproducing*, or acquiring objectivity and independence from the observer-designer. What they call 'emergent' are just 'accidental' (but interesting) results; e.g. accidental cooperation. But accidental results are really interesting only when they can be 'reproduced', and consequently when they are preconditions for learning or for functions (see Section 5).

'Gestalt emergence' can be seen as a special case of a wider category that also encompasses 'descriptive emergence'.

3.3. Descriptive 'emergence' and beyond

Complex systems, consisting of many active elements, can be described either in terms of the actions/properties of their components or at the level of the system as a whole. At this level it is possible to provide a concise description using new predicates for new properties and regularities which are 'emerging' because they are only at the global level.

This **descriptive view of emergence** is like Gestalt emergence, although not related just to perception but to the observer's conceptualisation and description. In any case, the interest and the goodness of the emerging phenomena is *only relative to the observer's aims*. It is a *pseudo-teleological* phenomenon.

> *My main question about 'emergence' is: is this notion necessarily relative to an observer and to her 'view' and evaluation; is it necessarily and only subjective; or is it possible to provide a scientific notion that is based not on the perception, description and interest of the observer but on independent causal effects of a pattern and even on the self-organisation and reproduction of the phenomenon in/by itself?*

My answer is yes, but we need time, some memory and some re-production mechanism, or at least some causal effect: an emergent structure is objective when there is some *specific causal effect on its environment due to the global, structural properties in themselves*; and it is objective and independent even in a stronger sense when *it reproduces thanks to these effects* (circular causality).[3]

In other words: the emergent results should be 'results', they should be interesting, useful, good, efficient, etc. The question is ''for what? for whom?'' For an evaluator? for a higher level macro-system assigning a role? or in themselves (= fitness)?

Later we will return to this problem while addressing the notion of function (Section 4.3); for the time being let us establish that there could be emergent phenomena playing a causal role in nature or society and — among those — emergent phenomena causally replicating and reproducing themselves. For this, of course, diachrony and history are needed.[4]

3.3.1. Cognitive emergence

Another type of 'emergence' is particularly interesting for cognitive social science and useful for my next argument. Suppose some fact which was just objectively and from outside determining an agent's behaviour, or that was only 'implicitly' or unconsciously operating within its mind, becomes represented in that mind in an explicit way. When a social subject (e.g., a social group) becomes aware of its previously ignored objective interests, or when an implicit rule or knowledge becomes explicit, or an unconscious motive becomes conscious,[5] there is a 'cognitive emergence' from a cognitively inferior level to some higher or meta-cognitive level (Conte & Castelfranchi, 1995; Castelfranchi, 1998c).

[3] This is related to Maturana and Varela's notion of 'self-referential systems' which I avoid for its broader and philosophical character, preferring more operational and traditional notions.

[4] Of course, this does not exhaust the problems of the notion of emergence (which are all the problems of reductionism — Beckermann, Flohr & Kim, 1992); but I hope that it will at least contribute (with others, such as Conte & Gilbert, 1995; Beckermann et al., 1992; Gilbert, 1995; Odell, 1998) to setting up a discussion within the MAS and Social Simulation domain.

[5] See 'bottom-up learning' in Sun, Merrill & Peterson (1998) and Sun (2000).

4. Social functions and cognition

The aim of this section is to analyse the crucial relationship between social ‘functions’ and cognitive agents’ mental representations. This relationship is crucial for at least two reasons:

(a) on the one hand, *no theory of social functions is possible* and tenable without clearly solving this problem (see Section 4.2);
(b) on the other hand, *without a theory of emerging functions among cognitive agents social behaviour cannot be fully explained*.[6]

In my view, current approaches to cognitive agent architectures (in terms of Beliefs and Goals combined with some model of *learning without understanding*) would allow us to solve this problem; though perhaps we need some more treatment of emotions.

In particular, functions install and maintain themselves parasitically to cognition:

> *functions install and maintain themselves thanks to and through agents’ mental representations but not as mental representations: i.e. without being known or at least intended.*

While the emergence and functioning of Social Norms also require a ‘cognitive emergence’, Social Functions require an extra-cognitive emergence and working. *For a Social Norm to work as a Social Norm and be fully effective, agents should understand it as a Social Norm* (Conte & Castelfranchi, 1995; Castelfranchi, 1998b). *On the contrary the effectiveness of a Social Function is independent of the agents’ understanding of this function* of their behaviour. In fact:

(a) the function can rise and maintain itself without the agents being aware of it;
(b) one could even argue that if the agents intend the results of their behaviour, these would no longer be ‘social functions’ of their behaviour but just ‘intentions’ (see Section 4.2).

Before analysing this problem I would like to consider it from another — very relevant — perspective, more familiar to economists.

4.1. ‘THE core theoretical problem of the whole social science’: The ‘invisible hand’ (Hayek and Smith)

> ‘‘*This problem* ⟨*the spontaneous emergence of an unintentional social order and institutions*⟩ *is in no way specific of the economic science* ... *it doubtless is THE core theoretical problem of the whole social science*’’ (*Hayek*, 1967).

I believe that Hayek is absolutely right, but that the problem is not simply how a given *equilibrium* is achieved and some stable *order* emerges. Is this emergence just an epi-phenomenon? Is this ‘order’ only from the observer’s point of view? To have a ‘social order’ or an ‘institution’, spontaneous emergence and equilibria are not enough. They must be ‘functional’.

In my view, Adam Smith’s original formulation of ‘THE problem’ is much deeper and clearer, provided that we take it seriously and literally. The famous problem of the ‘invisible hand’ is in fact not simply the problem of the emergence of some equilibrium, or of the emergence of compound, unpredictable, unintentional effects. The hard question is how:

> ‘‘⟨*the individual*⟩ *generally*, *indeed*, *neither intends to promote the public interest, nor knows how much he is promoting it* ... *he intends only his own gain* ... *and he is led by an invisible hand to promote an end which was not part of his intention*’’ (*Adam Smith*, *The Wealth of Nations*, *IV*, *ii*, 9).

As one can see, this view implies that:

1. there are intentions and intentional behaviour,

[6] The notions of ‘function’ and ‘spontaneous order’ are also important for artificial systems and applications. In fact, the problem of functional social order and unplanned cooperation, independent of a designer or some centralised authority but emerging from local views and interests of distributed agents, will be one of the major problems of MAS and cyber-society. It is part of the problem of ‘emergent computing’ and ‘indirect programming’ (Castelfranchi, 1998d; Forrest, 1990).

2. some unintended and unknown (long-term or complex) effect emerges from this behaviour,
3. but it is not just an effect, it is an end we 'promote', i.e. its orients and controls — in some way — our behaviour: we ''*necessarily operate* for'' that result (Smith, ibid.).

In my view this is the right formulation of the problem. And it is a problem because it is not clear:

- how is it possible that we *pursue* something that is not an intention of ours; that the behaviour of an intentional and planning agent could be goal-oriented, teleological, without being intentional; and
- in which sense the unintentional effect of our behaviour is an 'end'.

The real challenge is relating and unifying 'mental' and 'non-mental' goals, the internal and the external teleology of behaviour; and also how intentional behaviour may be — at a higher level — just goal-oriented (McFarland, 1983).

This is also fundamental for another version of the same problem. Social institutions play the role of a collective mind whose knowledge and decisions are distributed and partially unaware (in Hayek's view no centralised planning mind can succeed in governing societies). In this perspective, the problem is not only that of understanding how this distributed control works without any centralised planner or authority. A mind is characterised by goals (control and decision are goal-based notions). So, are there *ends* in a society — not simply effects and equilibria — that do not coincide with *intentions* of the individuals and how can those ends succeed in regulating the individual behaviour? (In Hayek's perspective only individuals (and small organisations) can have ends; ends are only subjective, explicit and deliberate.)

This problem appeared in other social sciences as the problem of the notion of 'functions' (social and biological) impinging on the behaviour of anticipatory and intentional agents, and of their relations with their 'intentions'. In fact, the same problems that troubled the theory of functions appear in Smith's theory and in Hayek's view of social order. For example, the view of the society or group as an organic 'order', or the 'positive' view of functions relative to this order (Castelfranchi, 2000a).

Hayek is right in characterising the long process of the emergence of social order and the formation of institutions in terms of 'adaptation' and 'selection', but I believe he is wrong in his optimistic view of such an 'order' and evolution. Not only does he use a questionable group-selection approach but his view of selection is pan-selectionist. The evolution of a society selects and records the positive results of the experience, after innumerable trials and errors, and only the positive features and the 'right rules' survive. All the behaviours that favour the development of the group persist, and they are replaced only when more efficient behaviours are elaborated. All the behaviours that result antithetical to the group cannot persist and are eliminated (Hayek, 1952, 1967, 1973).

So what emerges is not any order but a good or even the best possible order.

Since I try to relate the 'invisible hand' and the emergence of order and institutions to the notion of 'function', let us return to the troubles of this notion.

4.2. The notion of 'function' in the social sciences: main issues

The notion of 'function' is one of the most basic but also the most ill-defined, ambiguous and discussed theoretical notions of the social sciences. Of course I will not examine the different interpretations of the term, the very long and deep discussion in sociology and anthropology, and the different interpretations of social functions (Eisenstadt, 1990). We will only consider some crucial issues of functionalism, and some analyses that are strictly related to the problems that, in my view, computer modelling will be able to solve. Following Elster (1982) we can distinguish three kinds of functionalism in the social sciences:

- *weak functionalism* (Mandeville's view): institutions or behavioural patterns often have some advantageous consequences for some of the political or economical institutions–groups dominating society;
- *dominant or sophisticated functionalism* (Merton, 1949): the possible advantageous consequences of

an institution or a behavioural pattern explain the existence of that institution or behaviour;

- *strong functionalism* (Malinowski, 1954; Radcliffe-Brown, 1957): all institutions and habits in a culture/society must have some advantageous effect (function) that explains their existence.

Strong functionalism (mainly developed in anthropology) has been subject to many criticisms, for example by Merton (1949). In the strong functionalist view, three postulates hold:

- The *functional unity postulate*: ''the function of a particular social habit is its contribution to the functioning of the *entire social system*'' (Radcliffe-Brown, 1957): culture is considered as a whole. This view is contradicted by facts: Merton observes that the integration of institutions and behaviours is only partial and contradictory in a real culture/society. Moreover, this view presupposes an inclusive 'system' before any function. Institutions and behavioural patterns are reduced to 'organs' of an 'organism' (what is called the 'physiological' meaning of function; see later section).
- The *necessity postulate*: any social or cultural expression has its function; it must be useful for the social system. This should be demonstrated case by case, empirically, not assumed a priori. Assuming it as a postulate and for all social facts makes the functionality of social behaviours quite tautological and redundant. This also suggests a conservative view of societies, since any institution is necessarily good and useful.

Very interestingly, Merton introduces the notion of *eufunctions* (with positive effects) and *dys-functions* (with negative effects).[7]

- The *indispensability postulate*: there are certain requirements that have to be satisfied for a society to exist; cultural elements are specialised and irreplaceable for their specific function. On the contrary, sophisticated functionalism accepts the idea of *functional equivalents*: a set of possible different solutions for the same social requirement.

These are some of the basic problems of functions. Let me add that for both strong and sophisticated functionalism two problems remain unsolved:

1. *how* do ''advantageous consequences of an institution or of a behavioural pattern explain the existence of that institution or behaviour''?;
2. *relative to whom or what consequences are they 'advantageous'*?

For many authors (for example, Parsons, 1962, 1966) it is clear that they are advantageous for the social system as a whole, but in other cases it is the social system itself that has functions, and these functions are either relative to the individual (to provide their needs) or to the adaptation and survival of the system in its environment.

A very fallacious argument shared by functionalist social scientists is as follows: ''if *x* were not there, the system would collapse''. This might even be true, but such a condition is not sufficient at all for identifying or justifying a function; without rain, plants would disappear, but rain does not have the function of maintaining life; without a nose, glasses would drop down, but — as Voltaire ironically remarked — the nose is not providentially there *for* supporting our glasses.

Also, the application of evolutionary theory to the level of society, which has been used to solve these problems, is quite controversial, because of the lack of a clear distinction between the level of individual action and the level of the system, and of the unsolved problem of the relationship between functions and goals (both of the individual and of the social system) (Conte, 1985).

John Elster provides a more radical criticism of any possible functional theory. He claims that Merton's more subtle distinctions between positive and negative functions, and between 'manifest functions' (intended effects; conscious motivations of social actors) and 'latent functions' (unforeseen, unin-

[7] I will adopt a similar distinction, although generalising the notion of *dys-functions* and replacing this term by *kako-functions* (self-maintaining functions that are bad for their actors or for the system). In fact, *dys-functions* strongly suggests the physiological view that I reject and consider only as a special case of hetero-functions (see Sections 4.3 and 7).

tended, and objective consequences) are also full of problems. Merton's functionalism ''is *arbitrary* because the manipulation of the time dimension ... lets us find a way in which a given pattern is good; *ambiguous* because the distinction between the 'manifest' and the 'latent' may be ... (also) read as a distinction between transitional effects and steady-state effects; and *inconsistent*, because positive latent effects ⟨traditional 'functions'⟩ (being unintentional) could never dominate negative manifest effects'' (p. 459). On the other hand, if positive effects are manifest and foreseen, they are intentional and the notion of function is *superfluous*. I will come later to this point of Elster's objections.

Before abandoning the debate about functionalism, it is worth considering Van Parijs' position. Van Parijs (1982) looks for a mechanism similar to natural selection that could explain socio-cultural functions. He claims that *reinforcement* ''is by far the most significant mechanism for the sake of legitimating social-scientific functional explanation'' at the social level. ''... Whereas natural selection and its analogues always consist in the selection of some item ... through the selection of an entity ..., reinforcement consists in selecting an item (e.g. a habit) directly within the entity concerned (e.g. an organism) ... reinforcement involves the operation of some 'choice' criterion internal to the entity'' (pp. 498–499). This parallel between selection and learning is traditional in evolutionary epistemology (Campbell, 1974). I will accept this view, while also maintaining the necessity for environmental/evolutionary selection processes; second, applying this notion not only to individuals but to 'abstract agents' that could also be groups, organisations, etc.; and third, more important, rejecting Van Parijs' next interpretation of 'reinforcement' that invalids all his previous construction, making it liable to Elster's objection.

In fact, Van Parijs continues as follows: ''reinforcement requires the registration ... of the causal link between the item and its functions ⟨awareness!⟩, whereas no awareness whatsoever is required by natural selection''. But, if reinforcement requires the awareness of the positive effects, what about its conceptual necessity among *intentional* systems? and what about Elster's objection? We will see the possibility of a cognitive reinforcement mechanism that is not based on the awareness of the function.

Although important sociologists maintain that the concept of function is fundamental and unavoidable for the study of social phenomena — but are not able to propose a convincing notion and an operational model — other important social scientists such as Giddens (1984) propose to do without the concept of function, because of its tautological character, ambiguity, inconsistency or redundancy, and of many methodological problems that make the descriptions of concrete social functions arbitrary.

I think that *a consistent and operational foundation for functional analysis of social phenomena can be provided by the Multi-Agent-based SS approach*. Moreover, if we abandon the notion and the theory of functions we lose a very crucial aspect of social emergence, 'spontaneous order', and the 'invisible hand'. We will simply reduce them to collective, complex, unplanned effects (either positive or negative), as in Boudon's notion of 'perverse effects' (Boudon, 1977). In this perspective we would be unable to distinguish an emergence that is only in the eye of the beholder (epi-phenomenon) from a phenomenon playing a causal role, and in particular a self-reproducing and enforcing social effect.

First, a sophisticated theory of action and intention is needed (Section 5.2); second, we need a theory of 'learning without understanding' and of its relationships with high level cognition (Sections 5.1 and 6.2); third, we need a theory of the relationships between ends that are internal to the agent's mind (goals, intentions) and external ends (biological and social functions, roles, etc.) (Sections 4.5 and 6.1; Conte & Castelfranchi, 1995, Chapter 8; Castelfranchi, 1982, 2000b).

4.3. What are functions: beyond regularity, equilibrium, patterns, and the physiological metaphor

The notion of emergence should be circumscribed, and the same holds for the special case of social functions. Neither the notion of 'equilibrium' nor the notions of 'order', 'pattern', 'regularity' are enough, even if one does not consider equilibria, orders, regularities due to some external intervention.[8] As already observed, a regularity or an order can be just an epiphenomenon, interesting from the point of view of the observer. On the contrary, the emergent result must play a causal role, and influence the future.

So, not any kind of dynamic equilibrium is enough. A regular recurrent pattern, or some stable configuration produced by concurrent behaviours are not enough, as long as one shares a notion of function which is derived from the evolutionary teleological or teleonomic (Mayr, 1974) vocabulary (Millikan, 1999a,b) (see Section 3).[9] Equilibrium, stability, regularity, iteration must be the result of a causal loop and of selection.[10] More precisely, the emergent pattern of unintended effects must positively feedback to its causes (behaviours) and, thanks to selection or reinforcement effects, must reproduce itself by reproducing its own causes. Thus:

> *a* ***function*** *of a behaviour is a self-reinforcing and self-reproducing* ***effect*** *which selects and reproduces its own source.*

It is a very special kind of stable or repeated pattern.[11] It needs a mechanism of *replication* in time, i.e. several *occurrences* of the same entity/behaviour and possible variations of it, and a feedback mechanism to select some of these variations or to reinforce the corresponding behaviour.

In conclusion, in order to have a function, a behaviour or trait or entity must be *replicated* and *shaped* by its effects. Appreciation, usefulness for somebody, use, destination are neither sufficient nor necessary (Castelfranchi, 1982; Conte & Castelfranchi, 1995, Chapter 8).

According to this view:

(a) functions do not necessarily apply only to collective effects (like in Boudon's view), there can be functions of individual behaviours (unintended self-reproducing effects);

(b) functions are not necessarily 'useful' for the social system where the individual is acting; more in general they are not necessarily the functions of a sub-system within and for the system;[12]

[8] Even the notion of 'self-organisation' is not enough if defined as in Haken's *Synergetic* (Haken, 1988): a system is self-organising "if it acquires a spatial, temporal or functional structure without specific interference from outside" (p. 11). Again, this 'structure' could have no causal role at all, and just be in the eye of a beholder.

[9] I build on the biological finalistic concepts and I tried to reconcile them — in a principled way — with the notion of purpose and goal in a cybernetic and psychological sense (Castelfranchi, 1982, 2000b; Conte & Castelfranchi, 1995). This is also why Bickhard's notion of function, just based on omeostasis and recursive self-maintenance, is not enough for my theory, although I like his 'interactive' and pragmatic approach (Birckhard, 2000).

[10] Also, a circular causation is not enough. For having a function, an end, it is not enough that **something exists and persists thanks to its effects**. Consider, for example, the water cycle: ocean—evaporation—clouds—rain—river—ocean; it is a cyclic and self-reproducing structure, but not functional or teleological proper. Consider also Purton's counter-example (Purton, 1978): "In a rocky region a couple of more or less rectangular boulders stand embedded in the soil a few feet apart. As a result of a minor landslip a similar shaped boulder falls on the top of the two standing stones to form a table-like structure resembling to a megaliths of Stonehenge.... As time goes on the soil is eroded from the base of the standing stones so that but for the presence of the cross-piece they would fall inwards and the structure would collapse. Now consider on of the standing stones. It contributes to keeping the cross-piece up. But also it is there (maintained there) *because* it keeps the cross-piece up..." (p. 14). What is needed is a 'genetic' mechanism producing in time different *occurrences* of the 'same' entity, and a **selection mechanism operating on these variations** (see later). Purton contests the notion of function in biology. Although his counter-example does not really affect the biological notion, it is very good against a too simplistic definition just in terms of self-maintenance and circular causality.

[11] It is not the expression of a stable or recurrent cause that produces the same regular effects in time. For example, streets being wet and slippery is not a function of rain. In the same vein, the fact that at time t_2, after some shaking, the water in a container achieves the same equilibrium as before (at time t_1) is not a functional effect: the second equilibrium is not influenced by the fact that a previous equilibrium was there at time 1. **The notion of function presupposes some 'history' in which a previous occurrence (the past) explains the present.** The aggregation of fishes in a shoal or of birds in a swarm is probably a function of their behaviour, having positive effects and contributing to their survival and reproduction. But, however similar in terms of pattern, observable structure, and regularity, the aggregation of flies around some fruit, or of butterflies around a lamp in the night, is not a *function* of their behaviour.

[12] This is just a sub-type: what we will call the 'physiological' model and 'hetero-function' (see later and Section 7). It is also called a C-Function or Cummings-Function (Cummins, 1975), i.e. the causal contribution of a functional item to a complex process. For a very important debate on this notion, see Buller (1999).

(c) functions are not necessarily 'useful', they are not necessarily good either for the system or for the agent performing the behaviour; although the individual has his/her goals and preferences, their functional effects are neither intended nor chosen, and there is no guarantee that they will not be bad for the agent even if they reproduce themselves through his/her intentional behaviour! (Castelfranchi, 2000a).

Of course one can reject this notion of 'function' and prefer to restrict it to the more traditional 'physiological' notion (which is not so well founded — see below) or in general to the C-function conceptualisation. But what cannot be done is to simply miss the important phenomenon of self-reproducing effects of these self-referential (not-useful) 'functions', or to mix it up with the general phenomenon of emerging collective effects (perhaps asystematic and accidental ones, or mere epiphenomena, or regular effects reproduced by simple causal mechanisms).

4.3.1. Beyond the 'physiological' notion: heteronomous vs. autonomous functions

Beyond evolutionary theory (from which I derive my notion of function), and the cybernetic models (from which I derive my notion of internal goal and intention[13]) we have a third family of teleonomic notions: those of physiology, representing an organism as a functional device with parts (organs) playing a specific role and satisfying functions useful for the organism; or the very close notion used in design: the function of a part in an artefact (Castelfranchi, 1982; Conte & Castelfranchi, 1995; Cummings, 1975).

This second notion of 'function' (Millikan, 1999b) is the most popular and intuitive, and the most diffuse metaphor for social functions in sociology and anthropology.

In my view, this teleonomic notion *does not have an autonomous scientific foundation*; in particular, it cannot be understood independent of the evolutionary notion. In fact, organs have functions only because organisms have been selected by natural selection (or have been designed by some intelligence) and are adaptive. An effect produced by the organ or part is advantageous and is a function only if and because it favours the fitness of the organism; it reproduces just indirectly and subordinately, by reproducing the whole organism (Buller, 1999; Wimsatt, 1972; Wright, 1976).

Those functions are just functions of 'sub-systems', they are *subordinate* and relative to the advantages (fitness or goals/needs) of the higher system. But what about the relationship with that higher level, i.e. the adaptive functions of the organism? These are not subordinate to the utility of some higher system; they are *self-referential: i.e. they are 'good' or 'useful' just in a Darwinian sense, in that they succeed in reproducing the organism itself.* Why should this notion and mechanism not apply at any level of organisation, considering that any sub-system is a system and might have some autonomy? Perhaps this is not true or interesting with regard to the relationship between organs and organisms, or between cells and organs, but it is certainly important in social science given the *autonomy* of individuals and their behaviours relative to the social system and its interests.

Thus we will assume that each action might have a *self-referential function* (indifferent to the advantages and goals of the individual) and that every individual might have a self-referential function (indifferent to the advantages and goals of the social system). In other words, *supposing x is a part or a behaviour of X, my claim is that it is not necessary that x is advantageous for X, in order to reproduce itself through its own effects.*

When using the term 'function', or 'functional', one should always be obliged to specify: functional to what?, 'functional for whom?'. A lot of misunderstanding and vagueness is due to the lack of these specifications. However, one should also consider that a perfectly possible answer is: ''just to its reproducing effects'' and ''just to itself''.

[13] The main difference between goals and functions is that functions never 'control' an action or phenomenon during its development and production; they just select it a-posteriori. They do not currently and directly 'regulate' the behaviour. Indeed, goals make behaviour 'purposive' as a cybernetic internal control of it: an anticipatory representation of some effect used as a cybernetic set-point (Rosenblueth, 1968; Miller, Galanter & Pribram, 1960).

Let us now better characterise the notion of *heteronomous* (*hetero*-) functions, as opposed to the *autonomous* (*auto*-) functions:

> In *Autonomous* (or *Self-referential* or *Self-serving* or *Absolute*) *Functions*, some entity or behaviour or feature has been selected and reproduced (also) because of its own effects or better thanks to the effects of its previous occurrences.

For a good theory of emergent phenomena — in particular in the cognitive and social sciences — such a basic and biologically inspired notion of function is needed. The etiologic and selectionist approach to functional notions (Buller, 1999) which provides the right foundation should be generalised beyond the domain of *Proper functions* and biology. Evolution is no longer a biological notion and domain; it is a more general and powerful approach to be used in other diachronic domains. Selection and reproduction mechanisms provide a materialistic foundation for teleological notions in other domains. Also the notion of *social function* — originally grounded on an 'organismic' and 'physiological' view of societies (Durkeim; Radcliffe-Brown; Malinowski) — can be grounded on models of causal effects, feedback, selection and reproduction.

> By contrast, in *Hetero-referential* or (*Subordinate* or *Relative* or *Heteronomous*) *Functions* the effect of a given entity or feature *x* is 'useful' for the internal or external purposes (goals or functions) of *another* teleonomic system *X*.

Usually, *X* is a superordinated entity, the macro-system, and *x* is a 'part of' *X* playing a 'role' in it and in its 'functioning' (Castelfranchi, 1982; Cummings, 1975). *X* contains and maintains — in some way — *x*; *x* is reproduced thanks to its contribution to *X*.

However, *x* is not necessarily a sub-system of *X*; there might be a symmetric symbiotic relation between the two: *X* profits by *x*'s behaviour and *x* profits by *X*'s behaviour.

4.3.2. Hetero-functions and functioning

Given a system S1 — endowed with whatever kind of function F1 (consisting of the set of results *P*) — and given that S1 fulfils F1 and guarantees it through its internal processes, i.e. the activity of its sub-systems or parts or components, I call 'functioning' or 'working' those processes and activities of the parts that allow S1 to fulfil its function. This is the functioning of S1 and also of its components.

The contribution that any component S2 and any process is required to give to the global functioning, what it serves **for** and **in** S1, is the 'relative' or 'heteronomous' function of S2 in/for S1. More precisely: the activity of each component S2 — often together with other components — produces (among other irrelevant results) some result/effect *p* which is its contribution to the global result *P* (F1 of S1). The result *p* (useful for *p*) is the hetero-function F2 of S2 in and for F1 of S1.[14]

The existence, maintenance, and re-production of S2 in S1 (and in some sense by S1) should be — at least in part — justified and attributable to its producing *p*, i.e. to its function F2 in S1.

