

DRAFT - preliminary version
(unfinished translation from an Italian paper)

For a (Pessimistic) Theory of the Invisible Hand and Spontaneous Order

Cristiano Castelfranchi

Istituto di Psicologia - Consiglio Nazionale delle Ricerche *
Reparto di "Intelligenza Artificiale, Modelli Cognitivi e dell'Interazione"
cris@pscs2.irmkant.rm.cnr.it

"...La natura tutta e l'ordine eterno delle cose non e' in alcun modo diretto alla felicità degli esseri sensibili o degli animali. Esso vi e' anzi contrario. Non vi e' neppur diretta la natura loro propria e l'ordine eterno del loro essere. ... Gli enti sensibili sono per natura enti *souffrants*... poiche' essi esistono e le loro specie si perpetuano, convien dire che essi siano un anello necessario alla grande catena degli esseri, e all'ordine e all'esistenza di questo tale universo, al quale sia utile il loro danno, poiche' la loro esistenza e' un danno per loro, essendo essenzialmente una *souffrance*."
(Leopardi, *Zibaldone*, 4133, 9 April 1825).

0. Introduction

The main theses that I shall endeavour to defend are the following:

1. Hayek's view that the main, indeed the fundamental, problem of the social sciences is that of an order and of spontaneous institutions emerging as unnegotiated and unintentional effects of the actions of individuals is correct and of great importance.

In my opinion this theory is particularly topical today in view of the modernity of the approaches based on complexity, on dynamic systems, on the concept of emergence, and the antagonism with which they are juxtaposed to the dominant paradigm: 'cognitivism' (the mind as a symbolic system and as a teleological control over behaviour based on symbolic representations of the world and of action). However, I consider the true challenge not to be that of a paradigmatic revolution that replaces the symbolic approach with a dynamic one: it is rather that of a synthesis between the two (Castelfranchi, 1998a; 1998b; 1998c). At the social level this means exactly responding to Hayek's problem: how can it be that dynamic, emergent, non intentional effects govern the behaviour of cognitive and problem-solving agents whose actions are intentional and (at least partly) planned on the basis of their knowledge? I consider that the contribution to the science of the artificial (Artificial Intelligence - in particular Multi-Agent Systems, Artificial Life, Agent-based Social Simulation) will play a decisive role in solving this problem, which is the principal theoretical problem facing the social sciences (Castelfranchi, 1997).

2. Hayek grasps and retains the idea of Smith that what we unwittingly and unintentionally pursue are 'goals'. This intuition is fundamental (and, in my opinion, convergent with the problem of the theoretical foundation of the

* The present research was carried out within the framework of the collaboration project between the CNR and the Bulgarian Academy of Sciences on the topic "Cognition and Emergence", resp. Castelfranchi-Kokinov, as well as of ordinary IP-CNR projects.

concept of 'function' in the social sciences). Nevertheless Hayek does not provide an explicit and straightforward theory of this teleology and indeed, essentially - with his characteristic subjectivistic individualism - equates it with and reduces it to the psychological goals of individuals that may not necessarily be intentionally pursued. Indeed he assumes that the emerging social structures are persistent and self-reproducing precisely because they satisfy individual desires and individuals' conscious purposes.

Furthermore Hayek -like Smith- in acknowledging the teleological nature of the invisible hand and of spontaneous order, cannot help attributing to it a (positive) value judgment, a providential, benevolent, optimistic vision of this process of self-organization (that is -he of all people- commits the sin of ideologism).

The optimistic and positive view of spontaneous order, which I shall attempt to document, is based on his limited model of goals, intentions and actions. A much more sophisticated cognitive theory of action would be required to account for the emergence, organization and self-perpetuation of dysfunctions and malicious functions in human behaviour. Also the concept of 'effets pervers' (Boudon) is inadequate as it does not account for the actual teleological, functional nature (sect.8).

Hayek does not adequately explain *for whom* the emerging order is good and to what extent power differences are involved in maintaining it; he does not analyse the problem of the effects of our actions that are harmful to others and not to ourselves; he does not include in the theory the possibility that the social agents do not know what is best for them; he neglects the fact that desires and preferences are not a given (with respect to which the order is good) but are a product of the order itself; he makes use of models of group selection that are not clear, etc.

The central issue in this work is therefore whether it is *possible to recognize and account for the teleological, teleonomic, function character of the 'invisible hand' without having to adopt a teleological and providential view of society and of history.*

3. The proposed thesis is that it is possible and necessary to have a theory of teleological and functional behaviour (at both social and individual level) as a phenomenon distinct from both intentional behaviour and mere causal or chance effects. This is a distinction that Hayek intuitively grasped but did not resolve.

The fundamental problem is how to graft teleological but unintentional behaviours precisely on intention-driven behaviours. What answer can be given to Elster according to whom the idea of *intention* makes that of the *function* of behaviour impracticable and superfluous. How can intentional acts also be functional, that is, unwitting but *reproduced precisely as a result of their unintentional effects*. How can agents who intend their results and prefer what is good for them to what is bad for them not only produce also perverse and harmful side effects, but allow the latter to organize themselves and direct their behaviours.

The invisible hand, spontaneous order, precisely because it is an unintentional and yet emerging function, self-organizing and self-reproducing (through the individual behaviours), *is substantially indifferent to the goals and good of individuals*, and may be addressed equally to good and to evil (Leopardi's philosophic view is here opposed to Hayek's philosophy).

Hayek's optimism is theoretically unjustified.

1. From teleonomy to teleology?

As stated earlier, the central issue of the present work is the following:

Is it possible to acknowledge and theoretically account for the teleological, functional and teleonomic nature of the 'invisible hand' without having to adopt a teleological and providential view of society or history?

I argue that this is possible (and even necessary, in order to have an adequate scientific theory of spontaneous institutions and social functions), but that the greatest theoreticians of the invisible hand (from the Scottish philosophers down to Hayek) failed to disentangle teleonomy and teleology: they provide us with an optimistic and even providential view of the invisible hand as a *problem solver* and of spontaneous order as a public good. Nor did they succeed in distinguishing spontaneous, self-organizing finalism from subjective goals.

It is not important to understand whether this is for ideological reasons or (a non exclusive 'or') due to a (historically) limited theoretical baggage, in particular as far as the theory of mind, knowledge, intention and action is concerned.

Smith -in my opinion- clearly falls into the teleological trap, as does Hayek with his highly problematic concept of 'social order' (see sections 3 and 4).

Before discussing these limits, however, I should like to again emphasize the great importance of the intuition of the invisible hand, and the absolutely crucial nature of this theory for the social sciences.

As far as finalism is concerned, I shall argue that subjective finalism and *finalism without a subject* are two distinct concepts that can and must be disentangled and provided with an autonomous foundation, and that they are both scientifically practicable.

At the outset each teleological notion was incorporated in a finalistic conception of nature, man and history. No possible distinction could be made between what Mayr (1976) proposed to call 'teleonomy' precisely because he deemed it impossible to separate finalism from the concept of teleology (see also Wright, 1976; Wimsatt, 1972). This was the main reason for expunging 'final causes' from science for a long period of time. Modern scientific teleonomic concepts, contributed by both cybernetics and evolutionary biology, are not accompanied by any optimistic or evaluative view, or any end cause.

Nevertheless one unfortunate consequence of this distinction and foundation is actually the fact that self-referential teleonomic phenomena (such as spontaneous order, social functions, conventions, etc.) are not guaranteed to be functional to human needs, or to be good for subjective human purposes.

2. "THE core theoretical problem of the whole of social science": the "invisible hand" from Smith to Hayek

"This problem (of the unintentional emergence of order and of spontaneous institutions) is in no way specific to economics... *it is without doubt THE core problem of the whole of social science*" (Hayek, 1988)

I think that Hayek is completely right. But the problem is not simply how a given *equilibrium* is reached and how some stable *order* emerges.

Is this emergence a simple epiphenomenon? Is the "order" only in the eye of the beholder? (sect. 6). Emergence and spontaneous equilibria are not enough to ensure "spontaneous order" or an "institution. They must be 'functional'.

In my opinion Smith's original formulation of "THE problem" is much clearer and more profound provided he is taken seriously and literally.

The well-known problem of the "invisible hand" is not in fact simply a problem of the emergence of some equilibrium or other, or of composite, unpredictable, unintentional and stable effects. The essential question is how can it be that:

"(the individual) - generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it...., *led by an invisible hand to promote an end which was no part of his intention*" (Smith, *The Wealth of Nations*).

In this view:

- 1) there are intentions and intentional behaviours;
- 2) some unwitting and unintentional (complex and long-term) effects emerge from these intentional behaviours;
- 3) but these are not mere effects: they are "goals" that "we pursue", that is, that somehow orient and direct our behaviour: at a certain point we "operate" not accidentally but "necessarily for" this result (Smith);
- 4) and yet they are not part of our intentions.

In my opinion, this - with all its contradictions and technical difficulties - is the correct formulation of **THE** problem. And it is indeed a problem, because it is not clear:

- how can we *pursue* something that is not one of our intentions; how can the behaviour of an intentional planning agent be goal-oriented, finalistic, without being intentional; and

- in what sense can this unintentional effect of our behaviour be one of its "ends" .

The real problem is thus how to relate 'mental' and 'non mental' problems: how the intentional behaviour (action) can be - at a different level - simply goal-oriented, functional (sect. 6.2) (Castelfranchi, 1982).

This relationship is essential also for the Hayekian restatement of the same problem. In his view, the social institutions play the role of a collective mind¹ whose knowledge and decisions are distributed and partially unconscious (no central planning is capable of governing the development of society).

In this perspective, the problem is not so much one of understanding how this distributed control can function without a central planner or authority. Rather, a mind is characterized by goals ('control' and 'decision' are goal-oriented concepts). Therefore do *goals*, *ends* exist in society - and not just mere effects and equilibria - which are not simply the *intentions* of individuals? and how can these non-intentional goals regulate the behaviour of individuals?

¹ No true, problem-solving, planning 'collective mind' actually exists or can exist for Hayek. It is only an 'as if' construction, a convenient artifice.

In Hayek's view it is apparently actually only individuals (and small organizations) that can have true ends: the latter seem to be only subjective, explicit and deliberate (see 4.).

I believe that this same problem has cropped up in other social sciences as the problem underlying the notion of (social and biological) 'functions': how can functions regulate the behaviour of anticipatory and intentional agents? and what is their relationship with their intentions?

In actual fact, the same theoretical problems that have plagued the theory of functions appear in Smith's theory and in Hayek's view of spontaneous social order. For instance, the metaphor of society as an 'organism' or the connotation of behaviours as positive, useful and functional vis-à-vis the maintenance/evolution of this order, or the relationship with individuals' needs and their satisfaction.

One of the great theoretical breakthroughs we owe to Hayek is that of having characterized the process of emergence of social order and the formation of the institutions as an evolutionary process, in terms of adaptation and selection, although (except for the use of a rather pan-selectionist evolutionary model based on group-selection²) he gives a progressive view of this evolution. In his view, society selects and retains the positive results of experience after numerous attempts and errors. Only the positive features and the "correct rules" survive. The behaviours encouraging the development of the group are persistent and are replaced only when more efficient behaviours have been developed. All behaviours proving antithetical to the group cannot persist and are eliminated (Hayek, 1973). Therefore, what emerges of a stable nature is in the general interest, and favours the development of the group.

In the present work I shall first develop a criticism of this optimistic and tendentially teleological interpretation of the invisible hand and of spontaneous order (sect. 3 and 4) which in my opinion cancels out the ingenious discovery that it consists of an "end". I shall then attempt to provide a solution to the problem of the two finalisms, by means of a 'neutral' and self-referential theory of 'functions', and their relationship with intentions.

3. Friederic von Hayek's 'good' order

Smith explicitly states that the resulting equilibrium is in the "public interest" and to the good of the nation, and it is clear that the invisible hand providentially guides us to unconsciously pursue the general good. Moreover, he is the apologist of Mandeville, of private vice and public virtue: selfishness produces the general good.

