Engineering Social Order [1]

Cristiano Castelfranchi

National Research Council - Institute of Psychology
Division of "Artificial Intelligence, Cognitive and Interaction Modelling"
castel@ip.rm.cnr.it

Abstract. Social Order becomes a major problem in MAS and in computer mediated human interaction. After explaining the notions of Social Order and Social Control, I claim that there are multiple and complementary approaches to Social Order and to its engineering: all of them must be exploited. In computer science one try to solve this problem by rigid formalisation and rules, constraining infrastructures, security devices, etc. I think that a more socially oriented approach is also needed. My point is that Social Control - and in particular decentralised and autonomous Social Control - will be one of the most effective approaches.

0. The framework: Social Order *Vs* Social Control

This is an introductory paper. I mean that I will not propose any solution to the problem of social order in engineering cybersocieties: neither theoretical solutions and even less practical solutions. I want just to contribute to circumscribe and clarify the problem, identify relevant issues and discuss some notions for a possible ontology in this domain.
I take a cognitive and social perspective, however I claim that this is relevant not only for the new born *computational social sciences*, but for networked society and MAS. There is a dialectic relationship: on the one hand, in MAS and cybersocieties we should be inspired by human social phenomena, on the other hand, by computationally modelling social phenomena we should provide a better understanding of them.
In particular I try to understand what Social Order [2] is, and to describe different approaches to and strategies for Social Order, with special attention to Social Control and its means. Since the agents (either human or artificial) are relatively autonomous, act in an open world, on the basis of their subjective and limited points of view and for their own interests or goals, Social Order is a problem. There is no possibility of application for a pre-determined, "hardwired" or designed social order. Social order has to be continuously restored and adjusted, dynamically produced by and through the action of the agents themselves; this is why Social Control is necessary.
There are multiple and complementary approaches to Social Order and to its engineering: all of them must be exploited. In computer science one try to solve this problem by rigid formalisation and rules, constraining infrastructures, security devices, etc. I think that a more socially oriented approach is also needed. My point is that Social Control - and in particular decentralised and autonomous Social Control - will be one of the most effective approaches.

1. The Big Problem: Apocalypse now

I feel that the main trouble of infosocieties, distributed computing, Agent-based paradigm, etc. will be -quite soon- that of the "social order" in the virtual or in artificial society, in the net, in MASs. Currently the problem is mainly perceived in terms of "security", and in terms of crisis, breakdowns, and traffic, but it is more general. The problem is

> **how to obtaining from local design and programming, and from local actions, interests, and views, *some desirable and relatively predictable/stable emergent results*.**

---

[1] This work has been and is being developed within the ***ALFEBIITE*** European Project**:** *A Logical Framework For Ethical Behaviour Between Infohabitants In The Information Trading Economy Of The Universal Information Ecosystem.* - IST- 1999-10298.

[2] The spreading identification between "social order" and cooperation is troublesome. I use here social order as "desirable", good social order (from the point of view of an observer or designer, or from the point of view of the participants. However, more generally *social order* should be conceived as any form of systemic phenomenon or structure which is sufficiently stable, or better either self-organising and self-reproducing through the actions of the agents, or consciously orchestrated by (some of) them. Social order is neither necessarily cooperative nor a "good" social function. Also systematic *dis-functions* (in Merton's terminology) are forms of social order. See section 3.

This problem is particularly serious in open environments and MASs, or with heterogeneous and self-interested agents, where a simple organisational solution doesn't work.
This problem has several facets: Emergent computation and indirect programming [For90]; [Cas98a]; reconciling individual and global goals [Hog97]; [Kir99]; the trade-off between initiative and control; etc.). Let me just sketch some of these perspectives on THE problem.

1.1 Towards Social Computing: Programming (with) 'the Invisible Hand'? [3]

Let me consider the problem from a computational and engineering perspective. It has been remarked how we are going towards a new "social" computational paradigm [Gas91; 98]. I believe that this should be taken in a radical way, where "social" does not means only organisation, roles, communication and interaction protocols, norms (and other forms of coordination and control); but it should be taken also in terms of spontaneous orders and self-organising structures. That is, one should consider the *emergent* character of computation in Agent-Based Computing.
In a sense, the current paradigm of computing is going beyond strict 'programming', and this is particularly true in the agents paradigm and in large and open MASs.
On the one hand the agents acquire more and more features such as:

- ***adaptivity***: either in the sense that they learn from their own experience and from previous stimuli; or in the sense that there may be some genetic recombination, mutation, and selection; or in the sense that they are reactive and opportunistic, able to adapt their goals and actions to local, unpredictable and evolving environments.
- ***autonomy and initiative*****:** the agent takes care of the task/objective, by executing it when it finds an opportunity, and proactively, without the direct command or the direct control of the user; it is possible to delegate not only a specified action or task but also an objective to bring about in any way; and the agent will find its way on the basis of its own learning and adaptation, its own local knowledge, its own competence and reasoning, problem solving and discretion.
- ***distribution and decentralisation*****:** MAS can be open and decentralised. It is neither established nor predictable which agent will be involved, which task it will adopt, and how it will execute or solve it. And during the execution the agent may remain open and reactive to incoming inputs and to the dynamics of its internal state (for example, resource shortage, or change of preferences). Assigned tasks are (in part) specified by the delegated agents: nobody knows the complete plan. Nobody entirely knows who delegated what to whom.

In other words, nobody will be able to specify where, when, why, and who is running a given piece of the resulting computation. The actual computation is just emergent. Nobody directly wroten the program that is being executed. We are closer to Adam Smith's notion of 'the invisible hand' than to a model of a plan or a program as a pre-specified sequence of steps to be passively executed.
This is on my view the problem of ***Emergent Computation*** (EC) as it applies to DAI/MAS.
Forrest [For90] presents the problem of Emergent Computation as follows:

> 'The idea that interactions among simple deterministic elements can produce *interesting and complex global behaviours* is well accepted in sciences. However, the field of computing is oriented towards building systems that accomplish *specific tasks*, and emergent properties of complex systems are inherently difficult to predict and control. ... It is not obvious how architectures that have many interactions with often unpredictable and self-organising effects can be used effectively. ....The premise of EC is that interesting and useful computational systems can be constructed by exploiting interactions among agents.
> .....The important point is that the explicit instructions are at different (and lower) level than the phenomena of interest. There is a tension between low-level explicit computations and *direct programming*, and the patterns of their interaction'.
>
> Thus, there is some sort of 'indirect programming': *implementing computations indirectly as emergent patterns*.

Strangely enough, Forrest - following the fashion of that moment of opposing an antisymbolic paradigm to the gofAI- does not mention DAI, AI agents or MAS at all; she just refers to connectionist models, cellular automata, biological and ALife models, and to the social sciences.
However, also higher level components -complex AI agents, cognitive agents - give rise precisely to the same phenomenon (like humans!). More than this, I claim that the 'central themes of EC' as identified by Todd [Tod93] are among the most typical DAI/MAS issues.
Central themes of EC include in fact [Tod93]:

- self-organisation, with no central authority to control the overall flow of computation;
- collective phenomena emerging from the interactions of locally-communicating autonomous agents;
- global cooperation among agents, to solve a common goal or share a common resource, being balanced against competition between them to create a more efficient overall system;
- learning and adaptation (and autonomous problem solving and negotiation) replacing direct programming for building working systems;

---

[3] This section is from [Cas98a].

- dynamic system behaviour taking precedence over traditional AI static data structures.

In sum, Agent based computing, complex AI agents, and MASs are simply meeting the problems of human society: functions and 'the invisible hand'; the problem of a spontaneous emergent order, of beneficial self-organisation, of the impossibility of planning; but also the problem of *harmful self-organising behaviours.*
Let's look at the same problem from other perspectives.

1.2 Modelling emergent and unaware cooperation among intentional agents

Macy [Mac98] is right when he claims that *social cooperation does not need agents' understanding, agreement, contracts, rational planning, collective decisions*. There are forms of cooperation that are deliberated and based on some agreement (like a company, a team, an organised strike), and other forms of cooperation that are emergent: non contractual and even unaware. Modelling those forms is very important but my claim [Cas97] [Cas92a] is that it is important to model them not just among sub-cognitive agents [4] (using learning or selection of simple rules) [Ste80] [Mat92], but also among cognitive and planning agents [5] whose behaviour is regulated by anticipatory representations (the "future"). Also *these agents cannot understand, predict, and govern all the global and compound effects of their actions at the collective level*. Some of these effects are self-reinforcing and self-organising.
I argue that it is not sufficient to put deliberation and intentional action (with intended effects) together with some reactive or rule-based or associative layer/ behaviour, let some unintended social function emerge from this layer, and let the feedback of the unintended reinforcing effects operate on this layer [Par82]. The real issue is precisely that *the intentional actions of the agents give rise to functional, unaware collective phenomena* (e.g., the division of labour), not (only) their unintentional behaviours. How to build unaware functions and cooperation on top of intentional actions and intended effects? How is it possible that positive results -thanks to their advantages- reinforce and reproduce the actions of intentional agents, and self-organise and reproduce themselves, without becoming simple intentions? [Els82]. This is the real theoretical challenge for reconciling emergence with cognition, intentional behavior with social functions, planning agents with unaware cooperation. At the SimSoc'97 workshop in Cortona [Cas97] I claimed that only agent based social simulation joint with AI models of agents can eventually solve this problem by formally modelling and simulating *at the same time* the individual minds and behaviours, the emerging collective action, structure or effect, and their feedback to shape minds and reproduce themselves.
I suggested that we need more complex forms of reinforcement learning not just based on classifiers, rules, associations, etc. but *operating on the cognitive representations governing the action, i.e. on beliefs and goals*.
My claim is precisely that "the consequences of the action, which may or may not have been consciously anticipated, modify the probability that the action will be repeated next time the input conditions are met" [Cas97; Cas98c]:

> ***Functions*** *are just effects of the behavior of the agents, that go beyond the intended effects (are not intended) and succeed in reproducing themselves because they reinforce the beliefs and the goals of the agents that caused that behavior.*

1.3 Reconciling Individual with Global Goals

> "Typically, (intelligent) agents are assumed to pursue their own, individual goals. On this bases, a diversity of agent architectures (such as BDI), behavioral strategies (benevolent, antagonistic, etc.), and group formation models (joint intentions, coalition formation, and others) have been developed. All these approaches involve a bottom-up perspective. The existence of a collection of agents thus depends on a dynamic network of (in most cases: bilateral) individual commitments. Global behavior (macro level) emerges from individual activities/interactions (micro level). .... In business environments, however, the behavior of the global system (i.e., on the macro level) is important as well. Typical requirements concern system stability over time, a minimum level of predictability, an adequate relationship to (human-based) social systems, and a clear commitment to aims, strategies, tasks, and processes of enterprises. Introducing agents into business information systems thus requires to resolve this conflict of bottom up and top down oriented perspectives." [6] [Kir99]

---

[4] By "sub-cognitive" agents I mean agents whose behaviour is not regulated by an internal explicit representation of its purpose and by explicit beliefs. Sub-cognitive agents are for example simple neural-net agents, or mere reactive agents.

[5] Cognitive agents are agents whose actions are internally regulated by goals (goal-directed) and whose goals, decisions, and plans are based on beliefs. Both goals and beliefs are cognitive representations that can be internally generated, manipulated, and subject to inferences and reasoning. Since a cognitive agent may have more than one goal active in the same situation, it must have some form of choice/decision, based on some "reason" i.e. on some belief and evaluation.
Notice that we use "goal" as the general family term for all motivational representations: from desires to intentions, from objectives to motives, from needs to ambitions, etc.

[6] I would say between an *individualist* and a *collectivist* perrspective. Kirn proposes to deal with our problem an an organizational approach . This is quite traditional in MAS and is surely useful. However I claim in this paper that is largely insufficient.

This is another point of view on the same problem. It can be formulated as follows: How to reconcile individual rationality with group achievements?
Given goal-autonomous agents [7], basically there are two solutions to this problem of making the agent "sensible" to the collective interest:

a) to use *external incentives*: such as prizes, punishments, redistribution of incomes, in general *rewards*, for ex. money (for example to make industries sensible to the environmental problem you can put taxes on pollution), so that the agent will find *convenient* - relatively to his/her selfish motives and utility - to do something for the group (to favour the group or to do as requested by the group);

b) to endow the agent with *pro-social motives and attitudes* (sympathy, group identity, altruism, etc.) either based on social emotions or not, either acquired (learning, socialisation) or inborn (by inheritance or design); in this case there is an *intrinsic* pro-group *motivation*. The agent is subjectively rational - although not economically rational- but ready to sacrifice [8].

Human societies use both these approaches [9]; this is not casual. We should experiment advantages and disadvantages of the two, to see on which domain and why one is better than the other.

Experimental social simulation can give a precious contribution to cope with this problem (for ex. [Gil95; Con97; Kam00]]; but also good formal theories of spontaneous and non-spontaneous social order and of its mechanisms, and in particular theories of spontaneous and deliberated forms of "social control", will play a major role.
Clearly deontic stuff (norms, conventions, institutions, roles, commitments, obligations, rights, etc.) will have a large part in any "implementation" of social control mechanisms in virtual and artificial societies. However, on my view, this role will be improved by a more open and flexible view of deontic phenomena and by the embedding of them within the framework of social control and social order.
On the one hand, one should realised how social control and order has not only normative solutions (organisation and norm are not enough); on the other side -more important- one should account for a flexible view of normative behaviour, and for a gradual approach to norms from more informal and spontaneous phenomena (like spontaneous conventions and social norms, or spontaneous decentralised social control) to more institutionalised and formal forms of deontic regulation [Cas00].

2 Approaches and delusions

There are different "philosophies" about this very complex problem; different approaches and policies. For example:

- A Coordination media and infrastructures approach, where thanks to some infrastructures and media a desirable coordination among independent agents actions is obtained [Den99].
- A Social Control (SC) approach that is focused on sanctions, incentives, control, reputation, etc.
- An Organizational approach, relying on roles, division of labor, pre-established multi-agent plans, negotiation, agreements, etc.
- A Shared mind view of groups, teams, organisations where coordination is due to shared mental maps of the domain, task, and organisation, or to common knowledge of the individual minds
- A Spontaneous Social Order approach where nobody can bear in mind, understand, monitor or plan the global resulting effects of the "invisible hand" (von Hayek) [Hay67].

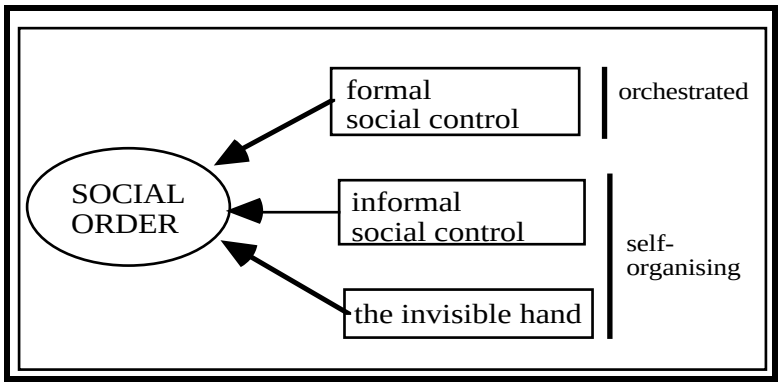
First, those approaches or views are not really conceptually independent of each other: frequently one partially overlaps with an other, simply hiding some aspects or notion. For example coordination media are frequently rules and norms; and the same is true for organisational notions like "role" that are normative notions. All of them exploit very much communication.

---

[7] i.e. self-motivated agents (self-interested but not necessarily selfish) that adopt goals only instrumentally to some goals of them (be these either selfish or altruistic).
[8] This might be also objectively rational, adaptive: it depends on ecological and evolutionary factors.
[9] There is a sort of intermediate or double-face phenomenon which, depending on its use or modelling, can be considered as part either of (a) or of (b): *internal gratification or punishment* (like guilt). If the agent does something pro-social *in order to* avoid guilt, this is selfish motivation and behavior (case a); but, notice that guilt feelings presuppose some pro-social goals like equity or norm respect!! If, on the contrary, the agents act fairly and honestly just *for* these pro-social motives (not in order to avoid regret), and them feel guilty when violate (and guilt is just a learning device for socialisation) we are in case (b). The psychological notion of intrinsic motivation or internal reward mixes up the two. Both perspectives are realistic and even compatible.

Second, in human groups -as I said- all these approaches are used, and Social Order is the result of both spontaneous dynamics and orchestrated and designed actions and constrains. It can be the result of SC and of other mechanisms like the "invisible hand", social influence and learning, socialization etc. But SC itself is ambiguous: it can be deliberated, official and institutional, or spontaneous, informal, and even unaware. I will completely put aside here education and social learning, and prosocial built-in motives (except for normative ones), although they play a very important role for shaping social order.

To be schematic, let's put at one extreme the merely self-organising forms unrelated to SC (ex.. market equilibrium) [10]; on the other extreme the deliberated and planned SC; in between there are forms of spontaneous, self-organising SC:



In IT all of these approaches will prove to be useful. For the moment the most appealing solutions are:

- on the one hand, what I would like to call the "organisational" solution (pre-established roles, some hierarchies, clear responsibilities, etc.. This is more "designed", engineered, and rather reassuring!
- on the other hand, the "normative" or "deontic" solution, based on the formalization of permissions, obligations, authorisation, delegation, etc. logically controllable in their coherence; [11]
- finally, the strictly "economic" solution based on free rational agent dealing with utilities and incentives in some form of market.

The problem is much more complex, and -in my view- several other convergent solutions should be explored. However, I will consider here only one facet of the problem. My view is that

> *"normative" SC but spontaneous and decentrealised has to play a major role both in cybersocieties and in artificial social systems.*

In IT there are some illusions about possible solutions.

The Illusion of Control: Security *Vs* Morality

The first, spontaneous approach of engineers and computer scientists to those issues is that of increasing security by certification, protocols, authentication, cryptography, central control, rigid rules, etc. Although some of these measures are surely useful and needed, as I said, I believe that the idea of a total control and a technical prevention against chaos, conflicts and deception in computers is unrealistic and even self-defeating in some case, like for building trust. Close to this illusion is the **formal-norm illusion** that I have already criticised (see later 5.).

The socio-anthropological illusion: let's embed technology in a social, moral and legal human context

In the area of information systems a perspective is already developing aimed at embedding the information system in a complex socio-cultural environment where there are -on the top of the technical and security layer- other layers relative to legal aspects, social interaction, trust, and morality [Har95; Lei96]. For sure this is a correct view. The new technology can properly work for human purposes only if integrated and situated in human morality, culture and law. However this is not enough.

For Intelligent Normative Supports (Agents)

I believe [Cas00b] that what is needed is some attempt

> *to "incorporate" part of these layers and issue in the technology itself.*

---

[10] Let's ignore here other factors of Social Order like constraining infrastrucuires for coordination (see section 4.).

[11] These two solutions can be strongly related one with the other, since one can give a normative interpretation and realisation of roles, hierarchies, organisations.

Especially within the intelligent and autonomous agents paradigm, I believe that it is both possible and necessary to model these typically human and social notions. In order to effectively support human cooperation - which is strongly based on social, moral, and legal notions- computers must be able to model and "understand" at least partially what happens among the users. They should be able to manage -and then partially "understand"- for ex. permissions, obligations, power, roles, commitments, trust.
Moreover, to cope with the open, unpredictable social interaction and collective activity that will emerge among them, the artificial agents themselves should base this interaction on something like organisation, role, norms, etc. This is in fact what is happening in the domain of agents and MAS, where these topics are in the focus of theoretical and formal modelling, and of some implementation. (Just to give some example see DEON Ws, ModelAge project, the IP-CNR work; Jennings; Moses; Tennenholtz; Singh; Boman; etc.].

3. Social Order

There are at least two notions of "Social Order".

Given a system of multiple entities, order is a *structure* of relationships
- among those entities or sub-units,
- of each of them with the whole, and
- with the environment of the system.

Such a trim, structure or pattern is "regular".
"Regular" means either:
α**)** corresponding to a given "regularity", i.e., to a norm in statistical or previsional terms; or
β**)** corresponding to some rule (Latin: regula), i.e., some norm or standard in a normative/deontic sense.

The α–meaning is broader and weaker: order is just a recurrent or stable emergent pattern. The β–meaning is stronger: the order is "intended" or at least desirable (from the point of view of some agent) or "functional". In any case, it is not simply accidental but finalistic/teleonomic.

Social Order is a non-accidental and non chaotic (thus, relatively predictable, repeated and stable) pattern of interactions in a given system of interfering agents, such that it allows the satisfaction of the interests of some agent A (or avoids their damage) [12]

Thus Social Order is relative to some system of interests or points of view, that makes it good or bad, desirable or undesirable. This point of view or reference can be:
- a single lay agent or group (internal or external to the order: observer);
- an agent with a special role: an authority, able (at least partially) to orient the system towards such an order;
- a shared goal, value, that is good for everybody or for most of the members (for example to reduce or avoid accidents; to guarantee the group survival; etc.);
- a real "common goal", in a strictly cooperative group.

Dynamic Social Order occurs when the stable macro-pattern or equilibrium is maintained thanks to *an incessant local (micro) activity* of its units, able to restore or reproduce the desired features. The global stability is due to local instability. Social order is very dynamic, especially with autonomous agents.

Given this notion of Social Order, let us examine some approaches and means for it.

4. Coordination media and infrastructures

Both "coordination" and "coordination media" are rather vague notions. "Coordination" can cover any good emerging social structure, any "order". I would like to use it not as a synonym of Social Order but in a more strict and delimited way [Cas99a].
"Coordination media" is also too broad. It can cover any device used/ful for coordination: communication of any kind, norms, (shared) mental models, representations and maps, roles, plans, trust, etc. etc. etc.
I do not find so useful such a broad set/container. I prefer to focus on more specific notions, for example norms (see later) or *constraining infrastructures* in a strict sense.

Virtuous coordination among the individual autonomous activities (producing a desirable global effect) is obtained sometimes thanks to coordination infrastructures, like for example the physical *barriers* for traffic in the

---

[12] The problem of social functions and functional order is far more complicated [Cas97]. Let's put it aside here.

cities (guard-rails, walls, side-walks, etc.) or like corrals for animals, etc. or like accessible or non-accessible blackboard structures, or pre-established communication channels among computational agents.
What characterise this means for coordination or better for SO, is the fact that

> **there is a *practical impossibility* for the agents to deviate from the desired behavior**: the action is externally constrained by its execution or success pre-conditions.

I propose to call them: *constraining infrastructures* or *barriers*.
Every action (either complex like a plan, or elementary) has certain preconditions for being materially executed or for succeeding. For example, the action of "switching the light on" has as its execution preconditions that there is an accessible and reachable switch, that it works, that I can move my arm appropriately, etc. Condition for success is that there is power in the cables, that there is a working lamp; that cables are connected, etc. If some of these conditions do not hold, even though I correctly perform the motor part of my action (switching the switch), I will not switch the light on.
By establishing certain "material" conditions in the environment that the agent is not able to change, I prevent it from doing a given action (or even from attempting to do it), and at the same time -since it is motivated to achieve a given goal- I channel its behavior in a given direction (towards the remaining practicable solutions) i.e.; towards what is practically permitted. Sometimes towards the only remaining possible move, that becomes "obligatory" since there are no degrees of freedom for the agent. For example, if you put me in a tube, and I want to go out, and you block one access (my way in), I'm *obliged* or better constrained to go straight head, as you want.

Barriers can be physical for physical agents (in the sense that the agent meets them physically in attempting to execute the action and failing). But they can also be non-physical. If the agent has no mental alternatives to that behavior, if it does not "choose" and "prefer", its behavior is constrained by a cognitive *practical impossibility* to do differently.
This is why true normative constraints (rules and conventions) are different: agents are free to obey to or violate them. [13]
The very difference can be characterised in the following way. Considering the **architecture** of an autonomous cognitive agent [Cas95; Con95] and how the goal-directed and belief-based intentional action is generated, we can say that:
the conformity of the behavior to the expected/desired standard, or better the elimination of the undesirable behaviors can be obtained (at least) in two very different ways:

> **i)** by means of the beliefs relative to **practical impossibility**: "I cannot do" "I'm not able" ("Necessary conditions are not there and I cannot create them").
> "*Barriers*" exploit this path.
>
> **ii)** by means of the beliefs relative to **utility and preference**: "it is not convenient for me" "too high risks"; "too high costs".
> *Norms* and *incentives* exploit this path.

Since those beliefs determine the persistence or the abandon of a given goal (at the desire or at the intention level of processing) giving/changing those beliefs we create the practical or the decision impossibility to perform an undesirable behavior.
Another important distinction, more general and applicable also to sub-cognitive rule-based agents, is between

- impossibility due to the internal stuff, to an internal lack of power of the agent (for example, the absence of a rule, of a skill, or the rule repertoire); and
- impossibility ascribable to external conditions and circumstances.

5. The feedback problem in Social Order

A very relevant problem (and an essential means) is how to model the ***feedback*** that must arrive at the local agent level (and enter its cognitive processes and decision, or its learning, or its reactive response) in order for it to adjust its behavior in the "right" way (relative to the desired global result). I mean the feedback from the global emerging structure to the individual agents.

Several theoretical possibilities should be explored.
a) the agents have/receive a representation of the global resulting structure or phenomenon

---

[13] A lot of what is called "infrastructure" are in fact norms, signals/messages to the agents (for ex. semaphores). Rules/norms entail in fact messages for giving the commands to the agents.

a1) they care for the global result and regulate their behaviors to achieve a specific global structure. This is the case for ex. of 'ring-a-ring-o'-roses' [14] where the agents adjust their own behavior by checking the global circular structure; they can also communicate with each other and prescribe the others certain behaviors in order to co-ordinate with each other.

a2) they receive the global result information but they adjust to it for their own purposes. For example, there is a traffic jam (or they receive the warming that there is a clogging on the highway) and they change their own route to avoid the row They do not care for reducing or not producing the jam, they just care of their own goal, but they adjust on a global-structure feedback.

b) The agents do not have a representation of the global result (they would not be able to understand it, or it is too complex and costly, etc.); they just receive some partial information about or feedback from the global result and they adjust their personal behavior (for their personal goals) to this feedback.
The most celebrated and beautiful example is from economics: the theory of price as information about a general market equilibrium: price is the necessary and sufficient information the agent has to locally know in order to decide; by this information and adjusted local decisions the global equilibrium (that nobody calculates or intends) is achieved.

However, these are just three basic possibilities; indeed a systematic theory of different kinds and functions of that feedback from the global to the local is needed.

Let me now focus on what I believe to be the most effective and important approach to SO, i.e. SC.

6. Social Control Approach to *Dynamic Social Order*

As I said at the beginning, precisely because agents are relatively autonomous, act in an open world, on the basis of their subjective and limited points of view and for their own interests or goals, Social Order is a problem. There is no possibility of application for a pre-determined, "hardwired" or designed social order. Social order has to be continuously restored and adjusted, dynamically produced by and through the action of the agents themselves; this is why social control is necessary.

6.1 Social Control (SC)

Let's take SC in a strict sense, not as any form of *social influence* on agents and of *socialization* (although obviously both can strongly contribute to social order). Let's consider SC only as the process through which, if/when an individual or a group derogates from the expected and prescribed degree of obedience to a norm, its behavior is led back to that degree of conformity [Hom50]. Social control is a reaction to deviant behavior, and is strongly related to the notion of 'sanction' [Par51].
However, let's also consider pro-active actions, prevention from deviation and reinforcement of correct behavior, and then also "positive" sanctions, social approval (also an implying implicit message/disapproval for deviant people) as part of SC.

I consider SC *any action of an agent aimed at (intended or destined to) enforcing the conformity of the behavior of another agent to some social norm* (in a broad sense, including social roles, conventions, social commitments, etc.) (see also the definition by Cos). [15]

Thus, not all socialization is SC (there are also other purposes and functions in socialization); not all social influence is SC. However, not only negative sanctions and post-hoc corrective interventions are SC.
As defined SC strictly presupposes:

a) some informal or formal, implicit or explicit, social or legal norm or convention in the society/group,

b) the possibility for the agent to deviate from it and to be led back (through psychological means) to the right behavior. SC presupposes autonomous agents, and possibly decision makers and normative agents [Con98a; Cas99b; Ver00], or at least some learning capability based on social rewards.

There are different forms of SC. It is useful to distinguish at least: deliberative/intended *Vs* unintended/functional SC; and centralized *Vs* decentralized SC.

---

[14] I take this nice example from a talk by Huhns at AOIS'99.

[15] I am calso lose to Johnson's [Joh60] view that SC consists in the action of all those mechanisms that neutralize deviant tendencies, either by preventing deviant behaviors, or -more important- by monitoring and inverting the motivational factors that can produce the deviant behavior. I only find too general the idea of SC as all the mechanisms that actually produce those effects. I prefer to restrict SC to those mechanisms aimed at producing those effects (functions) [Cas97].

	intended	unintended
centralized	A	B
decentralized	C	D

My claim (see also [Con98b]; [Cas99b]; [Kam00] [16] is that A is not the prototypical or the most efficacious form of SC in open and dynamic MAS or societies; a fundamental role is played by C and D, i.e., the spontaneous, bottom-up normative intervention of the distributed individuals, either conscious of their effects and intending to make another conform to the norms, or not intentionally oriented to this but in fact functional to this. For example, when Ag1 - a victim- complains for its pain or damage or aggressively reacts against Ag2 -the culprit- one of the effects and of the functions of these reactions is precisely to make Ag2's behavior conform to norms. Analogously, people just moving around in the city and observing the other people (what they are doing) are in fact exerting a form of non-voluntary SC (and this behavior actually reduces crimes).
Decentralized and in particular unintended SC (box D) are specially important in order to understand that also SC has some unplanned form and contributes to a merely emergent and self-organized Social Order.
Moreover, not only the monitoring and the intervention can be bottom-up and spontaneous (while the norm can still be formal and official); all normative bonds can be bottom-up, informal and spontaneous: the creation of conventions and norms, the establishment of rights, duties, permissions, etc. [17] [Cas00].
Given the features of infosocieties, computer mediated interactions, and of heterogeneous MASs my point is that we should understand and model these forms of SC, in order to engineer them in cybersociety. Formal and top-down forms of control cannot be enough.

6.2 Three Postulates

In sum, my analysis is framed by the following general assumptions:

- Social Order does not only depend on SC and is not reducible to SC

Other mechanisms (for ex. the so called "invisible hand", the emergent result of self-interest) produce social order, like the natural division of labour, or the market equilibrium. We do not define in fact social order as norm conformity. This is too strong. Even "desirable" social order is not simply or necessary norm-conformity. However, in this paper we have been mainly interested in a SC approach to Social Order, not in other mechanisms for emergent, self-organising Social Order.

- SC is not due to norms only and is not reducible to them [18]

There are various forms of and instruments for SC, beyond the explicit use of (social and legal) norms; ex. imitation; incentives; learning; etc.

- Normative means are not only based on or reducible to formal, top down, institutional norms

  The *informal* normative relationships, *the Micro/ Bottom-Up/ Decentralized/ Spontaneous Normative Social Control* are very important also for artificial systems.

---

[16] Kaminka and Tambe [Kam00] present interesting results.They claim that Agents in dynamic multi-agent environments must monitor their peers to execute individual and group plans. They show that a centralized scheme using a complex algorithm trades correctness for completeness and requires monitoring all teammates. By contrast, a simple distributed teamwork monitoring algorithm results in correct and complete detection of teamwork failures, despite relying on limited, uncertain knowledge, and monitoring only key agents in a team.

[17] Also the classical distinction between *external* and *internal* control (self-control) [Hom50] is very important. True norms are aimed in fact at the internal control by the addressee itself as a cognitive deliberative agent, able to understand a norm as such and adopt it [Con95]. Even more, norms are aimed at being adopted for specific reasons by the agents. The use of external control and sanctions is only a sub-ideal situation and obligation [Cas99c]. In artificial agents internal control is possible with real decision makers, be either norm-sensible agents or utility-sensible agents. Although self-control is for sure very important for Social Order and also for conformity to norms, I do not want to consider here it as a form of SC (see in fact our definition in terms of two different agents).

[18] Although -as I said- SC always presupposes some form of norm or convention, the norm itself is not the only means and strategy for obtaining conformity to norm. I can obtain conformity also by agents that lack any normative mind and understanding.

Recapitulation

- Social Order does coincide with either Social Control or Organisation; there are multiple and complementary approaches to Social Order and to its engineering: all of them must be exploited (thus modelled) in artificial systems.
- It is not possible to really "design" societies and the Social Order; more precisely it is possible only in some cases like in certain kinds of organisations; it is necessary to design only indirectly, i.e. to design frameworks, constraints and conditions (both internal ad external to the agents) in which the society can spontaneously self-organise and realise the desirable global effects.
- There is some illusion in computer science about solving this problem by rigid formalisation and rules, constraining infrastructures, security devices, etc. and there is scepticism or irritation towards more soft and "social" approaches, that leave more room to spontaneous emergence, or to decentralised control, or to normative "stuff" which is not externally imposed but internally managed by the agents.
- Social modelling will be the principal solution; it should leave some flexibility and try to deal with emergent and spontaneous forms of organisation; but there are serious problems like that of modelling the feedback from the global results to the local/individual layer.
- Social Control (and in particular decentralised and autonomous SC) will be one of the most effective approaches.

References

[Bin98] Binmore, K., Castelfranchi, C., Doran, J. and Wooldridge, M. Rationality in Multi-Agent Systems. The Knowledge Engineering Review, Vol. 13:3, 1998, 309-14.

[Bon89] A. H. Bond, Commitments, Some DAI insigths from Symbolic Interactionist Sociology. *AAAI Workshop on DAI*. 239-261. Menlo Park, Calif.: AAAI, Inc. 1989.

[Cas92] C. Castelfranchi and R. Conte. Emergent functionalitiy among intelligent systems: Cooperation within and without minds. *AI & Society, 6,* 78-93, 1992.

[Cas95] Castelfranchi, C., Guaranties for Autonomy in Cognitive Agent Architecture, in N. Jennings and M. Wooldridge (eds.) *Agent Theories, Architectures, and Languages*, Heidelberg, Spinger/Verlag, 1995

[Cas96] Castelfranchi, C., Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'96*, S.Francisco, June 1996, AAAI-MIT Press.

[Cas97] Castelfranchi, C. Challenges for agent-based social simulation. The theory of social functions. IP-CNR, TR. Sett.97; invited talk at *SimSoc'97,* Cortona, Italy

[Cas98a] Castelfranchi,C. Emergence and Cognition: Towards a Synthetic Paradigm in AI and Cognitive Science. In H. Coelho (Ed.) *Progress in Artificial Intelligence* - IBERAMIA 98, Springer, LNAI1484, Berlin, 1998 13-26

[Cas98c] Castelfranchi, C.,Through the minds of the agents, *Journal of Artificial Societies and Social Simulation*, 1(1), 1998, http://www.soc.surrey.ac.uk/JASSS/1/1/contents.html

[Cas99a] Castelfranchi, C., Modelling Social Action for AI Agents. *Artificial Intelligence*, 1999.

[Cas99b] Castelfranchi, C., Dignum, F., Jonker, C., Treur, J. (1999) Deliberate Normative Agents: Principles and Architecture. *ATAL'99,* Boston

[Cas00] Castelfranchi, C. Formalising the Informal? (invited talk). DEON'00, Touluse.

[Cas00b] Castelfranchi, C. Artificial liars: Why computers will (necessarily) deceive us and each other. Ethics and Information Technology, 00: 1-7, 2000.

[Cas-in press] Castelfranchi, C. and Tan, Y.H. (eds.) (in press) "Trust, Deception and Fraud in Artificial Societies" Kluwer.

[Con95] Conte,R. and Castelfranchi, C. *Cognitive and Social Action*, UCL Press, London, 1995.

[Con97] Conte, R., Hegselmann, R. and Terna, P. (eds) Studies in social simulation. Berlin, Springer, 1997

[Con98a] Conte, R. e Castelfranchi, C. (1998) From conventions to prescriptions. Towards a unified theory of norms. *AI&Law*, 1998. 3.

[Con98b] Conte, R., Castelfranchi, C., Dignum, F. Autonomous Norm Acceptance. In J. Mueller (ed) *Proceedings of the 5th International workshop on Agent Theories Architectures and Languages*, Paris, 4-7 July, 1998 (in press).

[Den99] Denti, E., Omicini, A., Toschi, V. Coordination technology for the development of MAS on the Web. In E. Lamma and P. Mello (eds.) AI*IA '99 Congress, Bologna, Pitagora Editrice. pag. 29-38

[For90] Forrest, *Emergent Computation*, 1990

[Gas91] L. Gasser. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence* 47: 107-138.

[Gas98] L. Gasser. Invited talk at *Autonomous Agents'98*, Minneapoli May 1998.

[Gil95] Gilbert, N., and Conte, R. (eds) Artificial Societies: the computer simulation of social life. London, UCL Press, 1995

[Gro95] B. Grosz, Collaborative Systems. AI Magazine, summer 1996, 67-85.

[Har95] Hartmann, A. (1995) "Comprehensive information technology security: A new approach to respond ethical and social issues surrounding information security in the 21st century". In IFIP TCI 11 Intern. Conf. of Information Security

[Hay67] F.A. Hayek, The result of human action but not of human design. In *Studies in Philosophy, Politics and Economics*, Routledge & Kegan, London, 1967.

[Hog97] Hogg L. M. and Jennings N. R. (1997) : Socially Rational Agents, in Proc. AAAI Fallsymposium on Socially Intelligent Agents, Boston, Mass., November 8-10, 61-63.

[Hom50] Homans, G.C., *The Human Group*, N.Y. 1950

[Joh60] Johnson, H.M. *Sociology: a systematic introduction*. N.Y. 1960

[Jen93] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review* 3, 1993: 223-50.

[Jon96] Jones, A.J.I. & Sergot, M. 1996. A Formal Characterisation of Institutionalised Power. Journal of the Interest Group in Pure and Applied Logics, 4(3): 427-45.

[Kam00] Kaminka, G.A. and Tambe, M. (2000) "Robust Agent Teams via Socially-Attentive Monitoring", Volume 12, pages 105-147.

[Kir99] Kirn, S. Invited Talk. *AOIS'99*, Hidelberg.

[Lei96] Leiwo, J. and Heikkuri, S. (1996) An Analysis of Ethics Foundations of Information Security in Distributed Systems, TR. Nokia TeleC, Helsinki

[Mac98] Macy, R. , In *JASSS*, I, 1, 1998.

[Mat92] M. Mataric. Designing Emergent Behaviors: From Local Interactions to Collective Intelligence. In *Simulation of Adaptive Behavior 2*. MIT Press. Cambridge, 1992.

[Mer49] Merton, R.K. *Social Theory and Social Structure*. Glencoe, Ill. 1949.

[O'H96] O'Hare, G. and Jennings, N R (Eds) (1996) *Foundations of Distributed AI*, John Wiley & Sons.

[Par51] Parsons, T. *The Social System*, Glencoe, Ill. 1951

[Sch95] Scheuch, E.K. *Controllo sociale*. In Enciclopedia delle Scienze Sociali, Treccani, 1995.

[Sin91] M.P. Singh, Social and Psychological Commitments in Multiagent Systems. In Preproceedings of "Knowledge and Action at Social & Organizational Levels", Fall Symposium Series, 1991. Menlo Park, Calif.: AAAI, Inc.

[Ste90] L. Steels. Cooperation between distributed agents through self-organization. In Y. Demazeau & J.P. Mueller (eds.) *Decentralized AI* North-Holland, Elsevier, 1990.

[Tuo93] Tuomela, R. What is Cooperation. *Erkenntnis*, 38, 1993, 87-101

[Ver00] Verhagen, H. *Normative Autonomous Agents*. PhD. Thesis, University of Stockolm, May, 2000

[Wag97] Wagner, G. (1997) "Multi-Level Security in Multiagent Systems", in P. Kandzia and M. Klusch (Eds), *Cooperative Information Agents*, Springer LNAI 1202, 1997, 272-285

[Woo95b] Wooldridge M.J. and Jennings N.R. (Eds.) 1995 *Intelligent Agents: Theories, Architectures, and Languages.* LNAI 890, Springer-Verlag, Heidelberg, Germany.