In Fig. 1 we can see both: the function of the system S1 relative to its own successful reproduction and maintenance, just at the same systemic level; and

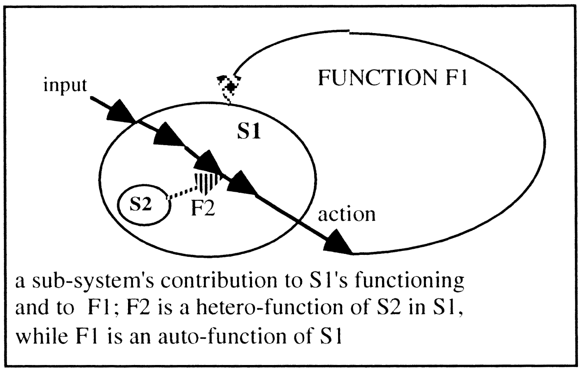
input
FUNCTION F1
S1
S2
F2
action
a sub-system's contribution to S1's functioning and to F1; F2 is a hetero-function of S2 in S1, while F1 is an auto-function of S1

Fig. 1. A sub-system's contribution to S1's functioning and to F1; F2 is a hetero-function of S2 in S1, while F1 is an auto-function of S1.

[14] 'Malfunctioning', 'out of order' or 'not working' apply when S1's working is either interrupted or modified so that it does not arrive at *P* and does not satisfy F1. This is due to internal causes, i.e. is recursively attributed to the malfunctioning of some sub-part or sub-process S2, which does not satisfy F2. Either some part is broken or some connection among the parts does not work.

the sub-functions of its component (S2) which are relative to S1 convenience and functioning.

How is it that S2 is reproduced, reinforced, created or maintained by S1 or in S1 thanks to its effect *p*? How does this positive feedback work? This is the fundamental issue to be solved for a theory of hetero-functions: the functions of parts, sub-systems, components, modules, organs, members, roles, etc. in a functional system.

The self-referential or absolute function F1 of a system S1, i.e. the function that contributes to reproduce the system *per se, independently of its functional contribution to some macro-system Sn,* is any effect *r* which does not contribute to the functioning of any functional macro-system but simply is able to feed back to S1 and reproduce it or its behaviour, thus reproducing itself. A functional sub-system can also merely have self-referential functions (sometimes in conflict with its hetero-functions). For example, any institution or organisation seems to tend to reproduce independent of its societal mission.

Malfunctioning and dys-function are not the same. A dys-function is a self-function of S2 noxious for S1, i.e. an effect *r* able to reproduce S2 and its behaviour, but contrary to F2 and then to S1 and F1. So, *r* is dys-functional and negative not for S2 *per se* but for S1 (or for an observer or user (O/U).

The problem of hetero- and auto-functions becomes more complicated when the system *X* or the agent *x* is endowed not simply with some functions but with some true goal. As for the hetero-functions one should generalise the definition as follows:

> *x part of, feature of, process in X has a function in X if it has been selected/designed because of its positive effect for some of X's goals or functions.*

Notice that a hetero-function can also be assigned by an observer or a user O/U (it is the *use* or *destination* of S1). In other words, O/U becomes the reference higher system of S1 and establishes that the 'good' result of S1 should be *P*: he wants S1 to provide *P*. This is a function of S1's which is just *relative* to O/U's evaluation or expectation, thus being its hetero-function for O/U and within the global system S1 + O/U.

So, 'functional' means either useful, advantageous for a given adaptive function, or for a given goal-oriented behaviour, or for a given goal and value.

As for auto-functions, the introduction of true goals creates the possibility of functions being either good/positive or bad/negative for *x* itself, relative to its goals. This is why we should distinguish between

- *dys-functions*, which are bad for the macro-system. They are auto-functions of *x* which are bad for *X* (for *X*'s goals or functions), and
- *kako-functions*, that is auto-functions bad at the same systemic level, for the goals of *x* itself, or for other functions of it.

So, for example, the tendency of institutions to maintain themselves beyond their mission is a dys-function of/for the society, but not for the institutions themselves; is not a kako-function of theirs.

4.4. The hard problem: intentions make social functions superfluous

As already mentioned, there is a sound model of functional behaviour for explaining goal-oriented (teleonomic) phenomena (besides the cybernetic model of purposive behaviour): it is the evolutionary model. A feature, or a behaviour *x* replicates itself (through biological reproduction, through genes) thanks to its effects. These effects are no longer *accidental* but become ends: adaptive functions, selective advantages, what *x* is useful *for*; *x* is such and such ''in order to . . . ''. This is the teleological/teleonomic (Mayr, 1974) vocabulary provided by evolutionary theory (Castelfranchi, 1982; Millikan, 1999a,b). Now, the problem is how to apply this model and vocabulary to behaviours and features that are not inherited or controlled by genes. We know that one can also model cultural phenomena, socially learned behaviours, techniques, habits, in terms of evolutionary models ('cultural evolution'; Cavalli Sforza & Feldman, 1981), but the problem is exactly here: what is the place and the role of the human ability to understand what one is doing, to evaluate what is good and what is bad, to foresee effects, to pursue them intentionally, within an evolutionary framework for cultural phenomena?

Thanks to simple learning mechanisms (like reinforcement learning) we do not have theoretical problems with animals, and we do not have problems in ALife and with swarm intelligence. Learning is a sort of selection mechanism incorporated within the organism itself, able to select, fix, and reproduce successful, 'functional' behaviours. There are mechanisms by which the 'agent' can select and reproduce the behaviour *because of its positive effects*, without understanding these effects and intentionally pursuing them next time.

Perhaps some problem can rise with the notion of 'positive' effects. As we saw before, in ALife, in robotics (Brooks, 1991), or in swarm intelligence models, quite often positive effects are *entirely traced back to the evaluation of the observer-designer*: 'positive' is the intended/desired effect of the implemented or simulated system. In order to have a real, observer-independent, emergence and functionality, one should have some *self-reproducing* organisation. Only self-reproduction allows a truly teleological vocabulary and the ascription of 'ends', 'advantages', 'functions' to the properties of the systems that are responsible for those effects that guarantee the evolutionary success of the system.

Anyway, when 'stupid' agents are at stake, either evolution or learning mechanisms or both suffice to account for functional phenomena, both at the individual and at the collective level. The real problem arises with cognitive-proactive agents, i.e. with intentional agents (Macy, 1998). In fact, we need a strange kind of behaviour: a behaviour that is goal-oriented (McFarland, 1983), teleological, but not goal-directed, non-intentional; and we need this within intentional agents.

As already argued, John Ester has definitely characterised the problem, which is of vital importance for functionalist theories in social science: for a functional explanation to be valid it is indeed necessary that a detailed analysis of the feedback mechanism is provided; in the huge majority of cases this will imply the existence of some filtering mechanism by which the advantaged agents are both able to understand how these consequences are caused, and have the power of maintaining the causal behaviour; however, this is just a complex form of causal/intentional explanation; it is meaningless to consider it as a 'functional' explanation. Thus, functional explanation is in an unfortunate *dilemma*: ''either it is not a valid form of scientific explanation (it is arbitrary, vague, or tautological), or is valid, but is not a specifically functional explanation'' (Elster, 1982, p. 480).

Let us tactically accept Elster's incompatibility claim between intentions and functions. I think that it is possible in fact to reconcile intentional and functional behaviours and explanations. With an evolutionary view of 'functions' it is possible to argue that intentional actions can acquire unintended functional effects. But, for the time being, let us restate Elster's problem and make it more radical as follows:

- Functions should not be *what the observer likes or notices*; they should be indeed observer-independent. They should be based on self-organising and self-reproducing phenomena. 'Positivity' can just consists of this. Thus, we cannot exclude phenomena that could be bad from the observer's point of view, from the involved agents' point of view, or for the macro-system's point of view. We cannot exclude Merton's negative dys-functions (or our more general notion of kako-functions) from the theory:

  the same mechanisms are responsible for both positive and negative functions.

- In fact, 'bad' or 'good' is necessarily relative to some teleological notion (ends, goals, standards), and we accept *two types of teleological notions*: evolutionary finalities and mental ends (motives, purposes, intentions). So, we also have to relate functions with non-evolutionary ends. We should account for functions that are 'good' or 'bad' relative to the goals and evaluations of the agents, and relative to the goals and values of the group. As we saw, 'functional' can also mean (in the 'physiological' perspective) ''useful for some goal of the macro-system (the organism or the organisation) and (re-)produced for this reason; fit for a given 'task' (sub-goal) assigned by the macro-system to the organ or to the role''. We have distinguished two levels of functionality: relative to a macro-system, and *per se*. Now, because we postulate *internal* (explicit) goals in certain systems (e.g., individual cognitive agents, or organisations) and we consider the relation (good or bad) between the effects of the actions and those goals, we meet the real problem.

- How is it possible that a system which acts intentionally and on the basis of the evaluation of the effects of its behaviour relative to its internal goals reproduces bad habits *thanks to* their bad effects? and, more crucial, if a behaviour is reproduced *thanks to* its good effects, that are good relatively to the goals of the agent (individual or collective), who reproduces them by acting intentionally, there is no room left for 'functions'. If the agent appreciates the goodness of these effects and the action is replicated in order to reproduce these effects, they are simply 'intended'. *The notion of intention is sufficient and invalids the notion of function*.

To found the notion of 'function', we should admit *some mechanism that reproduces the intentional action thanks to (some of) its effects, bypassing or independent of the agent understanding and pursuing these effects* (that can even be good for its goals and reproduced for that). The most relevant mechanism is some form of learning which is not based on an explicit understanding, like reinforcement learning.

However, putting a behaviourist reinforcement layer (van Parijs' mechanism of 'operational conditioning'; Van Parijs, 1982) together with a deliberative layer (controlled by beliefs and goals) is not a satisfactory solution: it is not enough to have the deliberative layer account for intentional actions and effects, and the behaviourist layer (exploiting conditioned or unconditioned reflexes) account for merely 'functional' behaviours. This is similar to what we have now in hybrid agent architectures: reactive layers competing with deliberative layers. In sociology, a rather similar reductive solution are functional 'habitus' (see below).

By contrast, our problem is indeed that *intentional actions do have functions*! Some goals and beliefs of the agents have functions. We should account for the functional character of intentional actions: goals that go beyond the intended goals, beyond the mental teleology, and succeed in maintaining — unconsciously — the behaviour.

But before addressing these issues, let us first conclude the critical discussion of sociological approaches to this problem.

4.5. Beyond functional habits: 'functional intentional actions'

Therefore, a serious problem is how to reconcile the 'external' teleology orienting behaviour with the 'internal' teleology governing it; how to reconcile intentionality, deliberation, and planning with producing or playing social functions. A simplistic solution is charging only the non-intentional, non-deliberate but merely routine behaviours with those functional aspects: according to such a view, role-playing would just be implemented in 'habitus' (Bourdieu & Wacquant, 1992). Thus, when a social actor is consciously deliberating and planning, he would not play a social role, he would be 'free'. I disagree with such a solution.

Reactive behaviours, conditioned reflexes, rule-based actions, and habits (or in a more sociological perspective, 'habitus' or role behaviours) can obviously have social functions. They can be — deliberately or unconsciously — shaped by the social environment either through reinforcements and instructions, or by imitative learning. However, in my view

(a) it is not true that social roles and functions are played, satisfied and produced only routinely, implicitly, by rule-based behaviour;
(b) it is not true that behaviours are either *functional* (then subjectively based on implicit knowledge, procedures, and automatic mechanisms; Bargh & Chartrand, 1999) or *intentional*; they could be both.

Social actors play social roles and satisfy their social functions also through their deliberate, intentional actions, however not deliberately. This requires a sophisticated model of intentions (see Section 5.2).

In Bourdieu's model (Bourdieu & Wacquant, 1992), for example, not only the social *field* where the actor is situated and its structural position produce its behaviour in a too deterministic way (see also Sun, 2000, for a criticism), but its behaviour in a role (i.e. — following the sociological tradition —

its behaviour as a social actor) is conceived too passively. The actors just follows the *rules of the game* by 'instinct', merely through some automatic 'habitus', that is, through bottom-level implicit, subconceptual processes. In such a way sociologists try to solve the puzzle of the unintentional fulfilment of social functions.

On the contrary, the real problem is modelling how we play our social roles (for example, the role of father or citizen) — while being unaware of the functional effects of our actions — not only with our routine actions but even when doing something creatively and deliberately for our own subjective motives.

In our model of functional social behaviour, the social actor is neither just an unconscious habitual role player, nor just an intentional pursuer of personal and original goals. Also his/her deliberate, intentional actions for his/her personal motives can implement social functions and roles. This does not imply that the actor is aware of such an implementation and intentionally realises his/her impinging functions.

How can our intentions serve higher aims (functions) that are not necessarily understood and intended? How can functions 'regulate' an intentional behaviour? How can our intentional behaviour have a functional self-reproducing character being selected by its effects without those effects being realised by the subject?[15]

There are two ways in which an intentional agent can serve a social as well as a biological function with its behaviour without having such a functional effect as intended or motivating effect.

