Università degli Studi di Padova

Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Scuola di Dottorato di Ricerca in Scienze Psicologiche
Indirizzo Scienze Cognitive
Ciclo XX

Cooperation Through Communication: Agent-Based Models and Experimental Results

Direttore della Scuola: Ch.mo Prof. Luciano Stegagno
Supervisore: Ch.mo Prof. Luigi Castelli

Dottorando: Gennaro Di Tosto

2 Febbraio 2009

Abstract

In human societies, reputation is a complex artefact. It relies on the tendency of group members to assess one another and its ultimate function is to perform distributed social control, i.e. the enforcement of social accepted norms and behaviours without the presence of a centralised institutional actor. This goal is accomplished by means of evaluations spreading (gossip) among the members of the same group. Hence, reputation is described as an information transmission process. The inputs of the process are the beliefs that group members autonomously acquire during social interactions, while its output is an emergent property for the evaluated agent, i.e. what an agent is believed to be (according to given social norms) as a result of the spreading of evaluations about him/her.

Since social actors interact in a non-linear way, Agent-Based Social Simulation (ABSS) is a candidate methodology for the study of gossip and reputation. By means of computer models, ABSS allows the design of the behaviour of artificial entities —the agents— and the local rules that govern their interactions. The simulated dynamic, then, allows to observe the properties that emerge at the level of the systems, which can be statistically analysed.

Simulations were implemented in scenarios that represent different forms of strategic interactions, allowing us to test how and to what extent circulation of information influence cooperative behavior. Two forms of communication mechanisms were tested: one with private information, the other with public information. Moreover, an additional case scenario is analysed: an Industrial district populated by artificial firms. Industrial districts (Ids) can be conceived as complex systems made of heterogeneous but strictly interrelated and complementary firms. One of the distinctive features of industrial districts is the tight connection existing between the social community and the firms: in this context, economic exchanges are mainly informed by social relationships and holding good reputation is an asset that may actually foster potential relations. In this work we modelled the effects of two kinds of social evaluations: namely Image (direct evaluations) and Reputation (reported evaluation). Likewise, we compared the effects of sincere and insincere information on the economic performances of the single firms and of the cluster as a whole.

In a different experimental settings —performed with groups of natural subjects interacting through a graphic computer interface— we analysed reciprocal forms of messages exchanges. The positive effects of communication on rates of cooperation is a robust experimental finding. When individuals can talk to one other, cooperation increases significantly. Proposed explanations to this phenomenon consider the formation of group identity, as well as the chance to make explicit commitments —where reputational and moral factors come into play— as fundamental causes. This research, however, looks at communication as a mean to establishing and enforcing cooperation among people; to our knowledge, no attempt has been made to analyze communication strategies, when communication processes are actually the place where cooperation evolve.

We developed a novel experimental setting in which participants playing a memory game (with numbers instead of images) could either play alone, or exchange messages containing the position and the value of the cards, so that those who received a truthful message could more easily get a match. This setting was conceptually modeled after the stag hunt game, a coordination game in which players do better if they coordinate their behavior with the behavior of others.

In our experiment, playing alone is faster than sending messages, but it leads to a quicker depletion of the available moves. On the contrary, sending messages is more time consuming, but it allows players to know the position of cards, provided the information is correct, without wasting moves in trying to guess.

Results show that the exchange of messages is a mutually beneficial activity, allowing participants to jointly discover the game board, to score higher and more efficiently. Cooperation through communication is conditional to receiving messages from other participants and is performed with this very expectation (as reported by the majority of the subjects afterwards). This strategic behavior could be explained according to two alternative frameworks: either a game-theoretical interpretation of reciprocity, analyzed as an imitation strategy (Tit-For-Tat); or a cognitive view in which cooperative behavior is regarded as a socially prescribed activity, and every deviation from the norm is punished according to the interpretation of the violation. The absence of retaliatory behaviour; and a tendency to exclude non-cooperative partners from further communication seems to exclude the possibility of an imitation strategy.

Abstract (Italiano)

Nelle società umane, la reputazione è un artefatto complesso. Essa si basa sulla tendenza dei membri di un gruppo di valutarsi reciprocamente e la sua funzione è quella di eseguire un controllo sociale distribuito, vale a dire l'applicazione ed il rispetto dei comportamenti e delle norme sociali accettate, senza la presenza di un attore istituzionale centralizzato. Questo obiettivo si realizza mediante la diffusione delle valutazioni (*gossip*) tra i membri dello stesso gruppo. Di conseguenza, la reputazione è descritta come un processo di trasmissione delle informazioni. Gli input del processo sono le credenze che i membri del gruppo acquisiscono autonomamente durante le interazioni sociali, mentre il suo output è una proprietà emergente per l' agente target delle valutazioni; vale a dire ciò che un agente è creduto essere (in base a norme sociali) a seguito della diffusione di valutazioni su di lui/lei.

Dal momento che gli attori sociali interagiscono tra loro in modo non-lineare, la Simulazione Sociale Basata su Agenti (ABSS) è stata scelta come metodologia per lo studio del gossip e della reputazione. Per mezzo di modelli computazionali, ABSS permette il design e l'analisi del comportamento di entità artificiali (gli agenti) e le norme locali che regolano le loro interazioni. Le dinamiche simulate, quindi, permettono di osservare le propriet emergenti a livello del sistema, che possono essere analizzate statisticamente.

Le simulazioni sono state implementate in scenari che rappresentano diverse forme di interazione strategica, e ci hanno permesso di testare come, e in quale misura, la circolazione di informazioni influenza il comportamento cooperativo. Due forme di comunicazione sono state testati: una con informazioni private, l'altra con informazioni pubbliche. Inoltre, un ulteriore scenario è analizzato: un distretto industriale popolato da agenti artificiali. Una delle caratteristiche distintive dei distretti industriali è il legame stretto esistente tra le imprese ed il loro contesto sociale: in questo caso, gli scambi economici sono principalmente informati da relazioni sociali e la buona reputazione di un'azienda è un bene che può effettivamente favorire potenziali relazioni. In questo lavoro abbiamo modellato gli effetti di due tipi di valutazioni sociali: vale a dire "Immagine" (valutazioni dirette) e "Reputazione" (valutazioni riportate). Allo stesso tempo, abbiamo confrontato gli effetti di comunicazioni sincere e insincere sui risultati economici delle singole imprese e del sistema nel suo complesso.

Parallelamente è stato intrapreso un lavoro sperimentale, i cui soggetti interagivano tramite computer, volto ad analizzare forme cooperative di comunicazione durante lo svolgimento di un task assegnato. L'effetto positivo della cooperazione sul tasso di cooperazione è un risultato sperimentale oramai solido. Quando i soggetti hanno la possibilità di parlare tra di loro, i comportamenti cooperativi aumentano significativamente. Le spiegazioni proposte per questo fenomeno vanno ricercate nella formazione di identita' di gruppo, come anche nella possibilita' di formare impegni espliciti, dove i fattori reputazionali e morali giocano un ruolo fondamentale. A tutt'oggi però lo stato dell'arte non contempla un'analisi delle strategie di comunicazione, nel caso in cui siano i processi

comunicativi il luogo in cui la cooperazione ha luogo ed evolve.

Il disegno sperimentale prevede la possibilità per i partecipanti, ai quali è chiesto di svolgere un compito simile ad un memory game (con numeri invece di immagini), di giocare da soli, oppure di scambiare messaggi con altri partecipanti contenenti la posizione ed il valore delle carte. In questo modo, chi riceve messaggi veritieri è facilitato nel compito. I risultati mostrano che lo scambio di informazioni è un'attività mutualmente benefica, che consente ai partecipanti di esplorare congiuntamente lo spazio delle possibilità, e di ottenere punteggi più elevati. L'invio di messaggi è inoltre un comportamento condizionato alla ricezione di messaggi dagli altri partecipanti, ed è un'attività svolta con questa precisa aspettativa sul comportamento degli altri, come riportato dalla maggiornaza dei soggetti nel questionario finale. Questa strategia comportamentale è interpretabile come una norma di reciprocazione. L'assenza di punizione (*retaliatory behaviour*), congiuntamente alla tendenza ad escludere partners non cooperativi dal processo di comunicazione, ad indicare che le deviazioni dalla norma sono trattate in base alla soggetiva interpretazione dell'utilità della norma.

Acknowledgements

I wish to thank all the people that helped and supported me during these years:

Luigi Castelli, my supervisor, for his support.

Rosaria Conte and Mario Paolucci, of the Institute of Cognitive Sciences and Technologies (ISTC), Italian National Research Council (CNR), for their encouragement, their scientific advice, and for having welcomed me in their lab.

Francesca Giardini, for the work we have done together.

Giulia Andrighetto, Luca Tummolini, Luca Giachi, Giovanni Pezzulo, Laura Barca, and all the young researchers of the ISTC, for our stimulating discussions, and their friendship.

Antonietta di Salvatore, for helping me with the statistical analysis.

Daniele Denaro, for developing the software used during the laboratory experiment.

My parents and Ele, for everything else.

The work presented in this dissertation was partially supported by the Italian Ministry of University and Scientific Research (MIUR) under the FIRB programme (project *SOCRATE*, contract number RBNE03Y338), and by the European Commission under the Sixth Framework Programme (project *eRep: Social Knowledge for e-Governance*, contract number CIT5-028575).

Contents

Chapter 1

Introduction

1.1 Purpose of the study

Reputation is the basis for trust. Commercial and electronic-auction web sites know this, as they adopt systems to collect users' evaluations to rate items and sellers. They use reputation to generate new profits, too: through recommendation engines they spread evaluations towards new users on the basis of the similarity of their interests.

Software solutions on the internet and the new communication media closely resemble solutions evolved in human societies to cope with problems of social order, interaction opportunities, cooperation and trust. Interestingly, the concepts and the design tools used to study and implement the interaction technologies of the near future share some relevant functional features with those used by researchers to investigate about the very beginning of hu-

man civilization and culture. Social life in other primates is confined in small groups, mostly composed of genetically related individuals. Human societies, instead, developed in large groups, where a significant part of the interactions take place between unrelated, often never met before, partners. Hence the question: what are the features responsible for this change, and how do we maintain the level of social order that sets us apart from our phylogenetic ancestors?

The study of the evolution of cooperation has come a long way since its inception, enriched by the contribution of several scientists committed to the study of social behaviour, and today can be certainly described as a multidisciplinary endeavor. The analytical framework built along with the scientific investigation of cooperative phenomena links the level of evolutionary processes and the level of strategic thinking, the history of the adaptations to our environment and the cognitive machinery that implement them — i.e. ultimate and proximate causality; and led one of the advocates of this project to state that '[...] recent theoretical and empirical developments have created the conditions for rendering coherent the areas of overlap of the various behavioural sciences' (Gintis, 2007, p. 1).

Although this position is far from been accepted without objections, it is true that the models developed for the study of cooperation have been fruitfully adopted in psychology, anthropology, economics, sociology to gather data about human social decision-making, in order to account for its evolutionary origin, its social function and effects, and its cognitive and emotional make up.

Among the processes that can promote cooperation, reputation is a distinguished feature of human societies. Knowledge about reputation can help us to predict, at least partially or approximately, what kind of social interac-

tion to expect and how that interaction could evolve; or to avoid dangerous interaction all together, when partner's reputation is bad. Reputation can be described as the result of the tendency of group members to assess one another. Analytically, reputation has been shown to be an efficient and effective way to promote social desirable behaviours. It is efficient because the use of social assessments produced by third parties saves us the cost of acquisition of valuable information (i.e. can compensate the lacking of personal experience). It is effective because it adds a cost to deviant behaviours; or, if agents have the goal to preserve their reputation, concerns about others' positive assessments will encourage them to conform to social prescriptions and norms.

The effects of reputation have been used to analyse the conditions that can favor the evolution of cooperation in the case of indirect reciprocity: with social interactions taking place between a growing number of agents —which increases the probability of anonymous, one-shot interactions— reputation can help to preserve social order, overcoming the shortages of the classical models of cooperation based on direct reciprocity, which assume dyadic interactions between agents. The same principle has been extracted and applied to the study of competitive settings, like e-markets, and is very common in the design of artificial societies by means of multi-agent systems (MAS). In these contexts the functions of reputation are implemented through so-called "reporting systems" (see section 2.1), which offer a way to accumulate the knowledge of the agents/users using a centralised control artefact. But there is another way to generate and process reputational knowledge: the creation and transmission of assessments of social agents among peers, i.e. gossip, which in turn is a decentralised mechanism based on communication. These assessments, or evaluations, are described as pieces of information regarding

other agents whose attitudes, behaviours and actions are assessed with respect to some specific dimensions or aspects. The purpose of this study, then, is to *provide an operational model to investigate the effects of the spreading of social evaluations among agents involved in strategic interactions.* This will be done by introducing, and by explicitly modelling, communication processes in the experimental design.

1.2 Reputation and gossip

One of the most significant factor distinguishing humans and the other primates to whom we are phylogenetically linked, is the extent of our reliance on culture: our ability to exist in nearly every ecosystem on the planet is primarily due to our capacity to acquire, employ, and elaborate on socially transmitted information (Henirch, 2004).

The use of socially transmitted information is an efficient way to cope with the environment. It allows one to build on the experience of others and to aggregate information from the behaviour of many individuals (Boyd and Richerson, 1985, chapter 7). Henrich and Gil-White (2001) argue that humans are imitators by default, or, as the authors put it, "default infocopiers": first they try to learn from others, then they improve through individual learning.

In an information-costly environment, the sources of information are selected from among the best ranking individuals: sources are ranked according to their skills, and infocopiers pay deference to them in order to buy proximity, which in turn improves copying reliability and fidelity. This mechanism leads to the emergence of prestige-biased variations. Prestige is defined as 'noncoerced, interindividual, within-group, human status asymmetries' (Henrich and Gil-White, 2001). In non human societies, status is almost always associated with force or force threat. The resulting social asymmetries are called dom-

inance hierarchies. Prestige is different from dominance because no force or force threat is involved, instead it is assigned to people who excel in valued domains. And since deference is very difficult to fake, it represents an honest signal, making skill ranking and prestige-biased imitation highly effective.

Thus, humans are always seeking information about their environment and constantly evaluating others as sources for knowledge. But a good part of this information refers to the social environment itself: the people with whom we inhabit it and the social norms that regulate it. And the knowledge we derive tells us how these people stand under the social norms in place, i.e. their *reputation.*

1.2.1 The functions of reputation

Reputation is usually considered to be a mechanism for enhancing trust and cooperation among strangers. And has been formalized by assuming that individuals who are "in good standing" in the community cooperate with others who are in good standing in the community. If an individual fails to cooperate with someone who is in good standing, he falls into "bad standing," and individuals in good standing will not cooperate with him. This type of situation is called *indirect reciprocity* (Alexander, 1987).

Theoretical models show that cooperation in sizable groups can be maintained if potential partners have information about a person's past behaviour and use it in making decision about interaction Panchanathan and Boyd (2004). This has been confirmed by experiments with economic games: alternating rounds of indirect reciprocity with rounds of a public good game (a setting in which individuals are asked to make contribution to a common pool, and subsequently share the benefits equally, regardless of the amount contributed), subjects contribute to the public good in order to acquire a prosocial

reputation for the indirect reciprocity rounds. However, rates of cooperation declines when rounds of indirect reciprocity are not expected anymore (Milinski et al., 2001), or when interactions become anonymous (Semmann et al., 2004).

Recent studies have also shown how concerns about reputation can reduce the opportunity for —or increase the cost of— opportunistic behaviour even when repeated interactions and identity disclosure are not an issue. Sometimes a simple cue of being watched (like eye-like spots) is sufficient to motivate people to being generous Haley and Fessler (2005); Bateson et al. (2006).

1.2.2 Gossip as a building mechanism for reputation

Gossiping, i.e. transmitting information and evaluations regarding a (usually absent) third party, is crucial for human societies, serving many different functions. The fact that gossip is an interesting phenomenon seems to be hardly questionable, although this argument has received little attention from researchers in psychology and cognitive science, compared to other social phenomena. This could be partially due to the difficulty of reducing this spontaneous and blurred flow of information to a controllable variable with predictable effects. Another motive for the paucity of interest raised by gossip may be the moral condemn this activity has received over the centuries, and by different societies around the globe.

However, a large and diverse group of scholars has given us detailed accounts of the functions of —and reasons for— gossiping. Gossip is a valuable source of information about the community, its members, its norms, values and habits, but it is also useful for mapping the social environment and for making inoffensive comparisons. According to the evolutionary perspective on gossip put forward by Dunbar, ‘gossip is the central plank on which human

sociality is founded. In reality, the cognitive demands of gossip are the very reason why such large brains evolved in the human lineage' (Dunbar, 2004, p. 109). Other authors emphasise the impact of gossip on crucial aspect of social life. If Coleman (1990) views the function of gossip as the creation and maintenance of norms, Elias (1974) considered this to be only one role of gossip; the other concerns social cohesion. In Elias's view, a form of collective social control is achieved by means of "blame gossip" used to sanction deviant in-groups. Stigmas and discrimination against out-groups are but an extension of "blame gossip" to outsiders, who are seen as competitors for scarce resources and who are targeted as dangerous and deviant in-groups.

Linking gossip and reputation makes the latter a highly dynamic phenomenon in two distinct senses: it is subject to change, especially as an effect of corruption, errors, deception, etc.; and it emerges as an effect of a communication process. Reputation is both what people think about targets and what targets are in the eyes of others. It is more powerful because it may not even be perceived by the individual to whom it sticks, and consequently it is out of the individual's power to control and manipulate.

1.3 Methodology: computational models and agent systems

In the social sciences, computational models may provide a great help in studying and explaining the connections between the individual and the social levels (*Micro-Macro Link*, see Schillo et al., 2001). Several examples illustrate how unexpected the aggregate outcomes may be. It can happen that different individuals become completely segregated in similar groups although their preferences are not particularly in favor of similar individuals —to cite one of

the first research studies to make use of computational modelling technique (Schelling, 1978). In this example the micro level and the macro level interact in a complex way, so that the final results cannot be derived studying solely the individuals or the social system.

Due to the complexity of its dynamic, cooperation and gossip comprise another scientific topic that could profit from this methodology.

Agent systems Computational models are models expressed as algorithms and implemented as software. From their inception, driven by technological innovation, they have evolved constantly, and, coupled with agent theory (Shoham, 1993), have been adopted as a new research methodology in different scientific disciplines.

With the term *agent* we usually refer to an entity endowed with:

- *(limited) autonomy*: agents have a behavioural repertoire that they can use proactively;

- *the capacity to adapt and react to the environment*: agents live in an environment, although artificial; and this environment is also a social environment, inhabited by other agents, with their own —potentially conflicting— behavioural repertoires. As the simulation goes on, agents can adjust the rule governing their behaviour through learning, or copy new rules from their neighbors. Other models adopt an evolutionary mechanism, so the distribution of the actions in the repertoires of new generations of agents is a function of their fitness;

- *involved in local interactions*: collective behaviours are emergent properties, in the sense that there is no central authority orchestrating the

system; i.e. the overall outcomes is not known, anticipated, or even desired by the individual agents.

Agents can be modelled at different levels of abstraction: single cells, entire organisms (animal or human), up to super-individual entities like groups, firms and institutions. Thus, agent-based models (ABMs) have been employed to investigate a variety of research questions. Social and political phenomena relevant to modern society have been addressed: marketing and consumer behaviour (**?**), the timing of retirement (Axtell and Epstein, 2006); as well as historical phenomena based on archaeological findings (a well known and replicated example is Axtell et al., 2002). Agent-based computational economics is an affirmed label that covers all the contributions of the agent community in the economics (see Tesfatsion and Judd, 2006).

Examples of the application of computer simulation can also be found in social psychology (e.g. Kalick and Hamilton, 1986; Nowak et al., 1990; Stasser, 1988). Smith and Conrey (2007) argue that ABMs might constitute a valuable asset for social psychologists on the ground of their ability to connect the different dimensions of the topics of this very research field:

- intrapersonal processes (e.g. decision-making, heuristics, memory effects, personality differences);
- interpersonal processes (e.g. mate choice, social influence, reciprocity);
- group processes (e.g. social norms, stereotype and prejudice, status differentiation);
- and cultural processes (e.g. cultural transmission, innovation diffusion, etc.).

Reasons for social simulation ABMs simulate a set of processes observed in the world to understand them better. One of the reasons for performing simulation is data validation: the construction of a somehow formal theory able to generate a set of predictions that fit a series of empirical observations. But prediction is not the only reason. As noted by Epstein (2008), aside from their potential predictive power, computational models are built for their explanatory power. An explicit representation of a theory can have a great impact on the following issues:

1. illuminate core dynamics;
2. illuminate core uncertainties;
3. help to generate testable hypothesis, and to discover new questions;
4. guide data collection.

1.4 Contribution and outline of the dissertation

The work of this thesis contributes to the social scientific literature enhancing theoretical and practical knowledge about how communication affects reputation formation and diffusion, and how it can sustain cooperation. It proposes research based on the cognitive account of reputation developed by Conte and Paolucci (2002) and it uses agent-based methodology to explicitly represent the agents and their interactions, as well as the to derive the social outcomes at the macro level.

The proposed models address three issues. The first is related to the implementation of social- and cognitive-inspired algorithms for the design and the analysis of decentralized rating systems to ensure trust among partners. The second issue deals with the testing of theoretical hypothesis about the

dynamics of reputation as an effect of rumors' spreading, i.e. third-party social evaluations where the source of the evaluation is unknown. Finally, using results from a laboratory setting with natural subjects, the issue of communication as a cooperative device is addressed, with messages exchange between subject being subject the rules of reciprocity.

This dissertation is organized as follows: Chapter 2 outlines the theoretical analysis of reputation and gossip that serves as a reference for the simulation models. Then, after an analysis of the dynamics of partner choice and reputation in Chapter 3, where, using agent-based models, we extend the study of the effects of reciprocity on the enforcement of cooperation to include communication, Chapter 4 shows the results of a simulative study testing the communication algorithm implemented and compares the results of the spreading of different types of social evaluation, namely: image and reputation. Chapter 5 presents a novel experimental setting in which subjects performing a memory game are offered the possibility to exchange messages. While testing for the effects of cooperative action on their performances, we were able to analyse how the subject reacted to a failure in communication and the outcomes of the adopted strategy in dealing with unreliable partners. