Cristiano Castelfranchi (1998) 'Through the Minds of the Agents'

***Journal of Artificial Societies and Social Simulation* vol. 1, no. 1, <http://jasss.soc.surrey.ac.uk/1/1/5.html>**

To cite articles published in the *Journal of Artificial Societies and Social Simulation,* please reference the above information and include paragraph numbers if necessary

Received: 28-Dec-1997 Published: 3-Jan-1998

This is a contribution to the Forum section, which features debates, controversies and work in progress. There are responses to this article.

Response

This is a response to Michael Macy's contribution to the JASSS Forum, Social Order in Artificial Worlds.

Keywords:
cooperation, altruism, rationality, cognition

1.1
I agree with several important claims in Macy's paper (in particular with the thesis that "the consequences of the action, which may or may not have been consciously anticipated, then modify the probability that the action will be repeated next time"). However, I disagree about Macy's central claim against the importance of the representation of "the future" (i.e. expectations, goals, plans, deliberation) in cooperation and social order. I trace back such a disagreement to the fact that I do not share several relevant presuppositions of Macy's argumentation. Let us see which are these presuppositions.

Intentional = Rational = Selfish?
Cooperation Or Altruism = Not Intentional?

2.1
Although the traditional game-theoretic approach is being criticised in Macy's statement, nevertheless an unjustified presupposition remains along his argumentation: *the usual and unprincipled identification between rationality and selfishness* is given for granted.

2.2
It seems to me that such an implicit identification between reason, goal-directed action, plans, deliberation and selfishness should not be accepted (Conte and Castelfranchi, 1995).

2.3
Rationality is not selfish, and neither altruism nor cooperation are necessarily unreasoned and unplanned.

2.4

Correctly interpreted classical rationality (rational decision theory) should say nothing about goals, motives, preferences of the agents. It should be just an empty shell, a merely formal or methodological device to decide the best[1] move, given a set of motives/preferences and their importance or order. Thus, being "rational" says nothing about being altruistic or not, being interested in capital (resources, money) or in art or in affects or in approval and reputation! Although everybody (especially economists and game theorists) will say that this is obvious and well known, in fact there is a systematic ambiguity and bias. By adopting a rational framework, we tacitly import a narrow theory of agent's motivation, i.e. the *Economic Rationality* which is (normative) rationality + economic motives (profit) and selfishness. Economists and game theorists are the first to be responsible for such a systematic misunderstanding[2]. The instrumentalist, merely formal approach to rationality should not be mixed up with the substantialist view of rationality: instrumentalist rationality ignores the specific motives or preferences of the agents. Thus *"utility" should not be conceived as a motive, a specific goal of the generic agent.* Utility is only an abstraction relative to the "mechanism" for choosing among the real motives or goals of the agent (Conte and Castelfranchi. 1995). I am puzzled also by the postulated mechanism or criterion for rational decision making. In particular I believe that also this mechanism is biased by the utilitarian philosophy and that there are several possible mechanisms more or less resource-driven or goal-driven (motivated rationality) (Castelfranchi and Conte, 1997). But this is not my point here. The point is that even by adopting the rational decision framework as it is *we can postulate any kind of motive/goal* we want or need in our agents: benevolence, group concern, altruism, and so on. This does not make them less rational, since rationality is defined *subjectively*! This might make them less efficient, less adaptive, less competitive, but not less subjectively rational.

2.5

Without arguing against such a presupposition, both cooperation and altruism (which are also made equivalent to one another) can only "emerge", can only be the result of selection, or of associative learning, they can just be due to emotions, or to learned rules and habits, or to social control and rewards, but never to reasoning and planning.

2.6

On the contrary, I believe that:

- On the one hand, the fact that genes are "egoist" says nothing about how the individuals (psychologically) are. Among the motives and the goals of the individuals there might be (either inborn or learned) some terminal, top-goals, some fully altruistic motives. So a (subjectively) rational agent can rationally *plan and deliberate for those unselfish goals*.

- On the other hand, agents can *deliberately cooperate* (in several ways: spontaneous and unilateral, under request, bilateral and by agreement, after a negotiation, etc.) both for altruistic motives and for selfish calculated advantages (instrumental goal-adoption). An important case deserves special attention: when the agents have and/or *believe they have common goals or common interests*.

2.7

Thus cooperation or altruism on the one side, and rational planned social interaction on the other side, are not two separate paths to sociality. There are altruistic impulsive or conditioned behaviours (goals) and egoistic impulsive or conditioned behaviours (goals), and there are rational plans and deliberated actions for altruistic goals, and rational plans for selfish goals. Cooperation is a large set of different kinds of behaviour partially overlapping with both: with the impulsive or learned and with the reasoned; with the altruistic and with the selfish.

2.8

Also the equivalence between "social order" and cooperation is somehow troublesome. To me *social order* is 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.

2.9

Analogously, I do not see why a strong relation between cooperation and altruism should be presupposed. Again a game theoretic unjustified assumption is allowed: the idea that cooperation necessarily exposes to free riding and pays a social cost, a contribution to society. It seems to me that

> [in Game Theory] a positive social action ("Cooperation") is an action which leads the agent to sustaining an intrinsically *social cost*. In other words, there is no cooperation without a penalty for the cooperative agent. This implicit conceptualisation of social action ties it up to the paradigm of bargaining. It characterises sociality as a costly and dangerous move, in which agents punish and reward each other at the same time, in which there can be no benefit without costs, in which agents face each other each trying to get away scot free. It is a view of sociality as a necessary evil, where agents are fundamentally opponents (as game-theorists indeed define them). Such a view is deeply engrained in the utilitarian philosophy. Actually, it is the only possible view of sociality if one takes a fundamentally utilitarian, that is to say formal in the sense previously defined, notion of interdependence. But it is by no means the only possible view of sociality. *Why should a "cooperative" move be by definition less convenient than other moves? why it should necessarily have an additional "social" cost?* (Castelfranchi and Conte, 1997)

Kinds and Levels of Cooperation: The Required Theory of Goal-Adoption

3.1

Not only it seems to me that cooperation shouldn't be identified with altruism or with social order, but, more importantly, I believe that cooperation is not an unitary notion or phenomenon. There are several different forms of cooperation, based on different mechanisms or motives, with different geneses, and explanations.

3.2

For a good theory of cooperation we need a subtler analysis of these forms (see e.g. Tuomela and Bonnevier-Tuomela, 1997; Conte et al, 1991). The broad and basic notion to be modelled is the notion of "social goal-adoption": the forms and reasons *why an agent decides to or acts in order to favour or realise a goal of another agent:*why it uses its own skills and resources for another's goal (adopting the goal of another). Notice that this is not necessarily an altruistic behaviour. There are **terminal** forms of Goal-Adoption where the adoption of the goal of the other is not instrumental to other internally represented goals, to further calculated advantages (like in pity, altruism, love, friendship, etc.). But *Goal-Adoption can be also* ***instrumental*** *to the achievement of selfish goals.*For example, feeding chickens (satisfying their need for food) is a means for eventually eating them. Instrumental Goal-Adoption also occurs in social exchange (reciprocal conditional Goal-Adoption).

3.3

Another motive-based type of Goal-Adoption (which might also be considered a sub-type of the Instrumental one) is strictly **cooperative** Goal-Adoption: *y* adopts *x*'s goal because he is co-interested in (some of) *x*'s intended results: they have a common goal.

3.4

The distinction between these three forms of Goal-Adoption is very important, since their different motivational bases allow important predictions on *y*'s "cooperative" behavior. For example, if *y* is a rational agent, in *social exchange* he should try to cheat, not reciprocating x's adoption. On the contrary, in cooperative adoption *y* is not interested in free riding since he has the same goal as *x* and they are *mutually dependent* on each other as for this goal *p*: both *x*'s and *y*'s action are necessary for *p*, so *y*'s damaging *x*would damage himself.

Modelling Emergent And Unaware Cooperation Among Intentional Agents

4.1

Although reasoning rationally does not equal being selfish, however, Macy is right when claiming that *social cooperation does not need agents' understanding, agreement, contracts, rational planning, collective decisions*. There are forms of cooperation that are deliberated and contractual (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 (Castelfranchi, 1997; Castelfranchi and Conte, 1992) is that it is important to model them not just among sub-cognitive agents[3] (using learning or selection of simple rules) (Steels, 1980; Mataric, 1992), but also among cognitive and planning agents[4] whose behaviour is regulated by anticipatory representations (the "future"). Also *these agents cannot understand, predict, and dominate all the global and compound effects of their actions at the collective level*. Some of these effects are self-reinforcing and self-organising.

4.2

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 and let some social unintended function emerge from this layer, and let the feedback of the unintended reinforcing effects operate on this layer (van Parijs, 1982). The real issue is precisely the fact 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 can one 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? (Elster, 1982). This is the real theoretical challenge for reconciling emergence and cognition, intentional behavior and social functions, planning agents and unaware cooperation. At the last SimSoc workshop in Cortona (Castelfranchi, 1997) I claimed that only agent based social simulation together 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.

4.3

I suggested that we need more complex form 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.*

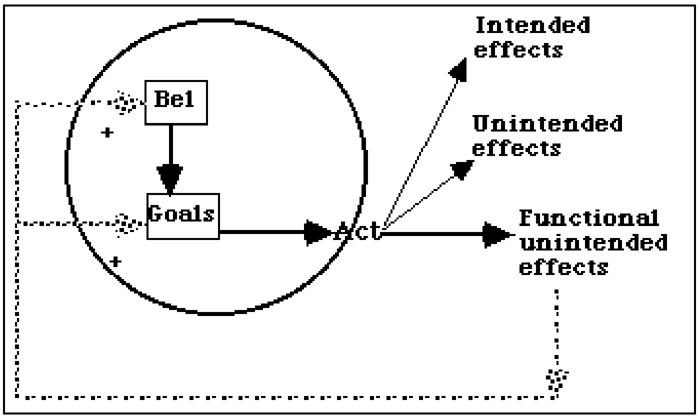
4.4

My claim was precisely that "the consequences of the action, which may or may not have been consciously anticipated, then modify the probability that the action will be repeated next time the input conditions are met", in agreement with Macy. In fact my sketched model (see Figure 1) was:

> 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. Then:

- First, behavior is goal-directed and reasons-based; i.e. is intentional action. The agent bases its goal-adoption, its preferences and decisions, and its actions on its Beliefs (this is the definition of "cognitive agents").
- Second, there is some effect of those actions that is unknown or at least unintended by the agent.
- Third, there is circular causality: a feedback loop from those unintended effects to increment, reinforce the Beliefs or the Goals that generated those actions.
- Fourth, this "reinforcement" increases the probability that in similar circumstances (activating the same Beliefs and Goals) the agent will produce the same behavior, then "reproducing" those effects.
- Fifth, at this point such effects are no longer "accidental" or unimportant: although remaining unintended they are teleonomically produced (Conte and Castelfranchi, 1995, ch.8): *that behavior exists (also) thanks to its unintended effects; it was selected by these effects, and it is functional to them.*Even if these effects could be negative for the goals or the interested of (some of) the involved agents, their behavior is "goal-oriented" to these effects.



4.5

I agree also with Macy's second point about "the probability that the associated rule will be replicated and diffuse". My example was in fact about the diffusion of behaviours like giving precedence at road crossings or leaving garbage around. I only disagree about *the role of cognitive representations* (beliefs and goals) in both reinforcement and reproduction, and diffusion. The problem is what kinds of mental mechanism the generic notion of "rule" can precisely cover[5].

4.6

So my position about the main claim of Macy's contribution ("Cooperation emerges not from the shadow of the future but from the lessons of the past") is as follow:

Cooperation emerges from the reflex of the past in the mirror of the future. The action remains anticipatory and goal-directed, but is influenced by (possibly not understood) lessons of the past.

Notes

[1] In Simon's more realistic view a "satisfying" move is enough.

[2] Just a couple of examples: consider, in economics, the literature on the "irrational" bias due to sunk costs (for example, Bazerman, 1990); in game theoretic approaches to the social sciences

consider Olson's claim about the irrationality of participation in social movements (Olson, 1965). In general one should remind Keynes' criticisms to the economists as "bentelhamist".

[3] 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.

[4] 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.

[5] Of course one might consider the relation between beliefs and goals generating and controlling behaviours as just a complex and flexible form of condition/action rule: where the condition is a configuration of beliefs to be checked up, and the action is a configuration of activated goals to be possibly planned. I don't think this is wrong, I just claim that cognitive representation must be analytically modelled to understand cooperation.

References

BAZERMAN, M. 1990. *Judgement in Managerial Decision Making*. John Wiley and Sons. 1990.

CASTELFRANCHI, C. 1997. Challenges for agent-based social simulation. The theory of social functions. IP-CNR, TR. Sett.97; invited talk at *SimSoc'97*, Cortona, Italy

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

CASTELFRANCHI, C. and Conte, R., 1997. Limts of Strategic Rationality for Agents and M-A Systems. In A. Cesta & P.Y. Shobbens (eds.) *Proceedings of the 4th ModelAge Workshop on "Formal Models of Agents"*,pp. 59-70

CONTE R. and Castelfranchi C. 1995. *Cognitive and Social Action*, UCL Press, London.

CONTE, R., Miceli, M. and Castelfranchi, C., 1991. Limits and Levels of Cooperation. Disentangling various tvpes of prosocial interaction. In Y. Demazeau & J. P. Muller (Eds.) *Decentralized AI - 2*, Amsterdam, Elsevier.

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

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

OLSON, M. 1965. *The Logic of Collective Action.* Cambridge, Mass. Harvard University Press.

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

TUOMELA, R. and Bonnevier-Tuomela, M. 1997. From social imitation to teamwork. In G. Holmstrom-Hintikka and R. Tuomela (eds.) *Contemporary Action Theory*, Kluwer, Dordrecht, Vol. II, 1-47.

VAN PARIJS , P. 1982. Functionalist marxism rehabilited. A comment to Elster. *Theory and Society,*11, 497-511.

Return to Contents of this issue