It is more surprising - in view of his sharp mind and critical spirit - to find in Hayek (as it seems to me) exactly the same fallacy as in Smith³. The emerging order is transformed from a mere dynamic equilibrium into something 'good' for mankind (and even something that one cannot seek to improve). It without doubt constantly acquires a positive *connotation*.

As I said earlier, the assumption on which Hayek's thesis is based is the classical "Scottish" one that intentional human actions produce unexpected and unintentional consequences, and that an equilibrium, a social order, emerge as the result of individuals' actions, without being consciously conceived of and pursued by (any) of them. And even more than this: Hayek actually adopts Smith's ingenious thesis according to which the emerging order is an 'end' we pursue (we are induced to pursue) even though we do not intend to. But while this time this spontaneous order, an unplanned dynamic, *equilibrium*, is presented simply/ neutrally as the "existence of certain *regularities* spontaneously produced without any planning or deliberation" (*The Counter-Revolution of Science: Studies on the Abuse of Reason*, 1952; abbreviated to AR, chap. IV), or as the fact that human movements "tend to conform to a well-defined *model*... that has not been planned by anyone" (ibidem); this notion very often (implicitly or explicitly) takes on a positive evaluative connotation (which among other things is already associated - and this is no coincidence - with the word 'order' used in its everyday meaning)⁴.

² Our civilization is based on a set of behavioural and ethical rules, the product of spontaneous evolution and selection based on the advantages of "a relative growth both in population and richness of those groups whose it happens to follow them??" (*The Fatal Conceit*). There is clearly a pan-selectionist view and a group selection approach. For some criticisms in this sense, see Vanberg, 1986; Rizzello, 1997.

³ A number of researchers claim that Hayek was "misunderstood" on this point. But what this actually shows is that Hayek is ambiguous and contradictory. It would be more correct to say (e.g. Rizzello, 1998) that his thinking contains "controversial" elements. However, this implies that there is not someone who understood them while the others "misunderstood" them.

⁴ And it is also associated with the term 'equilibrium' and the term 'regularity' but not, for example, with the term 'structure'. For this connotation see, for example: "If social phenomena do not manifest other *orders* except that which they receive from a conscious intentionality.... <on the contrary> a given type of *order* emerges as a result of the action of the singles, without being consciously pursued by any of them" (AR). Note the comparison between the two orders (the deliberate one and the spontaneous one) and the use of the same term. There is no doubt that the first order is subjectively *positive* for the agent as it is intentional and the agent prefers and pursues what it believes to be best for it. It is thus implicitly assumed that also spontaneous order is positive for agents. In *The Pretence of Knowledge* Hayek does not refer simply to spontaneous order but to "our civilisation" and

This equilibrium, this neutral regularity (which at most should refer to a criterion of 'self-referential' functionality/goodness - sect. 6.1) is assumed to be 'good' for people ('relative criterion' - sect. 6.1), and also better than any possible mentally plannable alternative, and in any case such as to defend oneself from.

For example in AR (chap. VIII) Hayek criticizes the inability of numerous researchers to understand "how the independent actions of numerous individuals can produce *stable sets, stable structures of relations, that serve important human purposes*, even though no one has intentionally directed them towards this".

Whereas 'stable sets' and 'stable structures of relations' can be neutral and descriptive notions and are indeed linked to the criterion of functionality, which I define as self-referential or absolute, the second statement (repeated on several occasions is apparently an explicit *positive evaluation* and refers to the subjective good of persons (but see note 6). Hayek has to claim that spontaneity is greater than any possible social planning, and so the product of this spontaneity and selection *must* compete with the product of the social project. For what purpose, for what reason is an individual or collective project born if not for the common good and for the solution of problems? The invisible hand *must* be a problem solver, and what is retained and handed down must represent good solutions to (individual and/or collective) problems.

Hayek claims -and this is very important - that purposes may exist which do not coincide with or are not the result of any intentional plan. Nevertheless, he fails to give us an explicit definition and a theory of this notion, nor is he able to distinguish it clearly from the subjective purposes of individuals and to give it a solid grounding. Ultimately his concept of purpose is based (nor could it be otherwise) (AR VIII, p. 106) on subjective finalism. He stresses the possibility of a spontaneous "ordered and *finalistic* arrangement"⁵ and claims that "something ordered and *consistent with a useful end*" may emerge spontaneously; but the problem is the redundancy of his theory between 'end' and 'useful': his 'end' does not go so far as to become a (self-referential) end in itself; it is necessarily 'useful', otherwise it would not be an 'end'. It refers to the needs, to the advantages of agents who have internal goals (goal-directed entities). Likewise, Hayek's order, or equilibrium, or regularity, or structure, is not only something more than a mere perceptual gestalt and descriptive construction, but also something more than a self-organizing and self-reproducing structure (and therefore self-referential and self-'motivating'): it is 'useful, good for human purposes'.

Hayek endeavours to render a self-referential criterion explicit and to characterize finalistic but unintentional phenomena. He correctly states that, in these cases "the end ...always consists in the preservation of a system, of a structure of permanent relations", although he immediately adds: "<structure> the existence of which we are accustomed to accepting as an established fact", thus again reducing the 'end' to a mere *subjective evaluation*, which is not conceptually necessary for the system's preservation.⁶ And he concludes -correctly- by again referring to the

characterizes the attempted / illusory control over the development of society by acting against spontaneity as 'baneful'). On page 118 (AR IX) he claims that "sometimes thanks to their spontaneous interaction social forces solve problems that the individual mind could never solve consciously". Therefore spontaneous order is a (good) "problem solver"; and these problems are problems of individuals and so their solution is good for individuals.

On p. 140 (AR, X) he states that it is necessary to consider "each individual as part of a process where its contribution is not guided but spontaneous, and where it collaborates to the creation of something greater than anything that any individual mind could never plan". Note the positive connotation of terms like 'contribution', 'collaborate', 'creation', 'greater than'. Here indeed we are approaching a kind of inspired panegyric of the invisible hand! (which neglects the fact that also wars, poverty, pollution, racism, etc. are products of the same mechanism).

I have included a few quotations from Hayek (to try and avoid misrepresenting him), but I wish to make it clear that my intention is not critical-philological in nature. My aim is not to establish "what Hayek really said" (who like all great men said things that are subject to interpretation and are not static), but rather on the one hand to develop and on the other to criticise a thesis and a view to which he contributed substantially; even if it were demonstrated that this was only his 'vulgate'. It is necessary to distinguish between the sinner and the sin (as was said by a man of peace), and my interest is in the sin, in its abstractness.

⁵ It seems that every order is necessarily also finalistic, while this is not true: it is true only for intentional, planned orders, and for orders that reproduce themselves by means of selection and feedback. For these two types of finalism see Castelfranchi, 1982; Conte and Castelfranchi, 1996 chap. 8.r

⁶ Note that this criterion of the "goodness" of social order differs from the previous ones. Here it seems that the order is good simply because it is consolidated, traditional, customary! Which is not the same as saying that it is good because it satisfies needs or because it has prevailed over other orders. In his justification of the goodness of the social order Hayek is far from being systematic, consistent and profound. He introduces heterogeneous and *ad hoc* dimensions. I shall briefly outline here the different unrelated arguments put forward as required to justify the goodness of spontaneous social order:

- it is regular, stable, an equilibrium;
- we are accustomed to it, we consider it normal, as a reference level;
- it is our civilization;
- it represents a selective advantage for the group that adopts it; it has been selected as good because the societies that adopt it are superior;
- it guarantees the survival of the greatest number of persons, not individual satisfaction (cf. Nardi, in the present volume);
- since agents act and learn as a function of their well-being, the emerging order is good for their purposes;
- it is a result that no single mind could conceive of or a single agent pursue;
- it is better than any order that could be planned and implemented as a project;

notion of 'function' although in its *physiological* meaning, that is, in an 'organism', and to the idea that the parts have the goal of preserving their wholes. Note how also the idea of 'organism' adds positive connotations to spontaneous social order.

The strongest case made by Hayek concerning the goodness of spontaneous social order is that " these self-persisting wholes/social structures represent a *conditio sine qua non* for the attainment of many individual aspirations: they go to make up the environment that makes possible the conception of our individual desires and allows them to be satisfied.The emerging social structures are 'useful' as they form the premises for future human development" (AR VIII). These institutions "*are a necessary condition for the attainment of conscious human purposes*" (AR VII note 5).⁷ Therefore:

- in the first place, social order - although unconscious and unintentional - is functional, is good/useful only as a function of the individuals' conscious purposes. In fact, there is only one valid teleonomic notion: the psychological one! There can be no *autonomous* teleonomy deriving from an evolutionary perspective.
- secondly, it is implicitly assumed that these emerging social structures are self-persisting precisely *because* they are 'useful', *because* they realize the individuals' desires and their conscious purposes.

I consider this to be the main argument provided by Hayek to support his optimistic view of emerging order. This is the main premise that is to be criticised on the basis of a cognitivist theory of individual action.

Hayek endeavours to ground this thesis on an evolutionary theory (see note 2) as well as on a specific conception of the micro-macro link. He claims - if I have understood correctly - that the order which emerges is good, and the best available, because it is the result of a selection made by the agents with an eye to their own good, and to their own unremitting actions and reactions based on their own preferences (subjective good), and/or on learning and tradition based on positive reinforcement, the selection of benefits, on evaluations, on the imitation of the most suitable (Rizzello, 1997; 1998)⁸.

If the agents fight unremittingly for their own individual good an emerging equilibrium cannot be contrary to their good.

*If the emerging good was not good for the agents, it would not stabilize, it would not be maintained, the agents would (consciously or unconsciously) rebel against it, they would react in the direction of their own good*⁹.

It follows that not only does a natural, unplanned order exist, but this order is also useful for individuals, and perhaps the best available among the possible variants, since man selects post-hoc (but does not create - at least at the collective level). If - in view of the variations that occur - it was possible to achieve a better equilibrium for and among these agents, they would find it; if more suitable solutions emerged they would be handed down (cf. sect.4.5). This thesis contains a number of fallacies which I shall examine in section 4.

3.1 In what sense is Hayek "evaluative"

That Hayek adopts an *evaluative attitude* rather than a scientific one towards 'spontaneous order' may be challenged. It is thus necessary to clarify carefully what represents an evaluative attitude and what is the one that Hayek undubitably takes and what instead is mere suspicion.

The evaluative attitudes considered in social science theory are usually not very highly structured. It is at least necessary to use the distinction introduced in the cognitive theory of evaluation and values (Miceli and Castelfranchi, 1992) between "involved" evaluation and "third-party" evaluation. Cognitive evaluation always in any case implies

- it is like an organism that maintains itself and survives by virtue of its functions.

⁷ These theses are -in my opinion- extremely close to several elaborations justifying the notion of 'function' in social anthropology (e.g.. Malinowsky).

⁸ Also here what is needed is a more critical and pessimistic view: in the first place, what is imitated is not that which is the most suitable but what is *subjectively perceived* as such (the selective environment of memes are individual minds, with their biases and their conformism); secondly, is what is imitated the most suitable, the most successful solution? or rather is it what is the most imitated that is the most suitable and the most successful?

⁹ This is a conception that has a long history, and which finds, for example, a very explicit precedent in Constant (who is also against the idea of a general 'will' that shapes society): "Ever since society has existed, certain relations are set up among men; these relations are in keeping with their nature since *if they were not in keeping with their nature they would not be set up*. Laws are merely these relations, observed and expressed.. They are not the cause of these relations, which indeed precede them." (quoted in Dupuy, 1992, p.12).

expressing a judgment, a belief concerning the usefulness/goodness/appropriateness .../ the power of y (what is evaluated) vis-à-vis a given goal S¹⁰:

(Believes x (Good-for y S))
(Believes x (Goal z S))

That is the evaluating subject x assumes that someone (z) has a goal (desire, need, rule, aspiration, etc.) and believes that y is good for that goal. However, this evaluation may be the neutral and indifferent evaluation of an expert (a consultant, for instance) who does not have that goal (or even the worried evaluation of someone with an opposite goal to that of z); or else it may involve the evaluation of someone who has and pursues the goal S; this is the case when z and x coincide (and in particular when S is an important and activated goal of x).

When I state that Hayek has an *evaluative* attitude I am saying

- i) that there is no doubt that he adopts the “third-party evaluation” with reference to “subjective human purposes”; furthermore, this is an explicitly positive and - as I shall demonstrate - optimistic evaluation;
- ii) that he is sometimes more or less explicitly involved and slips into a 'personal' evaluation (in the connotations of the terms he uses, or when he speaks of defending civilization);
- iii) that it is easy to slip like this from one evaluative attitude to another in this field.

In other words, Hayek's 'technical/objective' evaluation (that is, his 'third-party' evaluation, referring to the goals and values of others, not to the goals and values of the evaluator himself) forms a reasonable and compelling basis for a non-neutral and positive attitude towards social order. Indeed, *if spontaneous social order is good for participants, what other well grounded criterion could ever be proposed to evaluate its goodness or badness?* On what arbitrary or authoritarian basis can it be criticized?; with reference to the values of what authority or superior external mind?

In my opinion Hayek clearly slips from an 'objective' evaluation to a subjective one, as the former provides the basis and the supporting argument for the latter. It must be made clear however that I am making a criticism not so much of any 'value judgment' of his but actually of his explicit basic evaluation. I consider that Hayek's vision is not lacking in ideology so much because of the possible 'value judgment' but because it is based on a certain dose of 'motivated reasoning' (Kunda, 1990) owing to the fact that it neglects - not by chance, and systematically - extremely important aspects and problems that would have acted as counter-arguments to his thesis of the functionality of spontaneous order.

It is true that he explicitly refuses to apply a criterion of justice or injustice (which would refer only to individual actions)¹¹, but he is actually applying a criterion of efficacy and efficiency, a criterion of utility, which must necessarily be used to refer also only to subjective and individual goals.

4. Hayek's fallacies (or Hayek versus Hayek)

Hayek seems (deliberately or unconsciously) to ignore several consequences closely linked to his own fundamental theses, as well as several obvious objections to his conclusions. This is truly surprising in view of the subtlety of his reasoning, his irony against ideologies, his realistic view of man and of the limits of reason.

In this sense I would like to say that he tends to jump to acceptable conclusions, to be blind to inconvenient but obvious facts.

4.1 Whose human purposes?

In the first place, it is odd that Hayek himself, the champion of subjective individualism and of the inconsistency of every collective entity should suddenly become oblivious - in all his arguments concerning what is 'useful' - to the fact that subjective good is referred only to a given individual, that individuals have different degrees of knowledge and different preferences (and utilities). From which it necessarily follows that what is good and useful for one may be bad for another.

Therefore, when we say that the emerging order is “*at the service of important human purposes*” or that it is a necessary condition for attaining conscious human purposes” we must immediately ask ourselves “the human purposes of which individuals?!", “good for whom?!”. It is not possible to overlook this in an individualistic framework. Do purposes exist for mankind? Therefore, *even if it were true that this order were not only natural but*

¹⁰ For how also the evaluation concerning rules, canons and standards derives from and is reduced to this, see Miceli and Castelfranchi, 1992.

¹¹ For Hayek to say that a spontaneous order is good means slipping back into the atavic anthropomorphism and animism of primitive thinking, which attributes the creation of all social forms to the will of some agent or other. The expression “social justice” is meaningless; just and unjust apply only to personal voluntary actions (Hayek, 1976).

also a good as far as the agents are concerned (and I will deny that this is really true), it would be a good for some agents, but not for all.

- First, it remains to be demonstrated that these agents represent the majority of agents. This is anything but obvious and does not follow from any theoretical premise, since we cannot overlook the fact that agents, just as they have different degrees of knowledge, also *have different powers*. An agent who is more, or much more, powerful than another will have a completely different effect on the maintenance or modification of an emerging order. A minority of powerful agents (with similar preferences and interests) can have a much stronger influence on the emerging institutions than the majority of unempowered agents. Was it not Hayek who said that "those who possess all the means establish all the ends" (or was this just a criticism of socialism)? Therefore, those who possess many means establish many ends. We cannot overlook this fact while we more or less explicitly postulate that social institutions, traditions and social order are advantageous for individuals (which?!); or else are advantageous for the group: does the group have its own ends, separate from those of individuals?
- Second, while allowing that social order is good for the majority of agents involved, to what extent does this make it better than other possible ones, or something to defend? And why should others accept it? Why should disadvantaged individuals accept it?
- Lastly, why would a scholar consider it good and defend it, embracing the point of view of the advantaged agents (even when the latter represent the majority)? Is not this a value choice?

4.2 Ignorance of good

The limits and the individual and distributed nature of the agents' knowledge have important implications which Hayek glosses over. The latter include the possible *ignorance of one's own good*. Agents do not understand or know what is good or bad for them, or whether something better is feasible. In order for someone to act or react it is necessary at least:

- i. to believe it is possible to change the present state of the world, that is, it is possible to achieve a different state;
- ii. to know of at least one feasible alternative to this state of the world;
- iii. to know that this is also preferable, and thus have an objective (however local) to pursue;
- iv. to know what to do to attain this objective (possible intention);
- v. to believe one is capable and that adequate resources, conditions and opportunities are available for the action;
- vi. to believe one is in a position to cope with any risks or failure.

Hayek's view of intentional action is highly limited and inadequate. He fails to consider many cognitive preconditions of the action. He takes these cognitive preconditions for granted as well as the fact that the agents are capable of perceiving and reacting appropriately to what is bad for them.

Men can in fact very easily put up with what is bad for them, be unhappy and desperate, but be incapable of seeing any way out or any possible way of reacting or changing the conditions of their discontent, whether they are aware or not of being in a difficult situation and that better ones possibly exist (although they may not know how or believe they do not have any *practical possibility* or the *right* to change it). This is so true that even psychological mechanisms exist for adapting to adverse circumstances (see, for example, Elster's *Sour Grapes*).¹²

Moreover - as Hayek claims- the emerging order mechanism acts globally; it is too complex and cannot be perceived by individual cognition or decision; agents have limited understanding and can only act locally. But, in actual fact, individual choices are highly limited locally and dependent on the social order itself and its global mechanism. Where does the individual get enough room, freedom to protect his good (and thus to unwittingly forge the emerging order) if the constraints on him are so great (and his very wishes are the expression of this order)? Therefore *such an order might not be maintained through its goodness-usefulness but rather as a result of its opacity and the strong constraints it imposes on the local conditions and individual choices*. Individuals adapt or resign themselves to the order much more readily than they adapt the order to themselves and their own needs.

It could reasonably be claimed that the more spontaneous, less contractual, planned and designed the social structures are, the more difficult they are to influence and change.

4.3 Self-produced goals

Lastly, it should not be overlooked that unhappiness and dissatisfaction are related to the agents' goals (those that Hayek somewhat restrictedly calls "the conscious purposes"), that is their desires, perceived needs, plans, aspirations,

¹² Because of the fundamental psychological phenomenon of the aversion to loss and of the reference level, people tend to prefer the *status quo* to changes that would involve the risk of losses, even when the latter would be more than compensated by new good (Knetsch and Sidnen, 1984; Knetsch, 1989); moreover, people adapt rather well to conditions that are neither good nor just and quite rapidly perceive as normal and fair conditions that they initially perceived as unfair (Thaler, 1985; Kahneman, Knetsch, and Thaler 1986).

etc. However these are not given a priori, in a way that is independent of the actual satisfaction (Maslow's need theory), or experience and success (motivation channeling theory), or regardless of the culture and knowledge (consumer theory; Sen's criticism of the utilitarian approach to well-being). It follows that my level of wealth, knowledge and consumption determines my "conscious purpose" as a function of which I will be more or less content or unhappy, and more or less responsive to the existing circumstances. If I know only poverty and death and am unaware of what else is possible, I will act within this structural and mental framework. Unfortunately if social order is the unwitting result of our activity directed towards our subjective goals, our subjective goals are partly an unintentional product of the social order. There is a co-evolutionary relationship.

That man produces himself and his needs and desires is not just a historicist and marxist thesis: it is also a typical premise adopted by Hayek himself.

“...the development of the human mind is part of the development of civilization; and the state of civilization at any time determines the scope and possibility of human ends and values” *La societa' libera*. Seam, ed., 1998. ¹³

Nevertheless, and here another disavowal seems to have been performed, Hayek uses this argument only to criticise "The idea of a man deliberately constructing his own culture" as deriving "from a false intellectualism" (ibidem); it does not take into account the tautology that this co-evolutionary view introduces in his own theory of spontaneous order and its subjective goodness.

All this in order to pass over other very important cultural and moral aspects of the problem in silence: the fact that people perceive the spontaneous social order with which they have become familiarized as their own "civilization", that they are "accustomed to accept its existence as a fact" (AR, VIII). They thus perceive their individual situation through the categories provided for them by the order itself; their emotional reactions are based on this (perception of justice-injustice; satisfaction-dissatisfaction; entitlement or privilege; right or non right; envy or emulation; etc.). What incredible "stability" does this impression of "fact" and "our civilization" give to the social order and what positive or ineluctable connotation does it add to it? How can we thus believe that individuals react so reactively and systematically and unconsciously adapt social order to their needs?

Hayek himself provides very cogent arguments to explain why the social order can be very stable *without necessarily being good/useful for individuals*. But he does not use them in this critical direction.

However, to be quite honest, I do not consider the arguments I have mentioned to be conclusive (order is not good because individuals are powerless; because the good it satisfies is defined by it itself) ¹⁴. Hayek actually adopts a highly *subjective* individualism: he might object that all this is true but that *with reference to* the conscious purposes of the agents (including the respect of tradition, of justice -in its historical conception-, their given limited needs and knowledge, etc.) the order is good and satisfies them. ¹⁵

Indeed, my objection -based on a slightly more sophisticated conception of action and the mind- is intended to be more central: give the agents' current subjective goals the order may be bad precisely with reference to them although being self-organizing and stable just the same.

What is important in the last two arguments is merely the fact that *to the extent to which the stability of social order is due to its being satisfactory, this order produces the needs and the mentality that it must satisfy in order to be stable!*

But as we shall see *order does not emerge and become stable only because it is satisfactory*.

The question remains, however, of how subjective individualism is able to give such little weight to these terrible factors conditioning individual choice (and thus its collective outcomes) in its optimistic view of individual action, which is ignorant but *free*.

4.4 The irritating but inevitable notion of 'objective interests'

¹³ However, Hayek is highly unilateral. In the first place, the 'scotomization' or disavowal of the projectual, designed nature of so much of individual and collective human action (we are thinking of the Constitutions and the movements preceding them; or of organizations such as corporations or unions) and of so many human cultural "products", is indeed too much. In the second place, when he says for example. "... we rely [because of our ignorance] on the independent and concurrent efforts of the majority to propitiate the birth of what we want when we see it" (extraordinary!) he is a moving apologist of the market and not an objective scientist. How can he pretend not to know how many needs and desires are deliberately created, *planned*, and imposed by the large multinationals (Coca-Cola, automobiles, oil, and so on), which are by no means individual ("the many") but the few and very powerful, super-organizations (oligopolies). Or how can he not know how many products "we desire when we see them" because they were made deliberately, on the basis of definite surveys of what we would prefer? What kind of market has Hayek in mind? that of a *Candide* or that with an exorbitant weight of marketing, advertising, multinationals? Or else are these institutions not a *spontaneous* product of the market itself?

¹⁴ However, in my opinion, the argument of unhappiness and desperation experienced as such but in impotence remains valid.

¹⁵ Hayek actually wants it to be good also for the future development of human needs and faculties (and also this seems to be a value judgment).