(a) The *convergent* way, where the functionality impinges precisely on the intended effect (Fig. 2).

The function builds on top of the intended result (conscious aim): some effect of the intended effect is what matters. I personally want that my children learn to be obedient; this intended result contributes to produce obedient citizens.

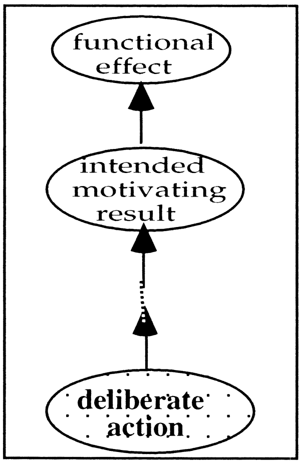

Fig. 2. The convergent way.

(b) The *divergent* way, where there is some functionality of the intended action but not of its intended effect (Fig. 3).[16]

One of the effects of the action (which subjectively is just a side-effect: unknown, unplanned or at least not motivating the action) is functionally rel-

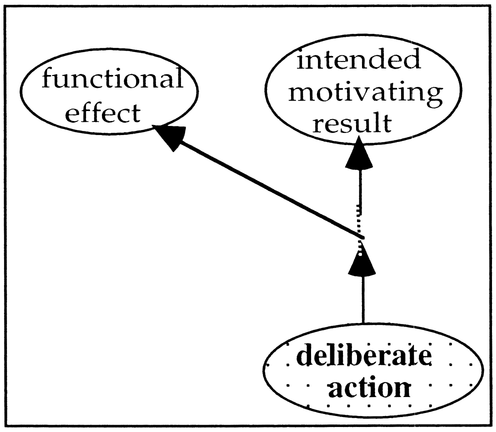

Fig. 3. The divergent way.

[15] As we saw, the mechanism I propose does not concern only hetero-function or C-function like playing social roles that are supposed to be useful for the society; more generally, it models feedback effects, reinforcement and selection on intentional behaviour, independently of this feedback being provided by the 'containing' macro-system and the effects being useful to it. There can be reinforcing effects and feedback independent of any useful result for the macro-system or for the individual.

[16] Of course the intended and motivating results also play some indirect functional role because they are useful and necessary for motivating and reproducing the functional behaviour. In this sense, they are functional: for example, pleasure in sex, or fear of sanctions for normative behaviour.

evant. One can have sex just for pleasure but the possible effect of having offspring is its biological function.

5. Cognitive requirements for a theory of social functions

In order to account for the functional character of intentional actions, from the cognitive point of view, on the one hand we have an *architectural* problem, on the other we need *a sophisticated model of intentional action*.

5.1. For a hybrid architecture: intentional behaviour and reinforcement learning

I think that basically the solution of such hard problems can be found in *learning*, and, in particular, in learning in a MA context and with a social feedback. But what about reinforcement learning in cognitive, deliberative agents? What would be needed is some form of reinforcement learning on top of intentional behaviour. How is this possible and non-contradictory?

It is well established in psychology and Cognitive Science in general (after, for example, Anderson, 1983) that there are different kinds of mental representation and knowledge: procedural vs. declarative, implicit vs. explicit, subconceptual vs. conceptual, etc. They are subject to different kinds of elaboration and learning. Action can also be based on those different kinds of knowledge: it can be either reason-based, driven by explicit expectations and beliefs, chosen on the basis of preferences and evaluations; or based on reactive rules and associations.

One might say that the prototypical view of mind in Cognitive Science implies a layer of 'low level mechanisms' which are merely associative, reactive, procedural, implicit and subsymbolic, and such a layer is placed below the layer of declarative mental representations, explicit knowledge, where judgement, reasoning, decision, etc. operate. Eventually and possibly an additional layer of meta-cognition or reflexive knowledge operates on the explicit cognition level.

The idea of such a hybrid architecture is basically correct and heuristic (see, for example, Strube, 1999), especially when the interaction between the two crucial layers and the relation between the two in learning are also modelled, such as, for example, in the DYNA model (Sutton, 1991) or even better in the CLARION model (Sun, 1997; Sun & Peterson, 1998; Sun et al., 1998). It is also the most promising path for our problem. However, it is important to realise how such a coexistence should be conceived. It is not enough to have a layered architecture (or a concurrent architecture — Kurihara, Aoyagi & Onai, 1997) where merely rule-based or reactive behaviours, which evolve by reinforcement learning, compete against a deliberative layer (which improves only by understanding, reasoning and deliberation) for controlling the agent's external action.

Since our thesis is that not only reflexes, routines, or habits, but also intentional actions play roles and produce functions, the hybrid architecture should show how this works.

So, let us claim, differently from this traditional view, that a number of low-level (automatic, reactive, merely associative) mechanisms operate *upon* the layer of high cognitive representations (beliefs, goals, expectations, reasoning, etc.). For example, the novelty and the interest of Damasio's theories does not lie so much in their treatment of decision making or in the role of emotions in rational choice. Rather, the valuable idea is that of a sort of classical associative and conditioning mechanisms which do not impinge on responses and behaviours, but precisely on high level mental representations, like alternative scenarios of choice, hypothetical goals, etc. To those high level explicit mental representations (beliefs, goals, etc.) more or less central affective responses are conditioned: 'somatic markers' and consequent 'mental' reactions of attraction or repulsion.

In sum, *low level associative learning mechanisms and reactive devices can operate upon high level mental representations* (Fig. 4).

This kind of architecture can do without the necessity of an infinite recursion of meta-levels, and some paradoxes of will, goals and meta-goals, decisions about preferences and decisions. At least in some cases, at the meta-level there are no true explicit goals (about ourselves, or about optimising utility, or avoiding pain and looking for pleasure, or about having coherent beliefs, etc.), but there are simple procedures, automatisms dealing with our mental representations. They are teleological (goal-

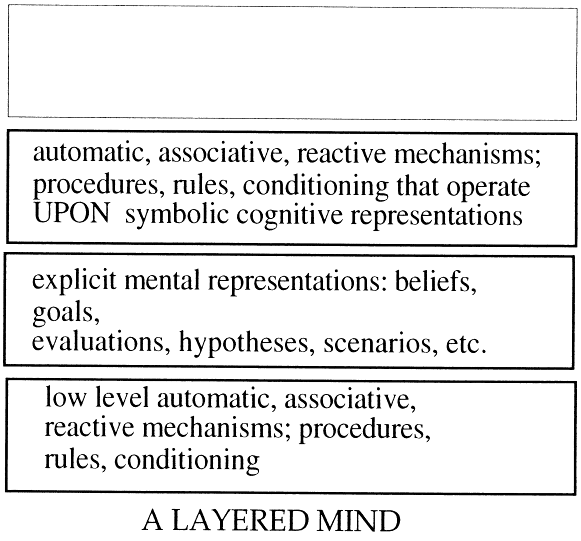
automatic, associative, reactive mechanisms; procedures, rules, conditioning that operate UPON symbolic cognitive representations
explicit mental representations: beliefs, goals, evaluations, hypotheses, scenarios, etc.
low level automatic, associative, reactive mechanisms; procedures, rules, conditioning
A LAYERED MIND

Fig. 4. A layered mind.

oriented) (McFarland, 1983) in the functional way, not in the intentional way.

To be true, this architecture is implicitly presupposed in the classic cognitive approach, based as it is on the idea of 'manipulation' and 'elaboration' of symbolic representations. Mere procedural rules, mere algorithms and mechanisms operate upon those explicit representations, manage them and transform them. This *hidden* procedural layer placed upon the symbolic one can — why not? — include reactive rules, associations and conditioned responses. Those can work as reinforcing, learning and selecting devices for building unintended functions on top of beliefs and intentions.

5.2. A correct view of intentional action and its effects

As we saw (Section 4.5) one can intentionally act while not intending all the effects (and then possible functions) of his/her action. Let us clarify how this is possible in a cognitive view of intentional actions.

Actions (α) have results (**R**); part of these results (a subset of **R**) are known; part of them are known *before* α, i.e. they are *expected* (**ER**). Intentions or better intended results (**IR**) are expected. Intentions, in fact, require an anticipatory representation of possible results of α. Moreover, those expectations must drive and select the behaviour, i.e. they *control* the action selection and execution. This is why action is 'purposive behaviour' (Rosenblueth & Wiener, 1968) with its internal teleology.

In the subjective perspective of the agent, the expected — and in general known — results of α can be either *negative* or *positive*. 'Negative' means that they are adverse to the agent's goals, while 'positive' means that they realise or favour some goal of the agent. The agent in fact has goals (wishes, desires, needs, duties, etc.) and s/he acts in order to realise them. So α is aimed at realising some goal through its results. Intended results (**IR**) are *positive* **ER** of α. However, not all the positive **ER** are **IR**. One can anticipate some positive (favourable) result of his/her action without acting *for* that goal and then for that positive **ER** but for another one. In other words,

> *really 'intended' are those results* ***for*** *which, in* ***view of*** *which,* ***in order to achieve*** *which one is acting.*

They are *motivating* the action: the agent chooses and performs α **iff** and until s/he believes that α will produce that specific **R** (Fig. 5).

It is also important to consider that deliberate *inaction, omission, decision of not doing something, is an action.* On the basis of **ER** and inaction we can introduce the useful notions of Passive intentions and Side intentions.

We have a *Passive intention* when I could prevent something from happening, but (because I like it or because preventing it would be expensive) I decide to let it happen by itself. I do not 'produce' that effect in the world: it is the effect (intended or not) of another agent or of another action. I do not

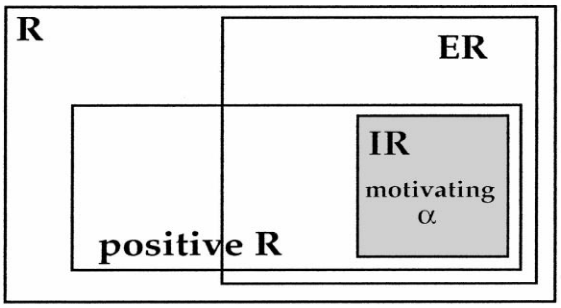

Fig. 5.

specifically act in order to achieve *q*; I could just prevent it.

Side intentions are a special kind of Passive intention; they occur when the expected non-motivating result (good or bad) is the side-effect of my own intentional action. In this case I could prevent it only by renouncing my action for my goal. It is 'passive' in the sense that it is not actively pursued; I just let it happen (as a consequence of my own action).

Beyond **IR** (motivating anticipated results) intentional actions have known and unknown, expected and unexpected, positive and negative results. All of them can produce complex emerging effects and support functions. The agents can even understand and be aware of emerging or functional effects of their actions. But even if those effects were 'positive', they would not become 'intentions' just for this (for an even more radical divorce from Elster's claim, see Section 8.2). Vice versa, functions must not necessarily be *positive* for some agent's goals.

6. How social functions are implemented through cognitive representations

After the previous characterisation of the critical points in the notion of social 'function', and the necessary specifications about intention and cognitive architecture, we can try to sketch the 'internal' mechanism(s) for external functions impinging on intentional actions.

I will first describe an abstract simplified model of 'auto-functions', be them either 'kako-functions' or 'eufunctions' relative to the goals or interests of the agents they emerge from. Second, I will exemplify this model in specific cases of kako-functions and of eufunctions. Third, I will extend the model to the social hetero-functions which are *useful for* and *reinforced by* a macro-system.

Why also use kako-functions? They are quite important. First, it is important to understand that the mechanism that installs a bad function can be exactly the same as that installing a good one. Second, this is also meant to distinguish a functional view of behaviour and society from any metaphysically-teleological, providential view (functions can be very bad and persist however bad).[17] Third, kako-functions are important theoretically because they cannot be explained in a traditional behaviourist learning framework: the result of the behaviour can be disagreeable or useless, but the behaviour will be 'reinforced', consolidated and reproduced.

6.1. The basic model

Functions are just effects of the agents' behaviour, that go beyond the intended effects and succeed in reproducing because they reinforce the beliefs and the goals of the agents that caused that behaviour. Thus:

- First, behaviour is goal-directed and reason-based, i.e. it is intentional action. The agent bases its goal-adoption, its preferences and decisions, and its actions on its Beliefs (this is the true definition of 'cognitive agents').
- Second, there is some effect of those actions that is unknown or at least unintended by the agent.
- Third, there is circular causality: a feedback loop from those unintended effects that strengthens, reinforces the beliefs or the goals that generated those actions.
- Fourth, this 'reinforcement' increases the probability that, in similar circumstances (those activating the same beliefs and goals), the agent will

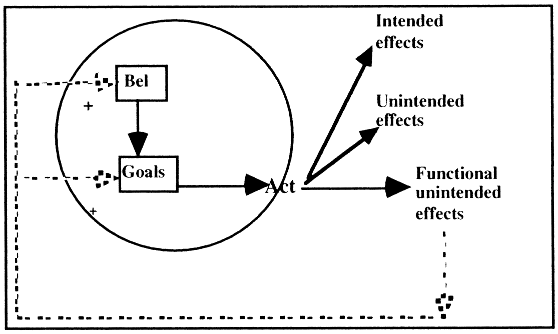

Fig. 6.

[17] Consider Marx's ironic (but true) thesis that courts and prisons in fact produce criminals, and by doing so they reproduce themselves. Thus, their real 'function' — very different from their official one — is the reproduction of criminality. Notice that, in fact, society (government, criminology, etc.) is now aware of this bad and even self-defeating 'function' but is unable to avoid and correct it (see Section 7).

produce the same behaviour, thus 'reproducing' those effects.

- Fifth, at this point such effects are no longer 'accidental' or unimportant: although remaining unintended, they are teleonomically produced: *that behaviour exists (also) thanks to its unintended effects; it was selected by these effects, and it is functional to them*. Even if these effects could be negative for the goals or the interested of (some of) the involved agents, their behaviour is 'goal-oriented' to these effects.

6.2. Cognitive 'reinforcement'

Reinforcement learning is based on the classic 'Law of Effect': the probability of unsuccessful actions decreases, while that of successful actions increases. However, this can be obtained through very different devices. I argue in favour of a cognitive variant of this law and mechanism: *since — in cognitive agents — action depends on goals and beliefs, goals and beliefs must be 'reinforced' in order to reinforce the action*![18]

It is typical of reinforcement learning to formalise the purpose or goal of the agent simply in term of a specific reward signal from the environment to the agent. This simplification is very general and practical for several purposes, but it is quite limiting for accounting for *high level reinforcement learning* in cognitive agents. It is necessary in this case to maintain explicit in the theory the goals and the beliefs of the agents (in general, its mental representations) and to model how the reinforcement acts on them.

Basically there seem to be two *Cognitive 're-inforcement' principles*.

6.2.1. Belief (expectation) reinforcement

Two different mechanisms can be hypothesised:

- *association* (*accessibility*): the *association* between the belief and that context or scenario is strengthened: the probability that the belief will be retrieved next time in similar situations increases; it will be more strongly activated and more available (accessibility bias);
- *confirmation* (*reliability*): some of the action's effects are perceived by the agent (even if not necessarily understood and causally connected to its actions) and they *confirm* the beliefs supporting the action: they give new *evidence* for that belief, increase its 'credibility', and reliability: they augment its 'truth' or the subjective probability of the event.

More precisely, what is reinforced are not simply beliefs, but *expectations* about the attitude and the behaviour of others, and about the effects of actions. Expectations are anticipatory mental representations. They are beliefs about the future, related to goals: 'positive' expectations, when we believe that p will be true, and we desire p; 'negative' expectations when we believe that p will be true, and we desire Not p (Castelfranchi, 1997a).

6.2.2. Goal reinforcement

Two different mechanisms can be hypothesised (analogous to the belief reinforcement mechanisms):

- *association* (*accessibility*): the success of the chosen goal, plan, action is recorded in the sense that the association between the goal-plan and that problematic context or scenario is strengthened: the goal/plan (solution) will be more likely retrieved next time in similar situations; it will be more strongly activated, more available and accessible;
- *confirmation* (*reliability*): the success of the chosen goal, plan, action is recorded; it increments a 'successfulness index' relative to that choice; or better some meta-cognitive evaluation of the value of the action. This memorised behavioural choice is 'confirmed': next time the probability to choose the same way (goal, plan, strategy, action) will be greater: it will be more preferable and reliable (we will trust it more).[19]

Since the probability that a goal or plan is pursued (chosen) depends on its 'success' and on its cognitive supports (reasons) (Castelfranchi, 1996) the action of

[18] But there is not only reinforcement learning to 're-produce' the action: consider, for example, persisting after failures.

[19] This is the specific mechanism that applies in my main example, but there are other Goal Reinforcement mechanisms. In the example of courts and prisons' function, the goals of the agents/institutions are only partially satisfied, and partially self-defeated. *What will 'reinforce' the goal is indeed its partial failure*. More precisely, failure (the 'habitual criminals') creates the conditions for a new activation of the same goals (to arrest, to imprison). Notice that not only the success but also the frustration of a goal can create appropriate conditions for its persistence or repetition.

the first or the second principal will determine *a reinforcement of that behaviour.*

(a) If a given belief activated or induced a certain goal (choice), the fact that this belief is more accessible and available, and more credible, more subjectively probable, will increase the probability that the same choice will be made in the future.
(b) If there are several alternatives, either as goals activated by certain beliefs, or as plans or strategies to reach an adopted goal, and if each already experienced alternative has a memory of its successes in execution, after a successful execution the probability to take the same way is again increased.

Notice that the agents do not necessarily intend or suspect to reinforce their beliefs or goals, and their own or the others' behaviour.

6.2.3. Restating

Let us now look at the same phenomenon with another perspective able to enlightening another — concurrent — mechanism.[20] Even without postulating any reinforcement and learning by the agent, *an effect that maintains or re-creates those* ***contextual conditions*** *that lead to that action, maintains or increases the probability for its re-occurrence.*

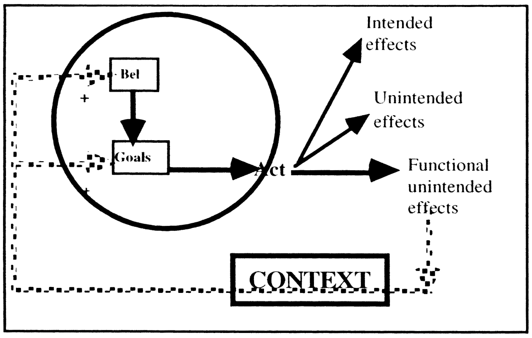

Fig. 7.

Any action is in fact 'interaction' with and in an environment (Castelfranchi, 1996), and motives, goals and actions are elicited by environmental stimuli, selected by and adapted to the contextual conditions.

Following our definition, *any regular — not accidental and occasional — effect which is self-reproducing through the re-statement or 'replica' of its action/cause, is a 'function' of that action.*

It is unnecessary to suppose a learning mechanism (reinforcement) on the goals and actions as the basis of systematic reproduction of the behaviour. There may simply be an effect on the context of the behaviour such that the choice mechanism will resort to the same behaviour.

This is also possible with rational decision: the unplanned and ununderstood effects of my (our) action creates external conditions and incentives — that I perceive — such that it is convenient for me to continue along the same behavioural way. Consider, for example, the spontaneous division of labour. The only thing the agent does not understand is the origin of local incentives and conditions — which are also effects of its own actions (Egidi & Marengo, 1995). This is the same mechanism of 'self-fulfilling prophecies'.

Clearly, at a more subtle level of analysis we are again speaking of *beliefs*. In fact, the agent has to *believe* that there are certain conditions and opportunities and that it is convenient for her/him to carry out a given action in these conditions. In any case, beliefs are preconditions of intentions (Cohen & Levesque, 1990; Castelfranchi, 1996, 1998e) and of their execution, and preconditions for reactive high level behaviour. Thus the reproduction of the action through the context maintenance or modification is also in fact a reproduction of the action through the maintenance or modification of some specific beliefs.

6.2.4. Emotional reinforcement: "Preferences need no inferences" (Zajonc)

Another plausible mechanism can account for this reinforcement of expectation-driven (intentional) behaviour: *the learned emotional reactions to anticipatory representations*. These mechanisms are both real and complementary to the former.

There is something in our body that unconsciously and automatically orients our decisions and actions,

[20] It is also possible that the behaviour is reproduced as a mere routine, habit, or trivial condition-action rule, and that the intentional justification is only a *post hoc* non-influent subjective interpretation of it. But this makes the problem trivial; in fact, unintended functionalities and organisations emerging from non-intentional behaviour are something clear, well known and theoretically harmless (Section 4.5).

attracting or repulsing us; giving us an unreasoned appraisal of possible scenarios (Miceli & Castelfranchi, 2000; Castelfranchi, 2000c). These unreasoned 'preferences' and choices can be accounted for by Damasio's 'somatic markers' (Damasio, 1994) that produce an immediate screening of the alternatives we — as proactive, anticipatory animals — are prospecting. They are generated by the emotions experienced in previous situations, and are *associated* (learning) with the prospect of future possible scenarios (not simply to the behavioural response).

So there is some sort of pre-decision, some *preselection* of possible goals, which is based on emotion and learning (positive and negative reinforcement on symbolic mental representations! — Section 5.1).

In the following we will try to apply this model of cognitive 'reinforcement' to both *absolute/autonomous* and *relative/heteronomous social functions*.

6.3. Examples

6.3.1. Absolute social function 1: car crossing

Let us see how it can happen that in a given town (say, Rome) a given interaction habit is established and reproduces independent of the intentions of the involved agents. The habit (which is typical of Rome) is the following. Although there is a traffic norm that assigns precedence to cars coming from the right, at a crossing one can often observe that the entitled driver stops and lets the other car (which has no precedence) cross. I would argue that this interactive behaviour is a self-maintaining and self-reproducing social bad habit; that it is produced by the intentional choices and actions of the drivers; but it is not intended by the drivers: it is just a (bad) function (kako-function) of their behaviour. I need two driver characters: the timid, and the aggressive. Let us now look at the timid character's beliefs and goals when arriving at the crossing.

> *The timid mind*. The timid believes (**B1**) that there is a certain number of 'aggressive drivers' (Class X) who might not respect traffic rules, and would try to cross a crossing even without having precedence. This belief has a certain strength or probability in his mind. He also believes (**B2**) that to be careful is better, and that to slow down (and in some cases stop) and letting the other cross is careful. Thus, he also has the goal (**G1**) of being careful and of letting the other cross if she is really trying to do so. The timid's character consists exactly of this belief and of the consequent preference to 'let the other win'. He also believes (**B3**), in this specific crossing situation, that the coming driver is probably an 'aggressive one' (since she is not slowing down enough): she is a member of Class X. Thus, his goal is instantiated: to slow down and (in case) let the other cross. This expectation and this goal induces a careful and hesitating behaviour. What happens at the same time in the mind of the 'aggressive' or simply 'impatient' driver?
>
> *The aggressive mind*. She believes (**B1**) that there are several slow, hesitating, uncertain drivers (Class Y) that waste our time (she could also believe that Norms themselves are stupid things and waste our time). She also believes (**B2**) that if one tries to cross — not slowing down — one succeeds because the other will give up. Thus she has the goal (**G1**) to try to cross anyway. When arriving at the crossing, she does not slow down in time, and, observing the careful behaviour of the other (that she herself is favouring), she will assume (**B3**) that the other is a member of class Y, and that he will not compete. Thus she will have the goal of not stopping and of crossing.

What is the result of this coordinated behaviour (based on the reciprocal understanding of the other's intentions)? The result is that the bad driver will cross before the other, but there will be no accident. However, the most important effect is that *the expectations* of both the drivers relative to each other's behaviour and to the success of their own behaviour (respectively: to pass without wasting time; to avoid an accident) *are confirmed*! More than this: basic and general beliefs are confirmed. In the timid-careful mind, the beliefs that "some people are not careful and are aggressive" receive more evidence and examples; the belief that to give up is careful and avoids accidents is proved; the decision to do so is successful.

In the aggressive-impatient mind, the beliefs about the existence of slow/timid drivers are strengthened; the beliefs that they will not compete and that there will not be accidents are confirmed; the strategy of trying to intimidate and pass in this circumstance has been successful.

Both agents, without (necessarily) understanding this, and without wanting this, *produce the effect of reinforcing their own behaviour and the other's beliefs and preferences*. It is very important to notice that even if one of the drivers were aware of these effects this does not imply that he or she wants such effects. Especially for the careful guy (for whom the function is *bad*), it is clear that he does not want to increase the number and the aggressiveness of aggressive drivers. Moreover, even if some driver, e.g. the aggressive one, had such a pedagogical intention (of incrementing timid behaviour), the phenomenon does not reproduce itself thanks to this intention!

Let us now reconsider this example also applying another possible reinforcement: the *emotional* one. Suppose that our timid driver arrives at a crossing in Rome believing that everybody will respect the rules. When he meets some aggressive transgressor he will experience surprise, fright and even worse: a crash. He will associate his negative emotions with that scenario. So, next time, arriving at a crossing in the same conditions, he will first be aroused and alarmed, and second he will feel some fear. In particular, the option/scenario of not slowing down is associated with fear and is negatively 'experienced'. So he will automatically be attracted by the other alternative of slowing down to see whether the other driver tries to cross and in that case let her go. The emotionally positive experience of this behaviour (everything goes well) will be associated with it and will reinforce it, increasing the probability that, in similar conditions, the driver will behave in the same way. The goal has been *emotionally reinforced*: it is more 'attractive' (and the alternative more 'repulsive') not for reasoned arguments but just in virtue of associations (it is somatically marked).

What is important is that this has to do not only with 'responses' or behaviours but also with anticipatory representations, and thus with a goal-directed and intentional behaviour. It is another way of 'reinforcing' true goals and not simply behaviours.

In any case, *the agents are unconsciously cooperating to reinforce each other's behaviour*. In this way the social phenomenon stabilises, reproduces, and *spreads* around through a typical MA learning process.[21] Notice that in such a MA learning situation the behaviour of agent A reinforces the behaviour of agent B and vice versa, although they are opposite and complementary, not in the usual way based on imitation, conformity, and 'social proof'. Also notice that there is no awareness/understanding of this and that awareness is not necessary for learning.

6.3.2. Absolute social function 2: dirty and clean streets

Let us now consider another example of a social (kako)function, which is based on completely different kinds of beliefs and goals. In fact, it is based on *social conformity and imitation*. Also, these kinds of mental representations are able to establish and reproduce an emerging global social phenomenon that is neither understood nor intended by the agents. Normally, this is supposed to refer to 'social norms' or 'social conventions', while I strongly disagree with such a view (Conte & Castelfranchi, 1995). In these cases we just have 'social habits' and 'functions'.

The problem is why people throw garbage (e.g., tickets, paper, cigarettes, etc.) in the street when this is a diffuse (bad) habit. The following is the mental set I postulate.

The agent assumes (**B1**) that this is a bad behaviour or even a forbidden one; he assumes (**B2**) that a lot of other people behave in this way; that (**B3**) this can somtimes be quite practical and easy; he assumes that (**B4**) his contribution to the amount of garbage is quite marginal and small (which is true). He has the goal (**G1**) to do as others do and until others do so (Bicchieri, 1990); or at least to do as others do and until others do so if this is useful and practical for his goals. On the bases of beliefs **B2**, **B3** and **B4**, Goal **G1** will generate another goal (**G2**) 'to leave garbage in the street', which overcomes the possible goal (**G3**) — based on **B1** — of not dirtying the city. The result of such a behaviour

[21] This coarse model (that is waiting for formalisation and simulation) predicts increasingly careful behaviour in prudence-oriented people, and vice versa: so the population should polarise into two groups. This is not necessarily the case in real life, because of other interfering factors, such as accidents and experience, or imitation: why should I not violate the rule? why should I always give up? However, these interfering factors do not prevent the maintenance of such a quite diffuse kako-function.

is that streets are dirtier; this is perceived and then it will confirm the supporting beliefs (**B2**, **B4**) and goal **G2**.

It is quite interesting to observe that exactly the same kind of beliefs, and an identical goal (**G1**), can generate in this case a eufunction: to maintain the city clean.

> *If everybody avoids throwing garbage in the street, and the streets are clean, then nobody is encouraged to throw garbage. In this case, to satisfy his goal* (***G*1**) *the agent adopts the goal* (***G*4**) *of 'not dirtying the street'. This will also be in agreement with **B*1**; but it is in contrast with **B*3*** (*and a related goal **G*5** of acting in a practical and easy way*).

Everybody reinforces others' behaviour. Notice *that the global effect is wanted and intended by nobody*; that the reinforcement effect is also unexpected and unintended. The behaviour is (reciprocally) reinforced by its effects. These effects are self-maintaining and reproducing through the reinforcement of their own causes. This causal chain *passes through the mind of the agents* (their beliefs and goals) but not through their consciousness and intention.

Unlike the previous example (Section 6.3.1), here reinforcement is also due to imitation, conformity, 'social proof' (Cialdini, 1993) and social expectations. This is a well known and much more studied mechanism of MA learning which plays a role in the emergence of a stabilised social habit. Clearly, agents can feel confirmed in their behaviour (reinforcement) by observing other agents doing the same.

7. Kako- or eu-functions: relative to whom or what?

In what sense is the 'clean-street' habit good and the 'dirty-street' habit bad? As we saw, 'good' *(eu) and 'bad' (kakos) must be relative to the goals or interests of some system/agent* (Miceli & Castelfranchi, 1989). In fact, so far we have referred the bad character of these functions to the involved agents' goals or interests. So, the habits of dirtying streets is bad relative to **B1** and its related goals or to the cleanliness and aesthetic interests of the agent. With respect to those goals and interests the clean habit is a eu-function. But I also understand that, since the agents' goals may be in conflict, so the functions may be eu — or kako — depending on the focused goal. For example, to maintain the streets clean is a kako-function relative to **B3** and goal **G5**. This seems counterintuitive, because, in fact, when we view an emerging function as good or bad, we do not usually relate it to the local utility of the agents the function is emerging from, but to the utility (goals and interests) of the macro-system in which the agents are situated or by which they are governed.

7.1. Hetero-functions and the macro-system's role

If in these two cases of kako-functions (which are bad for at least some goal or interest of at least some of the involved agents) we consider the macro-system (group, society), we can see that this self-reproducing 'function' of the agents' behaviour is also bad relative to the goals of the macro-system: it is not good for its functioning or for its goals and values (dys-function). That is why, in fact, the macro-system can try to repress those behaviours through norms. But the system is not always aware of the real cognitive micro-mechanisms.

What happens when some function is good for the macro-system? Will it necessarily acknowledge this and 'deliberately' encourage and reproduce this phenomenon? or will the phenomenon remain a mere 'function'? To have pure eu-functions, the macro-system has to remain either unaware of the positive effect, or, while being aware of it, it should not understand how to confirm or improve it, or be unable to intervene. However, the answer is not so sharp and simple, because there are several hybrid cases (but see also Section 8.2). As Nigel Gilbert states: ''members of human societies have the ability to monitor and orient to the emergent ⟨positive⟩ features of their own societies'' (Gilbert, 1995). But the problem is how this is possible.

There are four basic possibilities.

7.1.1. Passive and unconscious exploitation by the macro-system

In this case, the system is also unaware of the mechanism that produces the positive effect, and even of its being an effect of some micro-level behaviour.

For example, a primitive society can be completely unaware of the technical division of labour. One could take this as a 'natural' status, or ignore the process of micro-decisions that creates and reproduces such a specialisation. Since the effect is positive and self-reproducing, perhaps it is even more likely that people/the community do not have the problem of understanding it (it is necessity that pushes for intelligence!). Thus society does nothing to change or prevent the mechanisms from operating. This is a passive and unconscious form of taking advantage of functions. In this case the society or organisation is like individuals: it is just reinforced by, and exploits, the (positive) effects of functions, but *the latter do not emerge cognitively* at any level (Castelfranchi, 1998c).

Of course, if the model applies to the positive cases (eu-functions, advantages) it should also work in negative cases (disadvantages, dys-functions). In fact, there are several examples of self-reproducing bad global effects the macro-system is not aware of — at least for a certain period (Boudon, 1977).[22] Consider, for example, the individual and group unawareness of the cue traffic effect of their slowing down to look at an accident. Also, public authorities remained for a long time unable to understand such phenomena and their causes; the same is true in economics, etc. However, when the effect is very dangerous, the probability of some attempt to understand and manage it becomes greater (e.g., footnote 17).

7.1.2. Passive and conscious exploitation by the macro-system

Sometimes, the macro-system becomes aware of the positive effects of certain micro-behaviours and is in favour of their reproduction (e.g., mating). However, it cannot do anything to increment the process. So there is just a passive acceptance of the phenomenon. *Consciousness does not modify in this case the merely functional nature of the mechanism*, since *there is true intention only when/where there is the power of doing something*, and there is a decision to act.

7.1.3. Passive intentions at the macro-level

There is a 'passive' intention when an agent is aware of certain consequences of her actions, or of the effects of certain events, and she believes that she could avoid those effects (Section 5.2), but she decides (for several possible reasons, for example because she likes these effects) to let them happen. This is also possible at the level of the social system. A society or organisation can understand certain unintended effects of the actions of its members (e.g., communities of practice formation; or political discussions at pubs), and it has the means to interfere with and prevent them, but, since they are virtuous or irrelevant, or the intervention is too costly, it decides to let them go.

7.1.4. Blind cooperation between members and the macro-system

What is a function from the agents' point of view could be an intention from the macro-system point of view. The macro-system acts deliberately and consciously in order to reinforce and obtain such a behaviour by its members, but in so doing it does not use the members' understanding and their intentionality. Members remain unaware of the ultimate end of their behaviour, their internally represented goals are different from the relevant effects of the actions which really motivate the reinforcement they receive. Looking at their minds (Castelfranchi, 2000b) we just have a *social function* unintentionally but not accidentally reproduced (guaranteed) by their behaviours. Conversely, looking at the macro-system level, we have conscious planning to exploit such behaviour and such an ignorance.[23] Notice that, in

[22] Obviously, not all the combined (bad) effects of collective behaviours are (dys) *functions*. There are many global effects that are also repeatedly produced by agents just because they are not understood; others are reproduced although known by the agents just because they are related to actions that have more important desired results.

[23] Consumption activity by livestock does not cease to be just a phase of the production process only because the animals enjoy what they are eating (Marx, K., *Capital*, Vol. I, appendix, 1987). In the same way, workers accomplish their individual consumption activity for themselves, not for the benefit of the capitalists. However, they are in fact accomplishing a function of the system. Notice that even if workers discovered that they are also 'reproducing the labour force' for the capitalist, they could not do differently, or not enjoy their food.

this case, the feedback reinforcing mechanism from the global level back to the individual is not due to the same combined effects considered in the previous examples: this feedback is an *action* of the macro-system *aimed to* control the individual behaviour.

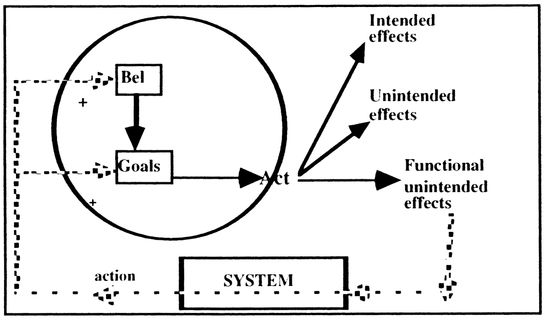

Fig. 8.

There should always be some 'closure', some way down (from the global or macro-system level to the individual mind), but in this case the individual behaviours (beliefs, goals) that are useful to the system (and reproduce it) are — thanks to their understood effects — reinforced and reproduced by the system: prescriptions are one means, but there are others such as socialisation, imitation, incentives, etc. Let us sketch a couple of brief examples.

Consider a mother and her child.[24] The mother wants the child to brush his teeth every evening, in order to avoid decay. The child does so in order to obey the mother and to make her happy; he ignores and could not understand the real function of his behaviour. What, relative to the intentional behaviour and the mind of the child, is just a function, in the mother's mind is an intended goal.

Exactly the same kind of relation often holds between government and citizens (Castelfranchi, 1991). Government pushes citizens to do something it considers necessary for the public utility, for some common interest, but it asks the citizens to do this by using rewards or sanctions. It does not rely on the citizens' understanding and sharing of the ultimate functions of their behaviours, and on their concern for the public welfare; it relies on the citizens' motivation for money or for avoiding punishment.

If we admit this very important kind of 'social function', we have to understand that it cannot simply 'emerge' from micro-interaction: *it should also 'emerge' cognitively* (become conscious at least at a certain level), and be deliberately reinforced/maintained; it can also be created artificially.

Authorities or groups could design certain functions of the actions of their members who will participate in such a process without being aware of it, or at least intending the overall phenomenon and its functions. This is, for example, the case of 'orchestrated or out designed cooperation' (Castelfranchi & Conte, 1992).

7.2. Norms and functions

This will bring us to very interesting hybrid social objects. Many social norms, for example, could remain mere 'functions' for the agents *as for their effects and aims*, while for the 'legislator' the norm's effects are supposed to be intentional.

As already stressed, norms, to work as norms, cannot remain unconscious in the addressee: the agent should *understand them as prescriptions and use them as such.* But in many kinds of norm-adoption (Conte & Castelfranchi, 1995), the agent can remain absolutely ignorant of the emerging effects of the prescribed behaviour.[25] In this case, from the point of view of the macro-system or of a legislator, the functional effects are in fact intentional effects; but from the side of the norm-regulated agents we have a mere 'function of the norm'. It is not strange to admit that *social norms have and guarantee social functions.* The same is true for intentions. Normative behaviour has to be intentional and conscious: it has to be based on knowledge of the norm (prescription), but this does not necessarily imply consciousness and intentionality relative to all the *functions of the norm.*

[24] To be true, the mother is not a system the child is a 'member of', however it is a system controlling the behavior of the other and 'putting' her own goals (tasks, and norms) above it.

[25] In some forms of norm-adoption, at least some of the functions of the norm are conscious and pursued by the agent. So, in the strict sense, one could refuse the term of 'function' (but see Section 8.2).

8. Why Elster and Hayek are wrong

On the basis of this cognitive characterisation of functions, let me now summarise the main points that Elster's criticism and proposal about the notion of function do not take into account. I will also discuss the limits of Hayek's view of spontaneous order as necessarily advantageous for the agents.

8.1. Elster's limits

Elster accurately identifies the conceptual basis of a methodologically correct notion of function, in the feedback mechanism and in the non-intended reproduction and reinforcement of behaviour. But he artificially limits his notion to the 'physiological' metaphor, i.e. to the hetero-functions subordinated to the interests of the overall social system. He ignores the possibility of effects that self-organise and reproduce independently of the goals or the advantages of the system, and bypasses the quite complex problem of who is really advantaged or not — and has power or influence on the system — when there are 'collective advantages'.

A theory of functions presupposes a theory of intentions since the latter compete with the former; Elster is perfectly right on this. However, Elster does not really have a theory as sophisticated as needed (Section 5.2). For example, his theory does not take into account *Side and Passive Intentions*. In fact, what is necessary for a theory of functions is that *the functional effect be not intentional* in the strict sense: it should not deliberately orient the behaviour, and should not be necessary or sufficient to act. However, is not necessary that the functional effect is unknown.[26] What is relevant for a theory of intentions is precisely that:

> *not all the known and anticipated (expected) consequences of an intentional action are intentions directing the action itself.*

Consider, for example, a mother with an excitable baby. She certainly does not want to spoil her baby. She is just upset by the fact that the baby wakes several times during the night, and has tantrums to go into his parents' bed. However, after some time she cannot bear the baby's crying and resists his pretence. By doing so she is unwillingly *reinforcing* the disliked behaviour of her baby and also the baby's next intention. It is a *vicious circle* ('vicious' relative to the mother's desires and values). Her behaviour has acquired a (kako)function that reproduces the behaviour in time. (From the baby's point of view this is in between a (eu-)function and an intention: it depends on our theory of the baby's mind; I personally believe that it is just instrumental learning, not really intentional behaviour.) Let us now suppose — as in fact is the case with many educated mothers — that the mother understands that she is herself creating this bad effect; nevertheless, she cannot resist doing so. In this case there is really a choice between *desired intended effects* (baby happy, stop crying) and damages (spoilt son), and the intended advantage must be higher than the *perceived* damages (as claimed by Elster). However, the expected negative effect — although anticipated — *is not an 'intention'* of the mother, while stopping the baby's crying is an intention. Nevertheless, the behaviour exists and reproduces itself also thanks to its negative effects and thus *in view of* it: it is a (kako/eu)function of that behaviour which is in fact *goal-oriented but not goal-directed towards such an effect*.

This not only applies to kako-functions, based on undesired but anticipated selective effects (Negative Passive or Side Intentions), but also to eu-functions.

Therefore, we can claim that:

> If x intends to do act α for G1 p and also predicts that effect q is likely to occur, and this effect q is good for x (i.e. there is another goal of x G2 — independent of G1 — implying q) this *does not* imply that G2 q turns into an intention of x governing α. It is a Side Intention (a special kind of Passive Intention — see Section 5.2).

[26] Boudon (1977) states this in his definition of 'perverse effects' discussing Merton's view. While Merton uses the term 'unanticipated consequences' (which puts together 'unknown' and 'unintended'), Boudon prefers the notion of 'undesired — although perhaps desirable — consequences', which I interpret as 'unintended'. However, Boudon also does not provide us with an analytical view of the different cognitive possibilities: Intended⇒Known, Not (Known)⇒Not (Intended), Known and Not (Intended), etc., and their combinations with desirability (Section 5.2).

This is an old and well known issue in ethics (Seneca): if I do charity and I feel good when I do charity, and I am aware that if I do charity I will feel good, this does *not* necessarily mean that my feeling good *motivates* my action, that I act *in order to* feel good. This does not mean that my *intention* in doing the action is to feel good.

My doing the action is conditional to my belief that it produces/allows p, and to having goal p,[27] while it is not conditional to the expected q. In other words, the expectation that p is necessary and sufficient for acting and activates my behaviour: it motivates me.

We can admit that *positive results can/will reinforce (as positive rewards) the behaviour without becoming their goal and intention.* Thus we could even admit conscious eu-functions: they are not conscious *as* functions, but just as effects. Although conscious, these functions do not become intentions.[28]

In sum, *both negative and positive effects that become functions might be known and expected by the agent, without becoming true intentions.*[29]

Any action reproduces thanks to its effects (through reinforcement learning, selection, restatement, or through understanding). Usually, in cognitive agents an action reproduces thanks to the effects that are *understood* (correctly perceived and attributed to or associated with the action) and then *expected* and *intended*. Nevertheless, there could be effects that, although not understood (and thus neither intended nor chosen), can reproduce the source-action. They reproduce and re-candidate the action thanks to the re-production and re-creation of some of *the internal or external conditions that activate, motivate or select that action*. However, *it is not true that, in order to reproduce, the effects must be 'good', i.e. useful for the goal of the agent or the macro-system*. This is where Elster and Hayek converge. On the contrary, effects can be good just in the self-referential sense that they are successful in reproducing themselves. Also, bad effects (i.e. bad from the point of view of the agent's goals or functions, or of the system) can self-organise and self-reproduce. This is what has to be proved in order to defeat the optimistic view of spontaneous social order and social functions.

8.2. The last challenge to Elster's claim: intentions overlapping functions (the conscious social actor)

Throughout this paper I have used Elster's claim (that if there are positively appreciated results of a behaviour and the latter reproduces thanks to the former, the notion of 'function' must be replaced by that of 'intention') as a useful methodological caution, and for making our problem and solution harder and more solid: *functional behaviour can be fully intentional although the functional effects are not intended*.

Now in order to provide a more complete view of the relationships between behaviour intentions and functions, I also argue against Elster's claim. It is true that in some cases the intentional explanation makes the functional explanation superfluous and redundant. However, they are not incompatible.

It is obvious — after Occam — that a simpler theory is a 'better' theory. However, since intentions and functions are independently founded and necessary for explaining different facts, why shouldn't/couldn't they coexist? Nature, mind and society are frequently redundant. Thus, however unnecessary, intentions can *overlap* functions, i.e. social actors can consciously and intentionally satisfy their social functions while those remain also 'functions'. For example, if one has sex for the purpose of having offspring (and also because one believes that this is the use and biological aim of sex), does her/his

[27] INTEND x to Do a for p = INTENDxp and INTENDx Do a and Intend x to do a, if and until x believes that a will produce p. For a more precise analysis of this conditional relation see Cohen and Levesque (1990), although they do not have a theory of *expected but non-intended results*.

[28] Even more than this: a function can be intentionally pursued by an agent but this does not eliminate its nature of 'function', because the intention **does not create** the finalistic behaviour, is **not necessary** for its functioning and reproduction; it is just **additional** and optional. This definitively overcomes Elster's objection (see below).

[29] Conversely, not all Passive Intentions (negative or positive) are functions. In fact, to be (come) a function, the expected result must have reproduced that behaviour through some feedback in the past. So, first of all it cannot be accidental but it must have occurred quite systematically in the past.

intention eliminate the biological function of her/his behaviour? Not at all, because that function precedes and is independent of the subject's intention: it does not require and is not affected by that intention. If a father understands the social functions of the behaviour of a father, internalises them, and realises them on purpose when acting as a father, does this make the social functions of fatherhood just a subjective intention? If awareness and intentionality do not dissolve biological functions, why should an overlapping intention eliminate the functional character and mechanisms of social functions?

The presence of intentions does not eliminate the functional mechanism. If this is true, the model depicted in Figs. 6–8 would represent sufficient but not necessary conditions for a functional behaviour. The 'functional effect' can occasionally be understood and even intended and motivate the behaviour, while remaining 'functional' both historically[30] (since its origin has not been intentional) and practically: being intended is not the necessary condition for, and the true mechanism of, its reproduction.

8.3. Hayek's optimistic view of selected effects and self-organising order

As already observed, Hayek grasps and preserves Smith's intuition that what we unconsciously and unintentionally pursue are 'ends'. However, Hayek does not provide any explicit and clear theory of such a teleology. Indeed, thanks to his subjective individualism, he basically identifies it with and reduces it to the psychological, subjective ends of the individuals, although pursued only unconsciously. He in fact assumes that emerging social structures are self-persistent and stable precisely because they allow the satisfaction of individual desires and conscious finalities.

Moreover, Hayek, like Smith, while recognising the goal-oriented character of his 'spontaneous order' or of the 'invisible hand', is not able to avoid an optimistic, beneficial, providential view of such a self-organising process. His ideologically positive and optimistic view of spontaneous order (Castelfranchi, 2000a) is allowed by a limited model of goals, intentions, and actions. As shown, a subtler cognitive theory of action is needed to account for the emergence, self-organisation and reproduction of dys-functions and noxious functions in human behaviour. Even Boudon's notion of 'perverse effects' is insufficient because it does not take into account their teleological or functional character. Hayek does not explain clearly enough *for whom* the emergent order should be good and how much the differences of power are responsible for its reproduction; he does not analyse the problem of the effects of our actions that are negative just for others; he does not account for the possibility that the social actors ignore their own interest; he bypasses the fact that desires and preferences (relative to which the 'order' is good) cannot be assumed as given but should be considered as produced by the order itself; he seems to use unclear models of group-selection, etc.

The crucial problem of this work is the following:

> *Is it possible to acknowledge and account for the goal-oriented, functional, teleonomic character of the 'invisible hand' without adopting a teleological and providential view of society or of history?*

The thesis of this paper is that it is both possible and necessary to provide a theory of processes and behaviours which is teleological and functional (both at the individual and social level) without being either intentional or simply casual or causal.

It is possible that actors who intend the results of their actions and prefer what is better for them, not only produce side-effects that are perverse and noxious, but let those effects self-organise and drive their own behaviours. The mechanism underlying the invisible hand, spontaneous order, precisely because it is non-deliberate, but just emergent, self-organising, self-reproducing (through individual behaviours) and self-referential, is basically *indifferent to the desires and the welfare of the individuals*. The resulting function can be *aimed at* either their good

[30] One should not forget that the notion of 'function' is a historical, evolutionary notion.

or at their evil.[31] In this perspective, Hayek's optimism is theoretically unwarranted and unjustified.

9. Concluding remarks

I hope that, after this long and tangled argumentation, it will be clearer why only Computational Social Science and, in particular, Multi-Agent-Based Social Simulation (SS) could probably deal with this kind of problem. Moreover, the task of SS is not only to predict emerging social effects or the experimentation of possible policies. I believe that the contribution of SS to the *theoretical* development of the cognitive and social sciences could be really remarkable. SS can provide not only an experimental method, but good operational models of cognitive 'actors', of the individual social mind, of group activity, etc. Models that can be richer, more various, and more adequate than those provided by economics, without being less formal. In particular, my focus on the core relation between functions and cognition was aimed at pointing out how the coming 'agent-based' approaches to social theory, using learning but deliberative agents, could deal with very old and hard problems of the social sciences and could re-orient them.

It seems possible to arrive — through the use of simulational models of minds and societies — at an operational notion of function that makes it scientifically sound and heuristic and improves our understanding of spontaneous and unaware functional cooperation among intelligent agents. In fact, my opinion is that the problem of reconciling in a principled way Cognition and Emergence will be the main challenge for the cognitive sciences in the next decade (Castelfranchi, 1998a,d). In other words, the problems to be solved are not spontaneous order or emergence *per se* (for example, with neural nets or rule-based agent), but accounting for them among intentional and rational agents.

This work has also tried to clarify how and why placing learning within a cognitive architecture is a very necessary approach for obtaining merely emergent intelligence and organisation among intentional agents; although my attempt to put them together may be debatable and yet unclear.

A final aim of this work has been to explain why a general theory of functions must be founded on the evolutionary model rather than on the physiological one, and why the basic notion of social functions must also include dys-functions and kako-functions. Self-organising social processes — not being chosen — are indifferent, in principle, to the agents' or group's goals and welfare; they are not necessarily 'functional' in the sense of 'useful', advantageous for something and somebody. Since the effects reproducing the behaviour are not realised and appreciated by the subject there is no reason for assuming that they will necessarily be 'good' for his/her needs or aims, or good for society's aims. Also, bad effects (for the individual or for the society) can reproduce by restating their conditions or reinforcing the behaviour. These are 'vicious circles' in individual and social behaviour. Contrary to Smith's and Hayek's claims, the emerging 'spontaneous social order' and the effects of the 'invisible hand' are not necessarily good and beneficial for the agents, although they can be self-organising and stable (Castelfranchi, 2000a).

Acknowledgements

This research forms part of a 20-year project pursued at the IP-CNR aimed at reconciling scientific teleological approaches and a theory of goal-governed agents with the theory of goal-oriented systems and of functional activities (Castelfranchi, 1982). This work on functions would not be possible

[31] Here Leopardi's view is opposed to Hayek's view.

> "*It is true that many things in nature proceed well, i.e. they proceed in such a way that they can preserve themselves and maintain, while otherwise they couldn't. However, an infinite number of them — both moral and physical things —* (*and perhaps a greater number than the former*) *proceed quite badly, and are badly orchestrated, with an extreme discomfort for creatures* . . . . *Nevertheless, since these things do not destroy the current order of things,* ***they go naturally an regularly badly****, and are a natural and regular evil.*" (*Zibaldone*, 4248, 18 *Feb*. 1823). " . . . *The whole nature, and the eternal order of things is in no way directed towards the happiness of sensible beings or of animals. Indeed it is contrary to their happiness. Nor their own nature and the internal order of their being is directed to that*" (*Zibaldone*, 4133, 9 *April* 1825).

without years of reflection and collaboration with Rosaria Conte and Maria Miceli on these topics. I would like to thank Rosaria, Maria, and Rafael Bordini for their precious comments. I am also in debt to Luke Kaven for several discussions on the subject and very useful philosophical suggestions that, for my limits, I was not able to really manage in this paper. A preliminary version of this paper was presented as an invited talk at *Simulating Societies '97*, Cortona. I would like to thank all the people who gave me comments, in particular Nigel Gilbert, Michael Macy, and Gerhard Weisbuch. This research was accomplished within the FIRMA European project (Freshwater Integrated Resource Management with Agents), 5th Framework, and within the ALFEBIITE: A logical framework for ethical behaviour between infohabitants in the information trading economy of the universal ecosystem – IST-1999-10298.

References

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

Alexander, J. C., Giesen, B., Muench, R., & Smelser, N. J. (Eds.), (1987). The micro–macro link, University of California Press, Berkeley.

Anderson, J. R. (1983). *The architecture of cognition*, Harvard University Press, Cambridge, MA.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist 54*, 462–479.

Beckermann, A., Flohr, H., & Kim, J. (Eds.), (1992). Emergence or reduction? Essays on the prospects of nonreductive physicalism, Walter de Gruyter, Berlin, pp. 25–48.

Bicchieri, C. (1990). Norms of cooperation. *Ethics 100*, 838–861.

Birckhard, M. H. (2000). Autonomy, function, and representation. http://www.lehigh.edu/~mhb0/autfuncrep.html.

Bobrow, D. (1991). Dimensions of interaction. *AI Magazine 12*(3), 64–80.

Boudon, R. (1977). *Effects pervers et ordre social*, PUF, Paris.

Bourdieu, P., & Wacquant, L. (1992). *An invitation to reflexive sociology*, University of Chicago Press, Chicago.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence 47*, 139–159.

Buller, D. J. (1999). Etiological theories of function: a geographical survey. In: Buller, D. J. (Ed.), Function, selection, and design, State University of New York Press.

Campbell, D. T. (1974). Evolutionary epistemology. In: Schilpp, P. (Ed.), The philosophy of Karl Popper, Open Court, La Salle, pp. 413–463.

Carley, K. M. (2000). Computational social science: agents, interaction, and dynamics. In: Invited talk at the WS on 'The simulation of social agents: architectures and institutions', University of Chicago, October 6–7.

Castelfranchi, C. (1982). Scopi esterni (External ends). *Rassegna Italiana di Sociologia XXIII*, 3.

Castelfranchi, C. (1991). Social power: a missed point in DAI, MAS and HCI. In: Demazeau, Y., & Muller, J. P. (Eds.), Decentralized AI, vol. 1, Elsevier, Amsterdam, pp. 49–62.

Castelfranchi, C. (1996). Reasons: belief support and goal dynamics. *Mathware and Soft Computing 3*, 233–247.

Castelfranchi, C. (1997a). Individual social action. In: Holmstrom-Hintikka, G., & Tuomela, R. (Eds.), Contemporary theory of action, vol. II, Kluwer, Dordrecht, pp. 163–192.

Castelfranchi, C. (1997b). Challenges for agent-based social simulation. The theory of social functions. In: Invited talk at SimSoc'97, Cortona, Italy, IP-CNR, TR. Sett. 97.

Castelfranchi, C. (1998a). Modelling social action for AI agents. *Artificial Intelligence 103*, 157–182.

Castelfranchi, C. (1998b). Through the minds of the agents. *Journal of Artificial Societies and Social Simulation 1*(1), http://www.soc.surrey.ac.uk/JASSS/1/1/contents.html.

Castelfranchi, C. (1998c). Simulating with cognitive agents: the importance of cognitive emergence. In: Sichman, J., Conte, R., & Gilbert, N. (Eds.), Multi-agent systems and agent-based simulation, Springer, Berlin, pp. 26–44, LNAI 1534.

Castelfranchi, C. (1998d). Emergence and cognition: towards a synthetic paradigm in AI and cognitive science. In: Coelho, H. (Ed.), Progress in AI, IBERAMIA'98, Springer, pp. 13–26, LNAI 1484.

Castelfranchi, C. (2000a). Per una teoria (pessimistica) della mano invisibile e dell'ordine spontaneo (For a pessimistic theory of the invisible hand and spontaneous social order). In: Rizzello, S. (Ed.), (a cura di) Organizzazione, informazione e conoscenza, Saggi su F.A. von Hayek, UTET, Torino.

Castelfranchi, C. (2000b). Through the agents' minds: cognitive mediators of social action. In: Mind and society, Rosembergh, Torino, pp. 109–140.

Castelfranchi, C. (2000c). Affective appraisal vs. cognitive evaluation in social emotions and interactions. In: Paiva, A. (Ed.), Affect in interactions, Springer, Heidelberg.

Castelfranchi, C., & Conte, R. (1992). Emerging functionalities among intelligent systems: co-operation within and without minds. *AI & Society 6*, 78–93.

Cavalli Sforza, L., & Feldman, M. (1981). *Cultural transmission and evolution. A quantitative approach*, Princeton University Press, Princeton, NJ.

Chattoe, E. (1998). Just how (un)realistic are evolutionary algoritms as representations of social processes? *Journal of Artificial Societies and Social Simulation 1*(3), http://www.soc-.surrey.ac.uk/JASSS/1/2/3.html.

Cialdini, R. (1993). *Influence. The psychology of persuasion*, Q.W. Morrow, New York.

Cohen, P. R., & Levesque, H. J. (1990). Rational interaction as the basis for communication. In: Cohen, P. R., Morgan, J., & Pollack, M. E. (Eds.), Intentions in communication, MIT Press.

Conte, R. (1985). Ancora sul funzionalismo nelle scienze sociali. Roma: IP-CNR, TR.3, 85.

Conte, R. (2000). The necessity of intelligent agents in social simulation. In: Invited talk at ICSS&SS-II, TRIP-CNR, Paris, September.

Conte, R. & Castelfranchi, C. (1995). *Cognitive and Social Action*, UCL Press, London.

Conte, R., & Gilbert, N. (1995). Introduction: computer simulation for social theory. In: Gilbert, N., & Conte, R. (Eds.), pp. 1–15.

Cummins, R. (1975). Functional analysis. In: Buller, D. J. (Ed.), Function, selection, and design, State University of New York Press, 1999, pp. 57–83.

Damasio, A. R. (1994). *Descartes' error*, Putnam's Sons, New York.

Drogoul, A., Corbara, B., & Lalande, S. (1995). MANTA: new experimental results on the emergence of (artificial) ant societies. In: Gilbert, N., & Conte, R. (Eds.), pp. 190–211.

Edmonds, B., & Dautenhahn, K. (2000). Starting from society — the application of social analogies to computational systems (special issue). *Journal of Artificial Societies and Social Simulation*, 3 (http://www.soc.surrey.ac.uk/JASSS/3.html).

Egidi, M., & Marengo, L. (1995). Division of labour and social co-ordination modes: a simple simulation model. In: Gilbert, N., & Conte, R. (Eds.), pp. 40–58.

Eisenstadt, S. N. (1990). Functional analysis in anthropology and sociology: an interpretative essay. *Annual Review of Anthropology 19*, 243–260.

Elster, J. (1982). Marxism, functionalism and game-theory: the case for methodological individualism. *Theory and Society 11*, 453–481.

Ferber, J. (Ed.), (1995). Les systemes multi-agents, InterEditions, Paris, p. iia.

Forrest, S. (Ed.), (1990). Emergent computation, MIT Press, Cambridge, MA.

Gasser, L. (1991). Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence 47*, 107–138.

Giddens, A. (1984). *The constitution of society*, Polity, Cambridge.

Gilbert, N. (1995). Emergence in social simulation. In: Gilbert, N., & Conte, R. (Eds.), pp. 144–156.

Gilbert, N., & Conte, R. (Eds.), (1995). Artificial societies. The computer simulation of social life, UCL, London.

Grosz, B. (1996). Collaborative systems. *AI Magazine*, 67–85.

Haken, H. (1988). *Information and Self-organization*, Springer, Heidelberg.

Hayek, F. (1952). *The counter-revolution of science: studies on the abuse of reason*, The Free Press, Glencoe.

Hayek, F. A. (1967). *Studies in philosophy, politics and economics*, Routledge & Kegan, London.

Hayek, F. A. (1973). *Rules and order*, Law, legislation, and liberty. A new statement of the liberal principles of justice and political economy, vol. I, Routledge and Kegan Paul, London.

Hedstrom, P. (1992). Is organizational ecology at an empasse? *Contemporary Sociology 21*, 751–753.

Hunhs, M. N., & Singh, M. P. (Eds.), (1998). Agent technology, foundations, applications and markets, Springer, Berlin.

Kurihara, S., Aoyagi, S., & Onai, R. (1997). Adaptive selection or reactive/deliberate planning for the dynamic environment. In: Boman, M., & Van de Welde, W. (Eds.), Multi-agent rationality. Proceedings MAAMAW'97, Springer, Berlin, pp. 112–127, LNAI 1237.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. T. (1990). On acting together. In: Proceedings of the 8th National Conference on Artificial Intelligence, Kaufmann, pp. 94–100.

Lomi, A., & Larsen, E. R. (1995). A computational approach to the evolution of competitive strategy. Working paper, London Business School.

Macy, M. (1998). Social order in artificial worlds. *Journal of Artificial Societies and Social Simulation 1*(1), http://www.soc-.surrey.ac.uk/JASSS/1.html.

Malinowski, B. (1954). *Magic, science and religion and other essays*, Doubleday, New York.

Masuch, M. (1995). Computer simulation. In: Nicholson, N. (Ed.), The dictionary of organizational behavior, Basil Blackwell, London.

Mataric, M. (1992). Designing emergent behaviors: from local interactions to collective intelligence. In: Simulation of adaptive behavior, vol. 2, MIT Press, Cambridge.

Mayr, E. (1974). Teleological and teleonomic: a new analysis; also appeared as: 1982. Learning, development and culture. In: Plotkin, H. C. (Ed.), Essays in evolutionary epistemology, Wiley, New York.

McFarland, D. (1983). Intentions as goals, open commentary to Dennet, D.C. Intentional systems in cognitive ethology: the 'Panglossian paradigm' defended. *The Behavioural and Brain Sciences 6*, 343–390.

Memmi, D., & Nguyen-Xuan, A. (1995). Learning and emergence: an introduction. In: Proceedings ECCS'95 — The first European conference on cognitive science, Saint-Malo, April.

Merton, R. K. (1949). In: Social theory and social structure, The Free Press, New York, p. 1963.

Miceli, M., & Castelfranchi, C. (1989). A cognitive approach to values. *Journal for the Theory of Social Behavior 2*, 169–194.

Miceli, M., & Castelfranchi, C. (2000). The role of evaluation in cognition and social interaction. In: Dautenhahn, K. (Ed.), Human cognition and social agent technology, John Benjamins, Amsterdam.

Miller, G., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*, Holt, Rinehart & Winston, New York.

Millikan, R. G. (1999a). Proper functions. In: Buller, D. J. (Ed.), Function, selection, and design, State University of New York Press.

Millikan, R. G. (1999b). An ambiguity in the notion of 'function'. In: Buller, D. J. (Ed.), Function, selection, and design, State University of New York Press.

Müller, H. J., Malsch, Th., & Schulz-Schaeffer, I. (1998). SOCIONICS: introduction and potential. *Journal of Artificial Societies and Social Simulation 1*(3), http://www.soc.surrey.ac.uk/JASSS/1/3/5.html.

Odell, J. (1998). Agents and emergence. *Distributed Computing October*, 4, www.distributedcomputing.com.

Parsons, T. (1962). *Towards a general theory of action*, Harper & Row, New York.

Parsons, T. (1966). *Societies*, Prentice-Hall, NJ.

Prietula, M. J., Carley, K. M., & Gasser, L. (Eds.), (1998). Simulating organisations: computational models of institutions and groups, AAAI–MIT Press, Menlo Park, CA.

Purton, A. C. (1978). Biological function. *Animal Behavior 26*, 653.

Radcliffe-Brown, A. R. (1957). *A natural science of society*, Free Press, Glencoe, IL.

Rosenblueth, A., & Wiener, N. (1968). Purposeful and non-purposeful behavior. In: Buckley, W. (Ed.), Modern systems research for the behavioral scientist, Aldine, Chicago, pp. 372–376.

Steels, L. (1990). Cooperation between distributed agents through self-organization. In: Demazeau, Y., & Mueller, J. P. (Eds.), Decentralized AI, North-Holland.

Stephan, A. (1992). Emergence — a systematic view on its historical facets. In: Beckermann, A., Flohr, H., & Kim, J. (Eds.), Emergence or reduction? essays on the prospects of nonreductive physicalism, Walter de Gruyter, Berlin, pp. 25–48.

Strube, G. (1999). Modelling motivation and action control in cognitive systems. In: Schmid, U., Krems, J. F., & Wysotzki, F. (Eds.), Mind modelling: a cognitive science approach to reasoning, learning and discmacroy, Pabst, Germany, pp. 111–130.

Sun, R., Merrill, E., & Peterson, T. (1998). A bottom-up model of skill learning. In: 20th Cognitive Science Society Conference, Lawrence Erlbaum Associates, pp. 1037–1042.

Sun, R. (1997). Learning, action, and consciousness: a hybrid approach towards modeling consciousness. *Neural Networks 10*(7), 1317–1331, special issue on consciousness.

Sun, R. (2000). Cognitive science meets multi-agent systems: a prolegomenon. Columbia: TR CECS Department, University of Missoury, April.

Sun, R., & Peterson, T. (1998). Some experiments with a hybrid model for learning sequential decision making. *Information Sciences 111*, 83–107.

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin 2*, 160–163.

Troitzsch, K. G. (1997). Social science simulation — Origin, prospects, purposes. In: Conte, R., Hegselmann, R., & Terna, P. (Eds.), Simulating social phenomena, Springer, Heidelberg.

Tuomela, R. (2000). *Cooperation*, Kluwer, Dordrecht.

Van Parijs, P. (1982). Functionalist marxism rehabilitated. A comment to Elster. *Theory and Society 11*, 497–511.

Virasoro, M. A. (1996). Interview on complexity, by Franco Foresta Martin. Trieste: SISSA, TR.

Wimsatt, W. C. (1972). Teleology and the logical structure of function statements. *Studies in the History and Philosophy of Science 3*, 1–80.

Wright, L. (1976). *Teleology explanations: an etiological analysis of goals and functions*, University of California Press, Berkeley.