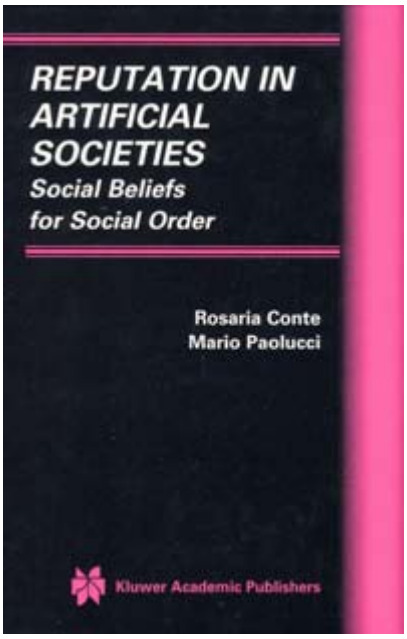
Reputation in Artificial Societies: Social Beliefs for Social Order

Conte, Rosaria and Paolucci, Mario
Kluwer Academic Publishers: Dordrecht, 2002
ISBN 1402071868

Order this book

Reviewed by **Flaminio Squazzoni**
Department of Social Sciences, Faculty of Economics, University of Brescia, Italy



An Initial Outlook: Moving into the "Structural Holes" of Disciplines

This book deals with reputation as a socio-cognitive mechanism able to strengthen collective action and promote social order, both in social systems and in artificial systems, such as infosocieties, e-institutions and online communities. In the book, the reader can find a thorough survey on reputation in many disciplines, a sound socio-cognitive theory of reputation, and some agent-based simulations that allow one to appreciate the theory put forward by the authors.

The subject of the book is intrinsically a transdisciplinary one and the possible applications of the theory described refer to many different fields. The transdisciplinarity suggested by the authors rests upon a "problem oriented" approach and a "process-oriented" method that allow them to overcome the unfruitful "turf wars" that are imposed by the division of labour among disciplines and sciences. The authors move with full awareness into the "structural holes" between different disciplines. They pass from evolutionary game theory to sociobiology, from cognitive science to economics, from political sciences to artificial intelligence, in search of a way to answer the question "what is reputation?" and to build a formal theory that can usefully be translated into agent-based models. From a theoretical and methodological point of view, this breadth is one of the main positive features of the book.

Image and Reputation

It is undeniable that reputation, far from being a "frivolous" issue, is an increasingly recognised subject in different disciplines, such as sociology, sociobiology, cognitive science, political science, economics, business, evolutionary game theory and so forth. This increasing interest comes from the fact that reputation and other related social mechanisms (such as trust, reciprocity, altruism, and mutual monitoring) are understood as an appropriate framework to institutionalise the smooth running of complex and decentralised societies. In fact, it is generally recognised that trust and reputation are essential conditions for reciprocity, and consequently for co-operation and collective action, even more so in social settings where institutions, social monitoring and control are distributed.

The reputation theory developed in the book is founded upon a synthesis of cognitive, social and institutional aspects. The core of the theory is the difference between "image" and "reputation" and the intersection between cognitive mechanisms and social transmission infrastructures. The authors stress that reputation cannot be understood as a "static attribute, rigidly codified as footprints of social hierarchy". On the contrary,

it has dynamic properties, because reputation attribution is a socio-cognitive mechanism that takes root in communication processes. In this respect, what the authors rightly emphasise is that one of the main limitations of the game theoretic approach is that it focuses on the "reputed agent" rather than on the "reputing agent". This therefore means that mechanisms of social transmission for "cognitive evaluations" are completely ignored, the complexity of interaction structures is reduced to a simple dyadic relation and reputation is viewed in terms of "effects" rather that being considered in terms of "processes". By contrast, if these mechanisms are taken into account, several intriguing consequences can emerge. This is basically what the book aims to do and show.

The reputation theory suggested by the authors is fruitfully translated into and formalised in agent-based models. Agent-based simulations are used as a theory-building tool. The validation of theoretical model micro-foundations is achieved through a comparison between different (and gradually more sophisticated) simulation settings. This is done without accepting the traditional approach to "simplicity" and to "unidirectional bottom-up" emergence that usually dominates the use of agent-based models in social science. As the authors rightly say, "good theories are not necessarily simple" but on this issue, see also Gilbert (1996).

A Summary of the Book

The book is composed of four parts. The first one focuses on the state of the art, with an analysis of the theory of co-operation as typically presented by evolutionary game theorists. The second part focuses on the problem of reputation transmission, one of the main interaction mechanisms that game theorists have failed to consider, and one that needs to be taken into account to explain the emergence of reputation-based social order. The third part focuses on several related mechanisms, such as altruism and reciprocity, which need to be reconsidered in a socio-cognitive perspective to support the theory of reputation described in the previous part. The fourth and closing section summarises the advantages, unresolved problems and possible applications of the reputation theory suggested by the authors, with a particular emphasis on applications in infosocieties and online communities.

Part I: "The State of the Art"

The first chapter permits the reader to acquire a panoptic view on theories and approaches to reputation as they have been developed over time in different fields and disciplines. As the authors point out, the first impression is that reputation has received growing attention in current scientific investigation (mostly in game theory and in the social sciences) but that many aspects of the phenomenon have been treated in an inadequate way. For example, the role of reputation transmission, both in direct and mediated interactions among social agents, has mostly gone unrecognised by game theorists, economists and sociobiologists. Such an impression of unbalanced development becomes stronger in the second chapter, where a critical analysis of the theory of co-operation and collective action in game theory and experimental psychology is presented.

One of the main pieces of evidence against the traditional view is that, as several experimental findings about social and collective dilemmas have highlighted, co-operation among agents is higher than game theory would suggest. Regarding this point, the authors draw on the well-known model of collective action suggested by Ostrom (1998). This model emphasises the role of reciprocity, reputation and trust as fundamental building blocks for a theory of co-operation. Even so, the Ostrom model fails to take into account important features of reciprocity and it does not produce a sound theory of the interaction between the building blocks of collective action it proposes. The third chapter focuses on the problem of the relevance of repeated interaction for the emergence of co-operation among agents, as it is laid out in game theory and in experimental psychology. The survey shows that the problem of co-operation does not find strong solutions within game theory. The authors show how game theory indicates that there is no rational basis for co-operation in the one-shot encounter. Furthermore, in repeated encounters, a number of factors may favour co-operation over mutual defection, according to uncertainty about the future (the famous "shadow of the future"), about the rationality of the co-player, and about the possible influence of social embeddedness. But, as different experiments have demonstrated, there is strong evidence that co-operation emerges in one-shot interaction, too. The chapter ends with some open questions (which are mostly related to the theory of embeddedness of social action) as a way of bringing social beliefs, social groups and reputation transmission into the traditional game theoretical framework.

Part II: "Reputation Transmission"

The second part of the book opens with a chapter in which the authors suggest an alternative perspective on reputation compared to that of game theory, namely the "reputing agent" perspective. The shift from "reputed agent" to "reputing agent" allows us to concentrate on social mechanisms of reputation transmission and to focus on social structures that contextualise agents' behaviour. Moreover, such a shift allows us to complicate the traditional view presented by game theory and experimental psychology, introducing some realistic cognitive and social mechanisms into the theory in the process.

The first chapter of this part consists of a description of a socio-cognitive model of reputation, in which reputation is understood in terms of the output of a social process of information transmission that works on an input that is called the "image of the agent". In this way, the authors emphasise an analytical difference between "image", conceived as a set of evaluative beliefs about a given target, and "reputation", conceived as the process and effect of transmission of the image. Both are "social mechanisms", because they concern properties of another agent (a "presumed attitude towards social desirable behaviour") and they may be shared by a multitude of agents.

Image is an evaluative belief, and it is thought to constitute a hybrid mechanism, because it concerns both a belief about a given entity ("it is good for, or can achieve, a given goal") and a social evaluation when the belief "concerns another agent as a means for achieving this goal". Essentially, reputation is a "meta belief", a "belief about other minds", without reference to the acceptance of nested beliefs and with agents that are embedded in a nexus of contemporary roles they can play (such as "evaluator, beneficiary, target, and third party".)

The decisions of agents are differentiated into two levels, namely the "epistemic level", where the decision is grounded in both image and reputation and focuses on "whether to accept a given belief" and the "pragmatic-strategic level", where the decision of an agent is to affect others' decisions, according to the mechanism of reputation transmission. This last mechanism raises the problem of "mimetic decisions" that is discussed in the third chapter of the second part. Finally, the authors describe different reputation dynamics. One is the propagation of social cognitive representations from one agent to another, that is to say what they call "transmission of reputation" (or "gossip"). Another is the extension of a given agent's reputation to other agents, that is to say what they call "contagion of reputation" (or "prejudice"), which works through socio-cognitive "typifications" (friend/enemy), allowing partner selection. The book and simulation models developed by the authors focus on the former dynamic mechanism, while the latter class of mechanisms will be investigated in further studies.

The second chapter in this part describes the results of a simulation model that the authors have developed to understand the effect of reputation in a population of agents composed of norm abiders and utilitarian agents and the reproductive benefit of normative behaviour. Without going into detail about the model, it is worth remarking that it is an "abstraction" in the sense that it allows us to explore a metaphor for a social arena in which agents interact. Agents move into a two-dimensional space (with a toroidal grid) and there is randomly scattered food. Fitness conforms to the following rule: each agent has a fitness value that grows when the agent eats and decreases when it moves, attacks or is attacked by others. Food units are associated with the fixed amount of fitness they provide when eaten. At the beginning of the simulation, agents and foods are randomly assigned their spatial locations.

The norm is viewed as a restriction on aggression, while the authors introduce three possible strategies that, in the first experimental setting, are associated with different types of agents: "blind aggression" (no constraints over aggression), "utilitarian or strategic aggression", where aggression is constrained by strategic reasoning (agents attack only those agents whose strength is not higher than their own), "normative strategies" (introducing a sort of "moral right" about food, with possession conforming to a rule of ascription, so that attacks are forbidden when agents are eating their own food). Essentially, blind agents attack anybody, utilitarian agents attack only the weaker and normative agents do not attack other agents who are eating their own food.

Different experimental settings are created. The first one is based on the comparison between different agents, while the second is based on the introduction of a mixed population. The simulation results of the

second simulation setting show a classic outcome: normative agents have several disadvantages with respect to utilitarian agents, because of the costs associated with normative action and the lack of sanction supporting infrastructures. In fact, normative agents sustain all the costs of aggression control, without the possibility of identifying and ratify the status of utilitarian agents.

To remove these limitations, the authors further complicate the model by introducing retaliation strategies that allow normative agents to identify and ratify utilitarian agents, something like the "image" mentioned above. The retaliation strategies are as follows: "absolute" (enforcing the norm with any agent), "quasi-rational" (attacking any utilitarian agent independent of its strength) and "rational" (behaving as a utilitarian agent with other agents of that kind). To implement these strategies, normative agents have to have access to information about the behaviour of other agents. This information has a binary structure (friend/enemy), according to prior compliance with the "moral right" norm mentioned above. A "reputation vector" is introduced, including a rule of "presumed innocence". (On principle, agents are supposed to be naturally norm abiding until their actions show the contrary.)

The results of this simulation setting (a mixed population equally composed of normative and utilitarian agents) shows a slight increase in utilitarian agent average strength and a slight decrease in the average strength of the normative agents. That is to say a more negative picture for norm adherence than that emerging from previous simulation settings. The conclusion is that "image" constitutes an insufficient social infrastructure to sustain normative behaviour. Because the simulation shows a profound loss of information about the utilitarian agents for normative agents, the authors point out that normative agents need to make use of something more than their personal experience in direct interactions.

This is the point where reputation matters. In a new simulation setting, normative agents have the capacity to exchange information about utilitarian agents. The simulation results show that reputation serves as a mechanism of cost redistribution for the normative population. Normative agents simultaneously punish the utilitarian agents and exercise an indirect influence on them, boosting the level of norm compliance.

Finally, in additional simulation settings, the authors focus on genetic mechanisms to study the reproductive advantage of a normative disposition by introducing inheritance of parents' knowledge about reputations and aggression routines. The results show that normative strategies are not evolutionarily stable when there is no transmission of information across generations. However, when inheritance of knowledge about normative reputation is allowed, then the reproduction rate of normative agents is greater than that of utilitarian agents.