Before going on to our central thesis, let us briefly review another point on which Hayek ignores his own arguments and seems to be influenced more by prejudice than by logic. Hayek's theses concerning our "astonishing" ignorance (of the conditions relevant to our action), on the limits of reason and on the distribution of knowledge, are truly fundamental. However, they forcefully give rise to consequences that he does not take into consideration adequately. In particular, the fact that (as we have said) it is not true that each person is necessarily the best judge of their own good! that the agent can afford not to know what their own objective individual interest is. This leads to *the possibility of another individual knowing better than x what is good for x*¹⁶, and perhaps trying to get him to pursue it ('tutorial' relation - Conte and Castelfranchi, 1996 : problem typical of education, medicine, laws, government, etc.). This possibility is always problematic for a liberal view but an inevitable consequence of these premises.¹⁷

4.5 Path-dependency, conventions and badness of the invisible hand

Lastly, it is necessary to consider the -now classical- objection concerning the intrinsically path-dependent, hard to reverse and binding nature of each "convention" as defined - in terms that are actually spontaneist and emergentist- in game-theory literature. It is by definition an equilibrium -often functional, useful for coordination but largely arbitrary- which no individual has any interest to deviate from -even if he does not agree with it- because he would be more harmed by it. The classical example is that of driving on the right or the left: even without laws and the police anyone going counter-current would be eliminated; but there are also much richer, more dramatic and more convincing arguments, for instance, infibulation in African cultures (Mackie, 1996). Being largely path-dependent, these conventions may easily be sub-optimal with respect to alternative viable concrete alternatives which have not however been followed. Even if sub-optimal or in any case worse than other non ideal possibilities, and even if considered worse by the individuals subjected to them who have glimpsed the alternative, these structures are nevertheless stable (Davis, 1985; Arthur, 1994). Indeed *the transition to the alternative would demand a coordinated and almost simultaneous change, which cannot take place without communication, agreement, shared projects*. The larger the number of individuals involved the more unlikely it is that there will be a spontaneous change for the better or an agreement and a common plan.

However, although bad or worse than what else is possible, these conventions are also 'useful' (at least at the time they are formed), catering for some needs of the individuals that contributed to establishing them, and thus more useful than detrimental (and hard to evaluate!). Therefore this objection - although important for depriving the invisible hand of its connotation of goodness or 'best possible' - is not my objection, and I shall not dwell further on it.

As I have said, the point I am making is more drastic and radical: a) *directing an action towards good does not mean that its effect will be good*; b) *equilibria may be set up not despite but thanks to their perverse effects*, and probably even simply harmful equilibria. In the case of the "conventions" the only certain element of this kind is the fact that the convention -although worse than a feasible alternative - is *retained also by its perverse effect of irreversibility*: indeed, it is precisely the fact that it is preferable not to deviate from it individually - which is the result of its having become a "convention" - which perpetuates it.

5 . Do actions directed towards good imply the goodness of the emerging order?

In general, but only in general, the agents' behaviour is related to their subjective good. I will accept this premise, even though it is too general, because my argument is intended to be more radical:

- *from the fact that individual behaviour is directed towards good (individual subjective goals) it does not follow that the emerging order is good (for individual subjective goals).*

¹⁶ And this postulating only x's current goals and preferences and his potential, future or normative goals.

¹⁷ Let us take into consideration also another line of reasoning that Hayek seems to reject. It is precisely the individual and subjective limits of knowledge can lead to cognitive and practical cooperation; to the attempt to construct *collective knowledge and goals* (groups, organizations). Since on the one hand individuals *follow* (anticipatory) *goals* -do not discover them only after the action; and also these goals are "ignorant", and other people's knowledge and points of view (goals) can improve them. On the other hand, a group may know more than the sum of the knowledge of its members. And its knowledge can also be better grounded and justified.

More exactly: even accepting the postulate that *individuals tend towards their own happiness or well-being*¹⁸,

- a) it does not follow that by so doing they produce their own good, either individual or collective, or objective or subjective. Subjective good and objective good do not coincide: there are limits to knowledge (ignorance, illusions, superstition, etc.) and to rationality (e.g. shortsighted preferences, framing, etc.) and thus in the goals adopted. As a result of which subjective good can be an evil or a lesser good with respect to an observer's point of view. Also subjective good is not guaranteed -given the limits and the cognitive distortions - since it is not necessarily true at all that the decision and the action, although directed precisely towards one's good, are adequate or rational (see below). Concerning the fact that good intentions do not necessarily imply good consequences, Hayek could obviously not agree more (it is his thesis on the limits of reason, of knowledge and of projects); his optimism resides in the fact of believing that the good solutions are selected and retained.
 - b) it does not follow that the emerging equilibrium and spontaneous order are good. Not only can "collective" results be bad - as in the classical prisoner's dilemma (why would spontaneous order be precisely a huge PD with many players? What would guarantee the public good (Smith) or the human purposes? Equilibrium may be bad also for the individual subjective goals, as not only can the good effects (for the individual and individuals) be self-organizing; also the negative effects (for the individual and individuals) can be self-sustaining and self-reproducing. It is not simply a matter -as is obvious- of producing also 'disorder' (which is combated by vital activity and action), or there being undesirable and harmful effects; there emerges precisely an *order* of harmful effects.
- *Self-organization is partly independent and indifferent to the interests and goals of the agents*

The order emerging from an individual behaviour directed towards good may be perverse (in the narrow sense, not in that of Boudon, cf. sect. 8.). Clearly additional unexpected evil effects exist, or evil effects combined with good individual intentions (Boudon, 1977) in which the intended good effects reproduced *in spite of* the negative consequences. This is true,

- both in the case in which the evil effects are not perceived or are not attributed correctly,
- and in the case in which they are perceived (Figure 1)

(in the second case the good effects must be subjectively more important and in any case preferred (for instance, be closer in time), or else are more conditioning/reinforcing than the evil effects)

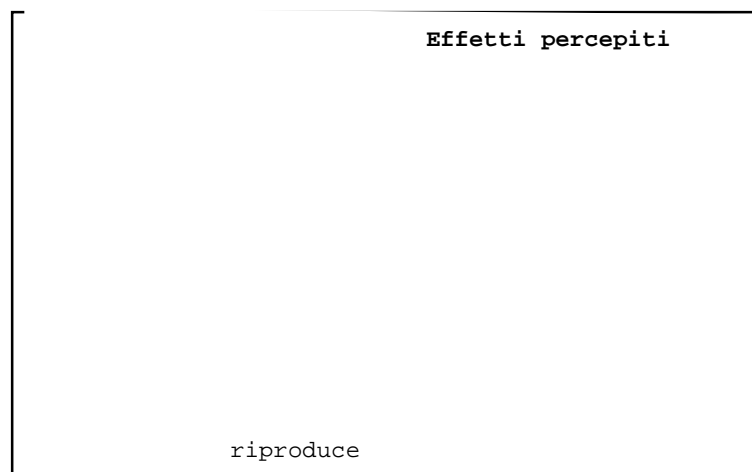


Fig. 1

Let us consider for example the negative effects consisting of a long line of automobiles and the slowing down due to the simple individual intention of rapidly glancing at an accident that has occurred in the other lane. Each

¹⁸ This premise is anything but obvious; one might disagree with it. Take for example the position of Leopardi set out in the introduction. Notice that I -although referring to Leopardi's conception of natural and regular evil (cf. par.7) - do not agree with his later philosophic view of evil as and end of nature, which emerges for example in the following text:

"All is evil. That is, everything that is, is evil; that each thing should exist is evil; each thing exists for the purpose of evil; existence is an evil and is ordered for evil; the purpose of the universe is evil; order and the state, the laws, the natural progress of the universe are nothing but evil, nor are they directed towards anything but evil" (*Zibaldone*, 4174; April 19 1826).

driver slows down only for a second and gives only one glance at the accident (without stopping), and yet because of the heavy traffic this causes a long line-up and a significant delay. This behaviour is repeated for several reasons: on the one hand, the majority of motorists do not realize that the significant overall delay is the result of a very short individual slowdown and, since the action achieves the intended positive effect (satisfy one's curiosity), it will be reproduced; on the other hand, even drivers who understand the collective perverse effect their individual actions may have are not capable of resisting the temptation and slow down, or else in any case have a quick look as in case they are obliged to slow down by the already emerging structure of the jam (they simply cannot do anything else).

But there are also harmful effects capable of self-reproduction (through the action) precisely *because of their negative nature* (Castelfranchi, 1997; 1998b; 1998d).

riproduce

riproduce

Fig. 2

A more detailed examination of this topic (sect. 7) calls for an excursus on the notion and theory of 'functions'.

6. From 'patterns' to 'functions': beyond the observer and the subjective goals

What kind of notion of "spontaneous order" do we need? Hayek's spontaneous order is ambiguous, and both its aspects are unsatisfactory. On the one hand, the order appears to be a merely observational type of emergence (as is currently customary to say about emergence and complexity - "in the eye of the beholder" - Forrest, 1990; Virasoro, 1998); on the other, it takes on a purposive-functional character, but only as far as the individuals' subjective goals are concerned. Nor is it clear how the order can be *self-reproducing* and if this gives it any autonomy from the observer and self-referentiality; nor if any finalisms that are independent of the subjective goals (and thus not intrinsically 'good' for someone) may be envisaged.

Hayek very rightly states that:

"What was important to achieve was a division into three categories that are included among

- natural phenomena, that is, *independent of human action*, and
- artificial and conventional phenomena, that is, *the result of human design*,
- a separate intermediate category including those *unintentional patterns and constants* found in human society and that represent the explanatory task of social theory.

(F. v. Hayek, *Studies in Philosophy, Politics and Economics*) (cited in Boudon, 197), p. 178).¹⁹

¹⁹ By means of this important distinction Hayek endeavours also to deflect Marx' irony directed at economists: "Economists have a singular method of procedure. There are only two kinds of institutions for them, artificial and natural. The institutions of feudalism are artificial institutions, those of the bourgeoisie are natural institutions Economists have a singular way of proceeding. For them only two types of institution exist, those of art and those of nature.... When the economists say that present-day relations -- the relations of bourgeois production -- are natural, they imply that *these are the relations in which wealth is created and productive forces developed in conformity with the laws of nature.*" (Marx, 1973, p. 182).

In my opinion, however, Hayek's 'spontaneous' category ends by playing a very similar (non rationally improvable) role and has very similar positive connotations; let us take the comparison between social order and a crystal: we know how a crystal is

Hayek nevertheless is not aware of the weakness and theoretical inadequacy of the notions of 'pattern' and 'constant', nor does he tackle the theoretical problem of the notion of 'function' and the important distinctions within the aggregated and constant unintentional effects of human action.

For instance, is the notion of pattern or of structure or emerging configuration (as we have said in sect. 4) referred only to an observer and thus only to his interests, discernment, values?, or else has it some degree of 'objectivity'? Should it not for example be possible to distinguish those configurations, the organization of which lies not only in the observer's eye (as for example in the case of the constellations in the sky), but has specific effects, is not just an epiphenomenon but *plays a causal role in the process*? For example the school or flight configuration taken on by many fish and birds (not intended by any individual and the result of simple local reactive rules) seems to perform specific functions, that is, to cause specific effects of deterrence or confusion among predators and in any case to reduce individual exposure to predation (or to other effects). And what may be said of those patterns that remain constant, that is, that are reproduced *precisely as a result of the causal effects they exert as patterns* (but without being understood and thus become intentional)? Do they not form a very important category that deserves to be distinguished from patterns or constants that are repeated but only as side effects of other phenomena that are repeated for reasons of their own? There are thus configurations that play a causal role and epiphenomenal configurations; included among the former are those that are actually *self-sustaining or self-reproducing (constant) by virtue of their own effects* based on a causal loop, a feedback -those that I call 'functions'- (while it is possible to be constant and repeated **without** using one's own effects).²⁰

These distinctions are no less essential than those proposed by Hayek, and indeed are useful in clarifying the latter; and it is in these distinctions -with the help of cognitive theory- that social science must formulate theories.

Even Boudon -the great theoretician of unintentional effects - does not make any such distinctions. He converges with Hayek, it appears, also in subjectivist individualism, that is, in reducing any possible functionalism and finalism to finalism in the psychological sense (cf. sect. 3).

On the one hand I am seeking a strong notion of "emergence" which is not merely observational but also a) plays a causal role in the phenomenon, b) is self-sustaining and self-perpetuating thanks to the effects it produces. On the other, on this basis, I invoke two types of finalism: the first of a psychological/mental (and possibly subjective) nature, while the second is literally an "*end in itself*", "functional" only to itself and its own reproduction, not to someone's good, and thus good or bad indifferently for the agents or their systems.

6.1 Order versus happiness: relative goodness and goodness in itself

Let us now examine the same problem from a different point of view.

There are two different and independent criteria with reference to which something may be deemed to be good or functional. They correspond to two types of 'telos', to two distinct foundations of finalism: on the one hand internal goals (both explicit and implicit), on the other, merely external goals or functions (Castelfranchi, 1982). These two different criteria and notions are often confused.

Let us call the first criterion "relative". It presupposes a goal-directed (and if possible, goal-regulated) entity X, that is, an agent possessing its own goals. What favours one of X's goals is "good for X"; what damages or threatens one of its goals is "bad for X".²¹

Note that what we have said does not mean that X knows or understands what is good/better for him. Good and bad apply even to completely stupid agents. It may be another agent (for example, the observer) who realizes that p is good for X (see sect. 4.4.). Therefore what is "good for X" is not necessarily "good according to X", and vice versa. There is an *objective* relative good and a *subjective* relative good. (Miceli and Castelfranchi, 1992). However, all goal-oriented systems have some form of appraisal or learning, or of cognitive evaluation in order to distinguish between what is good and what is bad for them, and to endeavour to procure one and avoid the other. They worsen their state of well-being and power when they encounter or produce something bad; and if they have a subjective life they suffer when then believe/feel that what they have is bad for them.

If their goals are suited to the world, they prosper and proliferate when they attain their subjective good; and -if they so believe - they are happy.

The second criterion is instead self-referential or "absolute". In this case "good" or "functional" simply means capable of self-reproduction, self-sustaining through a reproductive cycle and by virtue of a positive feedback within this cycle. There is not necessarily any reference to someone's needs, goals, well-being or happiness.

formed, but we cannot make it artificially, we can only create the conditions in which it will naturally/spontaneously form (see Gloria-Palermo, in this book).

²⁰ Given this notion (see also later) it is clear that *not all the stable emerging effects -good or bad - are functions*.

²¹ Since the other agents have more than one goal, some of which are active at the time while others are inactive, some in the short and others in the long term, some in conflict with others, etc., the (procedural or declaratory) evaluation of what is good or better is anything but simple and univocal.

These two notions are well illustrated (but also immediately confused) in a very interesting text by Leopardi (whose philosophy I intend here to oppose to Hayek's).

"Of course many things in nature are good, that is, they proceed in such a way that such things may be preserved and be lasting, which would otherwise not be the case. But an infinite number (and perhaps even more than the others) are not good, or are badly combined, both morally and physically, causing beings great inconvenience. ...Even though they do not destroy the present order of things, they are naturally and regularly bad, and are natural and regular evils. But from these we do not draw the argument that the workshop of the universe is the result of an unintelligent cause; rather from those things that function properly we believe we can consistently argue that the universe is the product of an intelligence " (*Zibaldone*, 4248 - 18 Feb. 1823).

In a masterly fashion Leopardi provides us with the second (absolute) criterion: "goodness" as self-organizing, self-reproducing and self-sustaining ("good, that is .. in such a way that such things may be preserved and be lasting, which would otherwise not be the case"). On the other hand, when speaking of evil, he introduces the criterion of *relative* and subjective ("causing beings great inconvenience"). He does not realize that the definition of "good" actually also implies and mingles with a relative criterion (the subjective well-being of beings), since he opposes a subjective evil to them. Nor does he realize that, with the criterion of good he proposes, evil itself turns out to be a good (in the absolute sense); it is functional, in so far as it preserves order (" ... they do not destroy the present order of things, they are naturally and regularly bad, and are natural and regular evils ").²²

In actual fact, the two criteria are independent and orthogonal to each other: there are functions versus entropic or non self-reproducing effects; and the functions may be good or bad²³ with reference to the goals of a given organism or goal-oriented system; and also the irregular effects may be good or bad for individuals or for their social systems. The tendency to confuse the two notions and to perceive the self-referential functions, orders or equilibria as good or providential for individuals or their society is very strong and widespread in both scientific and everyday thinking.

6.2 Intentional behaviour Vs functional behaviour

We finally come to illustrate the two finalisms in question, and then the notion of "function" in the general accepted meaning of the term (self-referent).

Many natural and social behaviours are purposive. However, they cannot be defined as being guided by "goals" in the narrow sense: for example, we do not want to attribute internally represented goals - of the intentional type - to all types of animals; nor consider the functional effects of social action (e.g. spontaneous division of labour) as necessarily deliberate; nor, lastly, to attribute a mind to society as a whole. Is there an available concept that can account for the teleonomic nature of behaviour without postulating goals within the system displaying it?

Finalistic systems: Goal-oriented Vs Goal-governed

There are two basic types of system having a finalistic (teleonomic) behaviour:

- *Goal-oriented systems* - Mc Farland, 1983), i.e. systems whose behaviour is finalistic, that is aimed at achieving a given result, which may not however be understood and expected, or explicitly represented inside the system: it is not a representation that anticipates the result which controls the behaviour itself from the inside.

A subcase typical of these systems is represented by *merely goal-oriented* systems based on rules (production rules, classifiers) or reflections, or releasers, or associations: they react to certain circumstances by deploying a given suitable behaviour (thanks to selection, learning or design).

- *Goal-governed systems* are instead a specific type of Goal-oriented system based on representations that anticipate the results. More exactly, a system or behaviour is defined as goal-governed when it is controlled and regulated in a "purposive" way by an internally represented goal, that is, by a "set-point" or "goal-state" (cf. Rosenblueth and Wiener, 1968; Rosenblueth et al., 1968)). The prototypical example is that of a heater-thermostat system. A goal-governed system responds to the external goals based on it (functions of its behaviour) thanks to its internal goals (Castelfranchi, 1982; 1998a).

It is essential to stress that the incompatibility between Merely Goal-oriented and Goal-Governed is not complete: a system may be Goal-governed for certain purposes (internally represented goals) and merely Goal-oriented for

²² This is what Leopardi realized later, when he actually stated that the evil suffered by sensitive creatures is necessary for the universe and that the universe is directed towards, ordered for, evil (cf. note 17).

²³ Merton (1949) already makes the distinction between *eufunctions* or *dys-functions*, albeit only with respect to the social system.

others. That is, also intentional systems can have functions that they do not pursue deliberately; either as effects of reactive or low level behaviours, or *as unintentional functions of their own intentional behaviours*. Intentionality does not (completely) replace Goal-oriented behaviours (unlike Elster's thesis that intentionality is incompatible and makes the notion of function superfluous) (see below).

6.3 The notion of 'function' as an effect selecting and reproducing its own cause

Clarifying the notion of function implies clarifying the relationship between functions and cognition. In particular, it is necessary to clarify the relation between the functions of the action and the beliefs and intentions guiding it. My thesis is that social functions are installed and maintained in a parasitic fashion vis-à-vis cognition:

- *the functions are installed and maintained thanks to and by means of the mental representations of the agents but not as mental representations: that is without being understood let alone intended.*

While, for example, in order for a social rule actually to function as a social rule and be fully effective, it is necessary for the agents to adopt it as a rule (Conte and Castelfranchi, 1996); conversely, the effectiveness of a social function is independent of the fact of whether the agents actually understand the function of their behaviour. Indeed:

- a) the function may be installed and maintained without the agents being aware of this;
- b) if the agents intended their own behaviour to produce these results, they would no longer be a "social function" of it but merely an intention (Elster, 1982).

I accept Elster's crucial objection to the classical functionalist notions of the social sciences²⁴, although I believe that functional and intentional behaviour can be reconciled. By means of an evolutionistic view of functions, it may be claimed that *intentional actions can lead to unexpected functional effects*. I would like to interpret and extend Elster's view as follows.

- As we cannot consider the emerging nature of the functions simply as *what the observer notices or appreciates*, but that it ought rather to be independent of the observer, and based on self-organizing and self-reproducing phenomena, the so-called "positivity" of the functions may simply reside here. That is, they are self-referential. Therefore we cannot exclude phenomena, effects, that could be bad, that is, negative from the observer's point of view, or of the agents involved, or of the social super-system²⁵. That is to say, we cannot exclude "negative functions" (kako-functions) from the system²⁶. It is possible that the same mechanism is responsible for both positive/useful functions and negative functions.
- How is it possible for a system that acts intentionally on the basis of an evaluation of the effects vis-à-vis its own goals, to reproduce bad habits precisely *as a result of* their bad effects? And even more crucially - if a behaviour is instead reproduced thanks to its good effects with respect to the (individual or collective) goals of the agent who reproduces them by acting intentionally, then there is no room for the "functions". If the agent appreciates the effects of his action and this is repeated for the purpose of reproducing these effects, they are simply "intentional", not functional.. *The notion of intention is sufficient to make that of function superfluous* (Elster, 1982).

²⁴ John Elster very clearly stated the terms of the problem, which is of fundamental importance for the functionalist theories of the social sciences: "...for a functional explanation to be valid it's indeed necessary that a detailed analysis of the feedback mechanism is provided; in the huge majority of the cases this will imply the existence of some filtering mechanism thanks to whom 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, the "unctional" explanation is in an unfortunate *dilemma*... either it is not a valid form of scientific explanation (it's arbitrary, vague, or tautological), or it is valid, but is not a specifically functional explanation" (Elster, 1982, p. 480).

²⁵ Positive and Negative for whom? There are several possibilities:

for the agent himself	+ + + - - - +
for another agent	+ + - - - + + -
for the System	+ - - - + + - +

There are obviously some *ambivalent* cases, that is, positive for one goal and negative for the others.

²⁶ Hence Merton's (1949) "dys-functions" (which are negative for the system) are only a sub-case. There are "micro" functions of an individual's behaviour related to the reproduction of the same behaviour (which may be negative for him) and "macro" functions of the individual's behaviour vis-à-vis the reproduction of the system in which it is contained and thus indirectly reproduce the behaviour (which may be negative for the system: dys-functions). I use the term kakofunctions to embrace them all: self-feeding bad effects at the micro and macro levels.

How can functions and spontaneous, unconscious cooperation (e.g. the division of labour) be constructed on intentional actions and individual decisions? How can results that are positive insofar as they are advantageous strengthen and reproduce an intentional behaviour without themselves becoming intentions of those actions? This is the real theoretical challenge to reconcile "emergence" and "cognition" (Castelfranchi, 1997; 1998b), intentional behaviour and social functions, planning agents and unconscious cooperation.

I think it is necessary to have complex reinforcement learning forms not merely based on classifiers, rules, associations, motor sequences, etc. but *operating on the cognitive representations governing the action, that is, on beliefs and goals*.

In this view "the consequences of the action, which may be more or less consciously anticipated, nevertheless modify the probability of the action being repeated the next time in similar stimulus conditions " (Macy, 1998). More exactly:

- *the functions are simply effects of behaviour which go beyond the intended effects but which can successfully be reproduced because they reinforce the agent's beliefs and goals that give rise to this behaviour.*

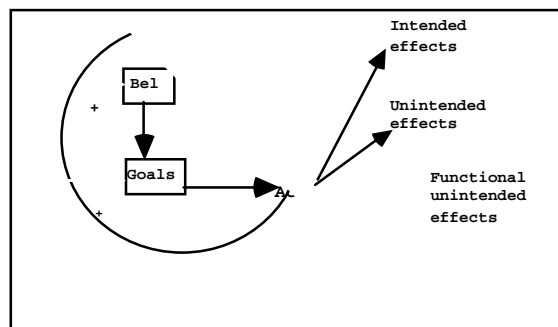


Fig. 3

Therefore:

- First, behaviour is 'goal-governed' and based on reasons; that is, it is an intentional *action*. The agent bases his goals, preferences and decisions on his beliefs (this is our notion of "cognitive" agent).
- Second, several effects of these actions are unknown to or at least unmotivating for the agent.
- Third, there is a circular cause (feedback loop) stemming from these unintentional effects which increases and reinforces the beliefs and goals giving rise to this action.
- Fourth, this "reinforcement" increases the likelihood that in similar circumstances (which activate the same beliefs and the same goals) the agent will produce that same behaviour, thus "reproducing" the same effects.
- Fifth, at this stage these effects are no longer simply accidental or negligible: although still remaining unintentional they are produced teleonomically: *that behaviour exists (also) by virtue of its unintentional effects, and is functionally related to them*. Even if these effects were negative for the goals or interests of (some of) the agents involved, their behaviour would be "oriented towards" them.

Note that the agent does not necessarily understand or suspect that, through his actions, he is reinforcing his own beliefs and goals, and therefore his behaviour.²⁷

Let us now examine the same phenomenon from another point of view, which will allow us to illustrate another similar mechanism. Even without postulating any reinforcement or any learning by the agent, seeing that each action is actually an "interaction" with an environment, and seeing that motives, goals and actions are elicited by environmental stimuli and adapted and selected by the actual conditions, *an effect that causally maintains or reproduces those conditions which lead to the action, will maintain or increase the likelihood of it recurring*.

²⁷ There is another plausible mechanism to account for the "reinforcement" of a behaviour guided by (intentional) expectations: *emotional reactions conditioned by anticipatory representations* (as possible goals and choices). I believe that both these mechanisms exist and are complementary to each other. There is something in our body/brain that unconsciously and automatically directs our decisions and actions, attracts or repels; it enables us to make an intuitive, unreasoned, appraisal of possible scenarios (Miceli and Castelfranchi, in press). These unreasoned preferences and choices can be accounted for by means of the so-called "somatic markers" (Damasio, 1994) or in terms of "evaluative conditioning" (Martin and Levey, 1978) which lead to an immediate pruning of the alternatives that, the proactive and anticipatory animals that we are, we set up. They are the product of previous emotional experiences and are associated (through learning) with similar future scenarios (not simply-it should be noted- with reactions and behaviours). In this way there is a kind of pre-decision, pre-selection of possible goals, based on affective responses and learning (positive or negative reinforcement).

It is not necessary for there to be learning and reinforcement mechanisms for goals and actions in order to systematically reproduce behaviour by virtue of its effects. It may be sufficient to have an effect on the context of the behaviour or choice such that the mechanism of choice would produce the same behaviour. This is possible also in a rational decision: the non understood and unplanned effects of our actions create external conditions and incentives - which we perceive - such as to induce us to repropose the same behaviour. (Figure 4)

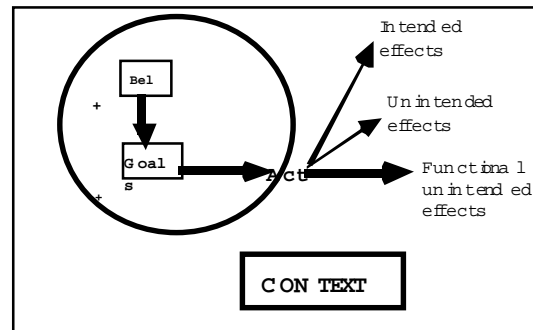


Fig. 4

In the spontaneous division of labour, for instance, agents do not understand the origin of the local conditions and incentives that induce them to become specialized, and which are they effects of their own actions.

Therefore this is the abstract and non evaluative notion of function that I propose:

every regular -non accidental and non occasional- effect that is self-reproductive thanks to the fact that it repropose or repeats its action/cause, is a "function" of this action, irrespective of whether it is good or bad with respect to the purpose of the action or the goals of the agent.

Indeed, each action reproduces itself thanks to its own effects (through reinforcement, or understanding, or the conditions, etc.). In *cognitive agents* the action is ideally reproduced thanks to effects that are *understood* (correctly perceived and attributed) and thus *expected* and *intended*. However, -and this is what was neglected by Elster - there may be effects that although not understood (let alone intended and chosen) that can reproduce (by reproducing themselves) the action/cause. The general principal is that:

- These effects reproduce and represent that action by reproducing or recreating some of the *internal or external conditions that activate, motivate, or select* that action and -in general- increase the probability of it occurring²⁸.

7. 'Natural and regular evil' (Leopardi versus Hayek): the stability of negative effects and evil-oriented functions

It is of decisive importance to realize that *negative effects can stabilize and self-reproduce like positive effects*. The only differences are that:

- a) Positive and negative effects are not equiprobable, since behaviour, intelligence and learning are oriented towards the production and preservation of the positive ones.
- b) It is slightly more likely that the negative effects will be discovered and specifically resisted (although not necessarily successfully, because of the "vicious circle" nature of the mechanism by which they are reproduced: suffice it to think of prejudice and marginalization, or the vicious circle among prisons that reproduce criminals that reproduce prisons, etc.).
- c) When discovered, positive effects become expected and often intended (Elster, 1982), thus reducing the "spontaneity" of the social order and the unintended nature of the institutions.

Nevertheless these differences do not eliminate the fact that:

the negative effects can stabilize and self-reproduce exactly like the positive effects.

²⁸ Clearly, at a more subtle level of analysis, we are actually still dealing with beliefs: the beliefs the agent holds concerning the context, the circumstances, the opportunities and the convenience. The preconditions for the intentions and the action are in any case beliefs (in cognitive agents); therefore also the reproduction of the action through the maintenance or modification of its context, is actually a reproduction of the action through the maintenance and modification of specific beliefs.

The necessary requisite for such self-reproduction is the installation of a loop (vicious circle): several of the effects have a feedback effect on their own causes (or rather on the causes of the action) contributing to or being sufficient for the reproduction of such behaviour. We see several examples at the level of individual²⁹, social and institutional behaviour. It is important that our view should reiterate that the problem of the perverse effects *is not simply a problem of composition and complexity*; it involves the theory of action in general: even individual action. Furthermore, one must not overlook the distinction between effects that become 'ends', 'functions' and those that are merely accidental or iterative (see sect.8).

7.1 Individual level

The wrong treatment and the persistence of pain

This is a trivial case in which the subject persists in the harmful behaviour also in view of the harm it does because he does not revise his causal beliefs (and does not know or seek any better alternatives). Let us assume I suffer from a physical ailment (for instance, heartburn or backache) and when I take a remedy I get temporary relief and a feeling of improvement but in actual fact I aggravate the situation and the ailment returns very quickly (e.g. bicarbonate or an incorrect gym exercise). I do not blame the 'remedy' for the return of the pain and thus persist in my detrimental behaviour. Note that any momentary relief is only the *sign* that I use to confirm to myself the efficacy (however limited) of my action. In actual fact, this sign is not even necessary: if I have complete confidence in a doctor I may persist - for some considerable time - in the useless or harmful behaviour even without any sign of its efficacy. Superstitious reasoning is a very similar example (see below).

It is obvious and definitory that each action should be reproposed or persist thanks (also) to a subjectively expected benefit. This is not the problem. The problem lies in the fact that the intentional nature of the action does not exclude or eliminate the fact that it may be functional and goal-oriented. That is, it does not exclude the possibility of it being reproduced by yet unknown or hitherto incorrectly attributed, even harmful, effects: the action is simultaneously intended (goal-directed) towards p, and functional (goal-oriented) towards q. And the two things do not take place in parallel; indeed the latter is parasitic on the former: the function is guaranteed thanks to the reinforcement and reiteration of the intention.

Superstition etc.: an anomalous reinforcement

Let us now consider behaviour based on false beliefs -which are very widespread in society- such as spells, witchcraft, horoscopes, etc. When the behaviour is "successful" (that is, accidental positive results may be attributed to it) it is reasonably confirmed, reinforced and, if necessary, intentionally repeated. But what is more important is that even when the practice or prediction fails, the subject may persist in the belief and in his behaviour simply as a result of erroneous explanations or attributions ("there is a certain interference from negative forces; either I did not understand something or didn't do it properly"; etc.). Note that it is precisely owing to its failure that the behaviour may persist and be repeated: in this case failure reinforces or maintains the goal, just as it does not weaken the beliefs and even reinforces them ("this proves that I am really under a negative influence!"; "if I have not won so far, the chances of my winning have now increased").

The anxious mother

Let us consider a mother with a somewhat nervous or sickly child, who absolutely wants to avoid 'spoiling' her child and yet is upset and worried by the fact that the child repeatedly wakes up in the middle of the night and cries until his parents take him into their bed. After a while, several unsuccessful attempts and much crying, she can no longer put up with the child's crying or to resist his tantrums, and gives in. This way *without intending to, she is reinforcing* the undesirable behaviour of her child, and also the forthcoming intention of the child to misbehave. It is a vicious circle - of course, obviously 'vicious' only as far as the desires and values of the mother are concerned; her intentional behaviour (to quieten/calm down the child) has taken on a kako-function which maintains and reiterates it over time³⁰. And this is true whether the mother is aware of the phenomenon or whether she is acculturated and

²⁹ This is also the case of the "neurotic paradox" (cf. Mancini, 1998) which, in a sense, I am proposing should be transferred from the individual level also to the interpersonal, group and institutional level (Castelfranchi, 1998b; 1998d). In neurotic behaviour I deny that there *must* be a "second advantage" that can account for its undesired repetition. The subject does not "choose" in a utilitarian way (he prefers, is attracted by, etc.) the secondary advantage or benefit, but chooses -as in any behaviour - to feel good, but to do so he may choose an inappropriate, counterproductive behaviour, and one that is in any compulsory (without any alternative). It is reinforced by the failure not by the secondary advantage. As Mancini points out, the real sub-problem to solve is why, despite the failure, one does not give up but persists, that is, the lack of or inadequate learning (which needs a theory of its own, in view of the numerous answers and cases, in the first case by distinguishing between cases in which failure is not perceived as such and those in which the subject is on the contrary aware of the failure).

³⁰ From the child's point of view this is a eufunction or an intention. It depends on the complexity of the mind that is attributed to the baby. I personally believe that it is not true intentional behaviour but operant conditioning.

understands that she is involuntarily reinforcing it (obviously in that case she must prefer the desired short term effects over the undesired ones - harm - in the longer term). On the one hand, the negative effect - although anticipated - is not an "intention" of the mother, it does not motivate her behaviour; on the other, *the behaviour exists and is reproduced also thanks to the unintentional and negative effects* therefore for them: the behaviour is goal-oriented but not goal-directed towards them.

The dysfunctional action (and also the neurotic behaviour) certainly has an objective, a benefit in view (provided it is not impulsive in nature) although it is not true that this benefit must necessarily be subjectively superior and preferred to the harm - not accidental and contingent - deriving to the subject from the behaviour adopted. This is true only when the harm is adequately appreciated, expected with reasonable certainty, attributed to the action and with all possible precautions: in this case the subject actually prefers the objective proposed (plus any expected advantages) to the predicted harm.

The general abstract model (in the sense that it corresponds to a range of micro-mechanisms) would be as follows:

an intentional action is repeated with a view to its goals although it is actually reproduced either by the (complete/partial) failure or by negative effects (perceived or not perceived; understood or not understood) that repropose the problem and reinforce the beliefs and goals on which the action is based.

7.2 Interpersonal and collective level

Hubbub in a restaurant or at a party

We are at a meeting, party, gathering in a restaurant of a large group of people having fun: more than 100 persons in a very large room, at different tables, drinking and eating. The persons present are involved in lively discussions or joking. The room becomes very noisy, the hubbub is very loud and you cannot hear what the person opposite is saying. In order to make yourself heard you have to shout. No one realizes that if everyone spoke softly it would be possible to hear at least as well and with less effort and annoyance. No one wants to make this din (collective effect) and considers their own contribution as minimal with respect to the whole, although everyone is producing it with the purely 'local' intention of speaking above the din and of being heard by their neighbour.

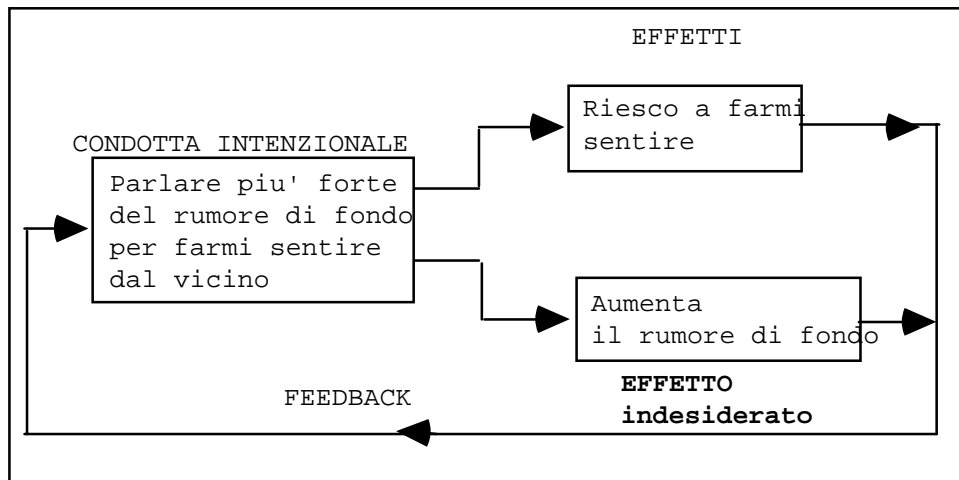


Fig. 5

Note how the din is reproduced thanks to its own effects and their impact on individual behaviour, and thus thanks to/through the latter. Note also that even if anyone realizes the absurdity of the matter, it would be to no avail to deviate from the others, as one would only lose out personally (one would not be heard). What has become established practice (shouting) is a kind of "convention". The only way out would be to stand up and, screaming at the top of one's voice or using a microphone, to ask everyone present to simultaneously speak more softly; or else the introduction of shared interiorized rules of correct behaviour; or lastly a social rule enforced by someone. (The example given is merely -on a small scale- the model followed by the arms race).

Street litter

A sees (*belief* - consider the model in Fig. 3) a lot of litter in the street, and thus thinks (*belief*) that the rule of not throwing paper and such like in the street is not very important in that community or else is a rule that is not respected. This reduces the strength of his goal to respect the rules (Conte and Castelfranchi, 1996) and not to throw litter in the street and so he too ends up by littering the street; moreover he considers (*belief*) that his contribution to the litter is only marginal. The problem is that by so doing he (both directly and indirectly: thanks to the bad example

and the spreading of the infringement) reproduces and amplifies the collective effect which, it should be noted, he may not desire and may consider negative. However, this collective effect confirms and reinforces his own preferences and thus his goal, and reproduces his behaviour like that of the others. It is not that he chooses and prefers the advantage of not having to make the effort to find a trash can to the feeling of annoyance at the litter in the street. It is rather that he realizes the circular nature of the phenomenon and his contribution, or else he considers the goal to have already been compromised ("in any case everyone does it, it's all dirty"). The behaviour is reproduced (also) by its own effects, which no one wants (it is obviously reproduced by the intention to make less of an effort).

Hostility leads to hostility

Let us consider A who has to interact with B, and towards whom he has feelings of fear and bias; for example, he is afraid the latter will display a hostile and threatening, and perhaps even aggressive, attitude. Simply in a partly reactive-emotional mode (out of fear-diffidence) and partly intentionally, in order to be circumspect and on one's guard, he adopts a closed, suspicious, hostile attitude. This attitude of his arouses in B diffidence and hostility, triggering in him a mirror image of his own behaviour. A *does not realize* that it is his own behaviour that produces B's hostility; he will not attribute this effect to his action. Indeed, he will interpret B's behaviour as a *confirmation* that his predictions and his beliefs concerning B were correct. The confirmation of these beliefs (now based also on evidence and experience) will repropose and strengthen the goal of being hostile and on one's guard. Of course A subjectively attains his goal (he believes, for example, he has avoided harm and that he has been prudent; perhaps he is wrong) and it is in view of this goal - for which his behaviour seems effective - that he repropose it, but this goal, just like the beliefs supporting and reactivating it, are reinforced by the effects of the action themselves (negative effects related to A's goals and fears: he creates his own phantoms). He may even create a real aggressor through his expectations ('Pygmalion effect'; Rosenthal and Jacobson, 1972).

7.3 Institutional level

Prisons and delinquency

Prisons, police and court-rooms reproduce themselves, reproducing delinquency; this is Marx's paradoxical and ironic observation (and it could be applied also to mental hospitals, medicine and various other social institutions. Naturally this is not the explicit institutional role of prisons, court-rooms and police: it is not the intention of the legislator or even of policemen, magistrates, etc. For the sake of simplicity let us assume that everyone has the best of intentions. And yet the institution produces this perverse effect (to train criminals and make them habitual offenders who 'stick together') against its intentions. In other words, it largely fails to achieve its intentions. And it is precisely this failure (the failure to achieve rehabilitation or the lack of deterrent power) to partly recreate the conditions that again demand new police, court-rooms, prisons. Also in this case *it may be precisely the failure of an action that causes it to be repropose cyclically.*

Of course, if it repropose the action it is in view of an advantage: its objective. Existing delinquency (which is partly reproduced) reproduces the *goal* of putting it down. It is true that the action partly attains its goals - and also for this reason it is repropose - and it is true that if the harm is known and understood (attributed to the action itself) the intended benefits must be greater than the expected harm. But in actual fact, there is no need to attribute to the System (as the Red Brigades did) any intention or any underhand and perverse intelligence, that is, to render the intention and calculation a mere self-reproducing (kako) function. Furthermore, also at this level not even a partial success of the action seems necessary: it is enough for the agent not to revise his beliefs concerning the effectiveness of the action and the absence of any better alternatives. It has been demonstrated for example that the death penalty is not a deterrent and in no way reduces serious crime, and yet its champions continue to repropose it in the belief that it is.

In other words, either the agent does not realize the (even partial) failure but its objective effects are such as to reinforce and reproduce the behaviour ('you see, I knew they were born delinquents, that only repression will work', etc.); or else the agent sees the failure but either it induces him to insist ('Yes, we have made them all worse, but only repression will work') or else he sees also the positive results and no better alternative ('Yes, prison makes them worse, but I can't see any other way; in any case, delinquency must be repressed').

In conclusion, the reproduction is often independent of goodness, in the sense that it is not the effectiveness that reproduces the behaviour; it is, at least partly, independent of the subjective, perceived or believed goodness. However, goodness may not be sufficient or perhaps even necessary³¹. Of course it is true that the goodness of the

³¹ I shall leave this question open, as it requires further investigation. In the case of intentional behaviour it is analytical for there to be a goal and an albeit mild positive expectation; however, there are mechanisms for reproducing unintentional behaviours and

effects reproduces the action (by means of reinforcement; intentionality; evaluative instruction, etc.), but also negative, harmful results may reproduce the agent's behaviour instead of discouraging it. This is the line of reasoning we had to pursue.

8. The insufficiently "perverse" effects of Boudon: functions or accidents?

The important concept of 'perverse effects' as defined by Boudon (Boudon, 1977) is not in my opinion in contradiction with Hayek's view. The term is not the most felicitous, because it implies that there are bad effects, which is not what Boudon intended. In actual fact he defines them as "undesired effects, whether or not they are desirable or predictable. They are not an objective pursued directly by the individuals" (notes on p.14, and p.15, plus further explanations on p. 19). It would be preferable to say "unintended" rather than "undesired". Indeed the critical case is that of positive effects (desirable, that is, coinciding with goals) and known, anticipated, but not motivating the action, nor necessary or sufficient for it. It is true however -according to Boudon's criticism- that Merton's notion (*unanticipated consequences*) is clean (if they are 'unanticipated' a fortiori they are 'unintended') although it does not provide a complete picture.

However, the 'ontology' proposed by Boudon does not provide an adequate picture either, and furthermore does not call for just an intentional, goal-oriented agent (as he states very clearly on page 17) but a sophisticated theory of intentions (which is not available).

Boudon also makes a very clear analysis of what is meant by 'positive' or 'negative', 'good' or 'bad', with respect to what and to whom. He identifies as many as 18 possible cases (page 16). Only that he overlooks one fundamental distinction: whether the unintentional good or evil I produce is the same for myself as for others. Something that actually makes a vital difference to the decision (as the effects, although not intended, may be expected) as well as in social theory (as we have said with reference to Hayek).

In particular Boudon fails to take into account the fact there are not just perverse effects, in the sense of being concurrent and global as well as unintentional, and that may be good or harmful, but that these effects -including the perverse ones- may be not simply stable but actually self-organizing and self-reproducing.

The main differences with respect to the view we have illustrated are set out in the following.

8.1 Only side effects?

Boudon equates perverse effects with *emerging, compound, aggregate, collective* effects³² (pp.), using the terms interchangeably. That is, he is not interested in the same phenomenon at the individual or interpersonal level. He considers that only the aggregate aspect is of interest to social science. This is a limitation

- on the one hand, *for a general theory of the phenomenon* (what is the general model of functioning? is it due to aggregation as such or more basically to limits of individual cognition/rationality and feedback mechanisms, to vicious or virtuous circles?);

- on the other, *for social theory itself*. Indeed, important phenomena, such as self-fulfilling prophecies, (on which Merton rightly focuses maximum attention) do not refer only to combined effects (e.g. lack of confidence in banks and their going bust; racial prejudice, etc.), but refer also to two-way relations (e.g. the Pygmalion effect between teacher and pupil, cit.) or oneselves (vicious circle of low self-esteem and pessimism and their self-fulfilling nature - Miceli, 1998). Not only does the interpersonal level occupy a rightful place in the social sciences, but also the other level. Indeed perverse macro-phenomena and vicious circles could indeed emerge from and be implemented in individual (and even intrapsychic) and interpersonal vicious circles (Castelfranchi, 1998d; 1998e).

8.2 Function = good for certain agents?

Boudon actually eliminates all the *functional* and *finalistic* aspects of the phenomenon. He is not even aware of the difference between his terminology and that of Smith ('the invisible hand'), that is, he considers the concept of 'end' proposed by Smith as over-emphatic or simply reducible to an 'accidentally good result'.

In essence, he either actually eliminates any functional concept or else reduces it to being 'useful for the purposes of the community' (sic!) or (perhaps) of the majority. It does not even dawn on him that there may be a different notion of 'end' that it is necessary to clarify, *a notion that cannot be related back or reduced to the goals and interest of the agents*, but has to be constructed in terms of selection and self-organization mechanisms.

mechanisms of reproduction of intentional behaviour not through intentions. However, what counts in this dispute on the self-organizing emerging effects is what reproduces the behaviour, whether failure or the harm done.

³² He often identifies perverse effects (that is undesired and (perhaps) not good) with social *disequilibria*. It is not made very clear: is equilibrium, stability, non change, harmony, a social 'end'? and if so for whom?

The thesis I have illustrated is that:

a) there are three (in actual fact, two independent) scientifically grounded notions of behaviour or *finalistic* phenomenon:

- *intentional* (more generally based on internal goals, goal-oriented) the operational definition of which is provided by cybernetics;
- *selectionistic*, based on goals or purposes 'external' to the mind, not represented in it, the model for which comes from the evolutionary theory in biology, but which must be generalized to take in social and cultural functions;
- *organismic-physiological*, that is, the notion of 'function' in a system and related to its functioning. In my opinion, this has no theoretical autonomy of its own but can be subsumed in the preceding one.

b) social functions cannot be related to unintentional effects that are good for some and even for 'all' (absurd) or even for the majority. Indeed, if this goodness is appreciated, the result tends to become actively or even passively 'intended', desired (and as Elster would say, the notion of 'function' is redundant); if on the other hand the result is good but unknown, not understood, but is a function, then it would be necessary for not-good functions to exist. To make it quite clear, for me what distinguishes *function* from *non function* is not that the unintentional (collective) effect is good but that it is self-organizing and self-producing by means of positive feedback, that is, by reinforcing, selecting, and reproducing the behaviour it has generated: *unintended effects that select their own causes*. If 'function' means 'effect that happens to be good' this gives rise to two ticklish problems: *good for how many?* is one enough? many? the majority? everyone? And again, *how good?* just good? more good than bad? goodish?

In our approach, in which functionality is kept distinct from goodness (and that is from the subjective goals of the agents), on the one hand, good and bad functions (exactly like unintended good and bad effects) are on the same plane: both may be self-organizing. Secondly, the function is not reproduced or maintained or repeated by virtue of its good effects (a risky approach owing to the boundary with intention) although it must be specified -as Elster rightly claims- what the feedback and replication mechanism is.

In my opinion the psychologistic and subjectivistic reduction of functions to what is 'good' for the agents carried out both by Hayek and Boudon actually amounts to a *liquidation of the notion of function* and completely blurs the philosophers' intuition that there is a form of autonomous 'end' at this level of organization. It would be like postulating goals and preferences in the bodily organs in order to speak meaningfully of their functions, or of relating the latter to explicit goals, intentions and preferences of the organism as a whole (e.g. a worm). A more reasonable solution would be to attribute adaptive functions to the organism (e.g. worm) based on an evolutionary model and then, in relation to these adaptive functions, define as 'functional' certain internal processes, features, or behaviours, and define as functions their useful effects as far as the adaptive functions are concerned. However, the same argument could be applied to agents having internal goals, i.e. desires, preferences and intentions, and the behaviour and effects of which may be good or bad also with respect to the latter. But these two dimensions of 'goodness' would be completely independent, orthogonal to each other. That is, there may be good effects that become functions (they self-organize and self-reproduce), good effects that are not functions (which are perhaps repeated but not by virtue of themselves), bad effects that are functions of the behaviours producing them and accidental (although repeated) bad effects.

c) It is actually necessary to construct a theory covering the relation between external goals and internal goals of the action, and it is necessary to revise -and not liquidate- the theory of social functions, and also to revise the structural relationship between biological functions and subjective goals of behaviour, as well as that between subjective goals (and freedom) and social roles and prescriptions (Castelfranchi, 1982; 1998a; Conte and Castelfranchi, 1996).

In my opinion this is the principal limit of Boudon's work: he emphasizes -also by means of striking examples- the great importance of undesired effects in the social field, but he tries to do this by setting aside the finalism intuitively grasped by Smith and others (and probably considering it as metaphysical) and in any case without feeling the need to provide a theoretical basis for the controversial theory of social 'functions' which is instead very closely linked to it??, or to clarify by means of suitable cognitive models the link between external goals affecting behaviour (adaptive goals, social roles, functions, etc.) and intentions which regulate it directly. He does this essentially without explaining how to reconcile the fact that the agent's intentions are causes (and free) by the fact that they are also effects. (p.19)

9. Brief conclusions

Hayek is definitely right when he claims that the invisible hand is the theoretical heart of all social sciences; he is right when he claims that the same results could never be achieved through centralized knowledge, decision-making and planning; when he claims that many results with positive effects for persons can be achieved unconsciously and emergently; and when he uses the notion of 'function', albeit to a limited extent. The dissent is related to the decline

of any critical attitude towards the invisible hand and spontaneous order; when tautologically each and every order that is stabilized, *provided that it is spontaneous*, becomes good for the agents. Lastly, at the theoretical and epistemological level, I do not approve of the systematic confusion between 'ends' in a purely self-referential, evolutionary and functional sense, and 'ends' in a subject sense. I have tried to argue that these two distinct notions can and must be kept separate and each has an autonomous foundation, and that they are each scientifically feasible. But -as I have said- one unpleasant consequence of this is that self-referential teleonomic phenomena (such as spontaneous order, social functions, conventions, etc.) are not guaranteed to be functional to human needs, to be good for subjective human purposes. We would have to have a much less providential conception of the invisible hand³³ and a much less optimistic one of spontaneous order. The invisible hand, like everything 'natural', is by nature *indifferent* to the good and the interests of individuals (Leopardi). I have opposed to this philosophic view that one held by Leopardi the philosopher, his pessimism regarding the "magnificent and progressive destinies" and "natural regular evil", a view that I consider to be more disenchanting and realistic. However, the opposition is not based solely on world views, on 'philosophies', but on concrete operational theoretical models of human action and its organization at the micro and macro level. In other words: it is on the scientific plane that I find the optimism highly limiting and unjustified.

As a cognitive scientist, I thus believe that the limits of and the contradictions in Hayek's illuminating theory fundamentally derive from the model of man he adopted (Hayek, 1952b; Birner, in the present book), a rather empirical type of man. It is praiseworthy that everything is grounded on a theory of agent, of mind and of action (as well as of evolution) (Rizzello, 1997), but this theory is too impoverished; in particular, the anticipatory, intelligent and plan-making nature of the human mind is excessively diminished (also for ideological reasons); he takes for granted a large number of problematic assumptions concerning mind and action. For Hayek man seems to be only a knower (more exactly, a perceiver) who learns. He is not a true problem-solver (as in modern cognitive theory), or if he solves problems, he apparently does so only by chance, by trial and error, and not with intelligence and foresight. Rationality plays a very limited role. As Carabelli and De Vecchi say (in the present book) "For Hayek in essence it is not reason but 'mere habit' that guides social processes"; the individual's mind has very little creative and processing capacity; his wisdom is based on tradition, on collective experience consolidated and selected over time. This is a very unilateral and convenient view of man, no less than that of a rational problem-solver with perfect knowledge.

Bibliographic references

- Arthur, B. 1994, *Increasing Returns and Path Dependence in the Economy*. Univ. of Michigan Press, Ann Arbor.
- Boudon, R. 1977, *Effets pervers et ordre social*, PUF, Paris (It. trans. *Effetti Perversi dell'azione sociale*, Milano, Feltrinelli, 1981)
- Castelfranchi, C. 1982, *Scopi esterni*. "Rassegna Italiana di Sociologia", XXIII, 3
- Castelfranchi, C. 1997, *Social Functions. A Challenge for Agent-Based Social Simulation*. Invited talk at the "Simulating Societies" Workshop- SimSoc'97, Cortona.
- Castelfranchi, C. 1998a, *Through The Agents' Minds: Cognitive Mediators Of Social Action*. Conference on "Cognitive Theory of Social Action" - Fondazione Rosselli - Torino, 11-13 June 1988. To be published in "Mind and Society", I, 1.
- Castelfranchi, C. 1998b, *Emergence and Cognition: Towards a Synthetic Paradigm in AI and Cognitive Science*. In H. Coelho (ed.) *Progress in Artificial Intelligence - IBERAMIA 98*, Springer, Berlin, pp.13-16
- Castelfranchi, C. 1998c, *Modelling Social Action for AI Agents*. "Artificial Intelligence", 6.
- Castelfranchi, C. 1998d, *Prevenzione. Tra pessimismo della ragione ed ottimismo della volonta'*. In R. Piccione (ed.) *Prevenzione*. Carocci, Roma.
- Castelfranchi, C. 1998e, *Il nevrotico cripto-utilitarista. Contro l'ideologia del vantaggio secondario*. "Sistemi Intelligenti" X, 2, 307-15

³³ Let us remember Kundera's remark that "Optimism is the virtue of the oppressors" (public discussion, reported by Pontiggia); "oppressors" in the sense of people at ease in a given order while others, with much less power, suffer hardship.

- Conte, R. e Castelfranchi, C. 1996, *La societa' delle menti*. Torino, UTET
- Davis, P. 1985, *Clio and the Economics of QWERTY*. "American Economic Review", 75, 332-7.
- Dupuy, J.P., 1992, *Introduction aux sciences sociales. Logique des phenomènes collectifs*. Paris, Marketing Ed.
- Elster, J. 1982. *Marxism, functionalism and game-theory: the case for methodological individualism*. "Theory and Society" 11, 453-81.
- Forrest, S. 1990, (ed.) *Emergent computation*. Cambridge, Mass. MIT Press.
- Hayek, F. 1952, *The Counter-Revolution of Science: Studies on the Abuse of Reason*, Glencoe, The Free Press.
- Hayek, F. 1967, *The Results of Human Action but not of Human Design*. In Hayek, F. *Studies in Philosophy, Politics and Economics*, London, Routledge & Kegan Paul.
- Hayek, F.A. 1973, *Law, legislation, and liberty. A New Statement of the Liberal Principles of Justice and Political Economy Vol. I Rules and Order*, London: Routledge and Kegan Paul. (It. trans. Il Saggiatore, Milano, 1986)
- Hayek, F.A. 1976, *Law, legislation, and liberty. A New Statement of the Liberal Principles of Justice and Political Economy Vol. II. The mirage of social justice*. London: Routledge and Kegan Paul. (It. trans. Il Saggiatore, Milano, 1986)
- Hayek, F. 1988, *Conoscenza, Mercato, Pianificazione*, Bologna, Il Mulino
- Hayek, F. 1998, *La societa' libera*. Seam, ed.
- Kahaneman, D., Knetsch, J.L., and Thaler, R. 1986, *Fairness and the Assumptions of Economics*. "Journal of Business", 59 (4) 285-300.
- Knetsch, J.L. 1989, *The Endowment Effect and Evidence of Nonreversible Indifference Curves*. "American Economic Review". 79 (5), 1277-84.
- Knetsch, J.L., and Sinden, J.A. 1984, *Willingness to Pay and Compensation Demanded*. "Quarterly Journal of Economics", 99 (3), 507-21.
- Kunda, Z. 1990, *The case for motivated reasoning*. "Psychological Bulletin", 108, 480-498.
- Mackie, G. 1996, *Ending footbinding and infibulation: a convention account*. "American Sociological Review, 61, 999-1017.
- Macy, R. 1998, *Social Order in Artificial Worlds*. In "Journal of Artificial Societies and Social Simulation", I, 1.
- Mancini, F. 1998, *La mente ipocondriaca ed i suoi paradossi*. "Sistemi Intelligenti", X, 1, pp .85-108.
- Martin, I. and Levey, A.B. 1978, *Evaluative conditioning*. "Advances in Behavior Research and Therapy", 1, 57-102.
- Marx, K. 1973, *La miseria della filosofia*. Roma, Editori Riuniti.
- Mayr, E. 1974. Teleological and teleonomic: A new analysis; also appeared as: 1982. Learning, development and culture. In *Essays in evolutionary epistemology*, H.C. Plotkin (ed.). New York: John Wiley.
- Merton, R.K.1949, *Social Theory and Social Structure*. N.Y. The Free Press (It. trans. *Teoria e struttura sociale*, Il Mulino, Bologna, 1974)
- 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, 343-90.
- Miceli, M. 1998, *Autostima*. Bologna, Il Mulino

- Miceli, M. e Castelfranchi C. 1989, *A Cognitive Approach to Values*. "Journal for the Theory of Social Behaviour", 2, 169-94.
- Miceli, M. and Castelfranchi, C. 1992. *La cognizione del valore*. Milano: Franco Angeli.
- Miceli, M. & Castelfranchi, C. (in press). The role of evaluation in cognition and social interaction. In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology* (pp. 225-261). Amsterdam: John Benjamins.
- Rizzello, S. 1997, *L'economia della mente*. Laterza, Bari
- Rizzello, S. 1998, *Economic Change, Subjective Perception, and Institutional Evolution*. "Metroeconomica" (forthcoming)
- Rosenblueth, A, Wiener, N. and Bigelow J. 1968, Behavior, Purpose, and Teleology. In *Modern systems research for the behavioral scientist*, Buckley, W. (ed.). Chicago: Aldine.
- Rosenblueth, A. and Wiener N. 1968, Purposeful and Non-Purposeful Behavior. In *Modern systems research for the behavioral scientist*, Buckley, W. (ed.). Chicago: Aldine.
- Rosenthal,R. and Jacobson, L. 1972, *Pigmalione in classe*, Milano, Angeli.
- Thaler, R.H. 1985, Mental Accounting and Consumer Choice. "Marketing Sci". , 4(3), 199-214
- Vanberg, V. 1986, *Spontaneous Market Order and Social Rules: A Critical Examination of Hayek's Theory of Cultural Evolution*. Economics and Philosophy", 2.
- Virasoro, M.A., 1996. *Intervista sulla complessita'*, ed. 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, pp.1-80
- Wright, L. 1976, *Teleology Explanations: An Etiological Analysis of Goals and Functions*. University of California Press, Berkeley,.