The third chapter focuses on the description of a mimetic model of reputation transmission that allows us to answer several questions: whether and why reputation information is transmitted; what are the mechanisms and reasons that explain the decision to transmit; to whom is information transmitted; whom does the information concern and how is it conveyed, that is to say what are the channels and the contextual properties that facilitate reputation transmission? Through the mimetic model, three important dynamics are observed, namely the likelihood of "gossip", "cynicism", and "leniency" in a population of agents.

Part III: "What Reputation is Good For"

The first chapter of this section focuses on an ambitious and challenging question, namely what is the evolutionary rationale of reputation. Firstly, the authors present a survey on evolutionary theories of altruism and reciprocity, mainly from sociobiology. The authors describe the types of reciprocity that the existing literature focuses on (direct reciprocity in small groups with high-density populations and indirect reciprocity in large societies with low-density populations) and the recognised paths in the evolution of reciprocity ("close kin advantage", "punishment", "retribution" and "group benefit"). The result of the survey is to show that the evolutionary sciences do not pay attention to cognitive aspects in the evolution of social behaviour. By means of an analysis of cognitive approaches to the so-called "adaptive mind", the authors suggest that sociobiological theories (as well as cognitive science theories) fail to take into account the problem of the co-evolution of mental and social structures, and the importance of "social cognitive artifacts" such as institutions.

The second chapter focuses on some possible functions of reputation transmission as a "social cognitive artifact" in the evolution of reciprocal altruism, for example, in contexts of low probability repeated interaction. The authors suggest an interesting analysis of "gossip" as a source of informational reciprocal

altruism. "Gossip" is viewed as a tool to enlarge reputation information. Reputation is viewed as a "secondary effect of reciprocal altruism", because it allows trustworthy partners to be selected and their identities to be kept hidden from utilitarian agents and because agents have a higher probability of survival and reproduction if they provide material and informational help to one another.

The authors identify two mechanisms that give rise to informational reciprocal altruism: the first one is that material help costs are usually higher than communication costs, and, consequently, that the incentive to defect is stronger at the material than at the informational level; the second one is that the power to provide material help is less frequent that the power to information, so that informational altruism is more likely to occur that material altruism. The authors present a model of the co-evolution of material and informational reciprocity and direct and indirect reciprocity. Without going into the details of this point, it is worth observing that gossip and reputation cannot be understood in terms of "mirror strategies" in a traditional game theoretical framework. Gossip and reputation account for prescriptive or moralistic retaliation as well as providing a tool which distributes the retaliation costs socially among agents.

The final chapter focuses on a simulation model of "false reputation", where the new results conform to what might be expected in the mimetic model mentioned above. The simulation models previously described assume that normative agents keep a record of utilitarian agents (what the authors call the "image"), retaliating against them in later encounters and that they exchange information about the reputations of others (what the authors call "reputation transmission"). In the simulation settings previously described, information is not complete but it is correct while the "false reputation" model allows for the introduction of errors and bluff. In this case, normative agents keep a record of both social categories (normative and utilitarian agents). When a stronger neighbour refrains from attacking a particular agent, the former might be recorded as a normative "type", even if it was actually a utilitarian that decided, according to self-interested rationality, not to attack. This change to the assumptions of the model introduces many possibilities for incorrect information, acts to bias information about "image" and affects reputation transmission. In fact, reputation transmission works through updating the list of utilitarian agents in the memories of normative agents. Normative agents accept information only from those they believe to be normative. They reject it when it comes from "reputed" utilitarian agents. Once an "image" is accepted, it will be used to update the list, even with no information about the target. In the case of contradictory information from two or more neighbours, one of them is randomly chosen. Finally, the simulation introduces some "copying errors" in the updating of the list.

The effect of noise introduced at cognitive ("image") and social ("reputation") level is to bring about two different types of social bias: what the authors call the "inclusive error" ("social optimism" or "leniency"), that is to say the case of "false good reputation" assigned to utilitarian agents, and the "exclusive error" ("calumny" or "social cynicism"), that is to say the case of a "false bad reputation". Simulation results show that "false good reputation" ("inclusive error") allows a relevant advantage for utilitarian agents, while "false bad reputation" is not so disadvantageous for normative agents. This last outcome is interesting, suggesting what the authors call "the asymmetry between calumny and leniency". This indicates an empirically plausible prudential algorithm: "spread news about others' bad reputation even if it is uncertain, since calumny is preferable to no reputation transmission" and "do not spread news about good reputation unless it is certain, since no reputation transmission is preferable to leniency". In sum, the authors point out that simulation results conform to the following rule: "social cynicism is apparently less dangerous that both social optimism and, more importantly, silence!"

Part IV: "Advantages of the Present Approach"

The first chapter of this part focuses on the social impact of reputation, with particular attention given to the transfer of knowledge to policymaking and social and institutional monitoring. The second chapter focuses on reputation in infosocieties, online communities like eBay, Sporas (Zacharia and Maes 2000) and Histos (Zacharia and Maes 2000). The last two are non-commercial research systems that have been built to improve the realism of reputation brokering systems and also for development of MAS applications. What matters is that these systems demonstrate the inconsistency of the pessimistic game theoretic predictions about co-operation and reciprocity, but that they do not exploit all the interesting implications of a reputation theory-based system. The conclusion is that "in a world where legal sanctions are hardly applicable, social order can but depend upon immaterial, symbolic sanctions", and thus that reputation should be viewed as an

ICT supporting institution. Unfortunately, the reader cannot avoid the impression that this chapter on ICT applications is not as well developed as the others and that further investigation of this topic is clearly needed.

Open Questions and Some Considerations from the Perspective of the Reader

The book is absolutely recommendable, both because of the issues it addresses, the methods the authors suggest and the simulation results that are presented. The potential audience is composed of social scientists (even those with no background in agent-based simulation and artificial societies), game theorists (who might be a bit disappointed with respect to the way game theory is usually practised), policy makers (and experts) in infosocieties and businessmen and managers with an interest in corporate reputation, not in search of some magic formula, but hoping for deep and constructive theoretical stimuli.

Clearly, the multidisciplinary aspects of this project and the lack of a sound theory of reputation in the social sciences have obliged the authors to launch into an extensive survey of various literatures that could daze the less keen reader. For example, frequently the substance of theoretical reflections is let drop and revived in a later chapter. To follow the discourse in a cumulative way, the reader is often obliged to jump from page to page and from chapter to chapter. This is the reason why the reader has to be willing to accept an "active bricolage" method of reading. To the great relief of the reader, the authors have provided a very detailed index and extremely useful recapitulation sections at the end of every chapter. This should overcome the negative reaction of some readers, faced with a book that has sound appeal but is intrinsically difficult in its content. In this respect, the reviewer feels he can remind readers that really challenging social science is always hard to apprehend and that the book is well worth the investment of effort.

The book raises a lot of unresolved questions but this is what a good book has to do! I will try to summarise these questions briefly, avoiding further time spent on praising the book and (generally speaking) the clearly productive research program on agent-based social simulation that the authors (together with other members of CNR in Italy) are undoubtedly going to continue to pursue for many years (Conte and Castelfranchi 1996, Castelfranchi 2000).

The first question concerns the relation between models of social phenomena and models of artificial societies (such as infosocieties). I would argue that there are relevant differences between the two. These differences need to be more carefully taken into account from a social science point of view. The second problem concerns the validation of agent-based simulation results.

Concerning the first issue, right from the outset, the book feeds an interplay between two different (even if quite interrelated) levels of analysis and application: reputation in social systems and reputation in engineered social systems, such as infosocieties. While the interest in infosocieties and online communities as social laboratories is unsurprising, in view of the fact that designers of artificial systems face difficult challenges in dealing with the problems of social order and collective action, it is impossible to deny a fundamental difference between social systems and infosocieties.

Infosocieties are engineered social artifacts with an identifiable optimisation goal and a clear functional strategy. The desirable positive behaviour in infosocieties is a function of the optimisation that is the goal of the system. In social reality, the desirable positive behaviour is non-transparent and is exposed to continuous negotiation, while the optimisation goal is unidentifiable. The system does not have a clear and unique goal, and the *emergence* of a norm, rather than mere compliance is the more interesting mechanism to focus on. Moreover, in social reality, norms can equally be a source of social *disorder*.

Such an interplay permeates the structure of the simulation models presented in the book. For example, in social situations (by contrast to the models presented), normative agents can form a clan, based on trust, reciprocity, altruism, reputation (and so on) and still intentionally pursue wicked goals. On the other hand, what the authors call "cheaters" or "utilitarian" agents can be a source of new (and ultimately desirable) normative regimes. Desirable positive behaviour (enforced by trust, reputation, altruism, and monitoring) can lead to inefficient or even disastrous collective results. Norms can be collectively inefficient, while one of the main forces leading to institutional change could be exactly the breaking of rules and examination of norm-abiding groups. In this context, for example, think about the individual who has allowed us to disrupt the norm-abiding Enron clan. Should they be thought of as a "cheater" or as a "utilitarian?"

In this respect, the literature on reputation and the models suggested in the book run the risk of producing the same opaqueness, difficulties and aporias that the literature on social capital has already encountered (Uzzi 1997, Portes 1998). To overcome such a risk, the authors need to take into account, in future developments of their work, a more complex theory of norms in social settings. This is to say that, from a social science point of view, the authors need to take deeper account of some interesting issues such as, for example, the role of reputation in the emergence of norms in social settings. Interesting phenomena arise where agents compete on the basis of alternative normative goals. The role of reputation mechanisms as agents of norm evolution and drivers of institutional change also require deeper consideration.

The second open question concerns the problem of validating the agent-based simulations presented in the book. This is a general question that bears on the use of agent-based models in social science.

First of all, the simulation models described in the book can be squarely classified as stark abstractions of social reality. They work as metaphors for generic social dilemmas but display no reference to a specific empirical reality or to a specific class of empirical phenomena. (For a brief note on the classification of agent-based models, see Boero, Castellani and Squazzoni 2004, sections 1.7 and 1.8.) The question is as follows: should the internal coherence of an abstraction and the comparison between simulation settings be thought as an effective validation of micro-foundations for agent-based models? Of course not.

Thanks to the theoretical quality of the models described in the book and to the precise formalisation of the reputation theory suggested by the authors, a desirable solution for the validation problem could be the use of experimental methods. This is one possible solution but perhaps not the only one. For example, the authors should set up experiments on their reputation theory using real humans in different contexts (infosocieties, firms, public social settings) or with real humans interacting with artificial agents and so on. Validation of simulation results should be taken up with assessing the micro-foundations of the reputation theory suggested by the authors. Probably, the authors have already thought of doing this. It would be a very desirable development of the research program described in this very stimulating book.

References

BOERO R., M. Castellani and F. Squazzoni 2004. Micro behavioural attitudes and macro technological adaptation in industrial districts: An agent-based prototype. *Journal of Artificial Societies and Social Simulation*, 7, <http://jasss.soc.surrey.ac.uk/7/2/1.html>.

CASTELFRANCHI C. 2000. Through the agents' mind: Cognitive mediators of social action. *Mind and Society*, 1:109-140.

CONTE R. and C. Castelfranchi 1996. Simulating multi-agent interdependencies: A two way approach to the micro-macro link. In K. G. Troitzsch, U. Mueller, N. Gilbert and J. Doran, editors,*Social Science Microsimulation*. Springer Verlag, Berlin.

GILBERT N. 1996. Holism, individualism and emergent properties: An approach from the perspective of simulation. In R. Hegselmann, U. Mueller and K. G. Troitzsch, editors,*Modelling and Simulation in the Social Sciences from the Philosophy of Sciences Point of View*. Kluwer Academic Publishers, Dordrecht.

OSTROM E. 1998. A behavioural approach to the rational choice theory of collective action. *American Political Science Review*, 92:1-22.

PORTES A. 1998. Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, 24:1-24.

UZZI B. 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42:35-67.

ZACHARIA G. and P. Maes 2000. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14:881-907.

Return to Contents of this issue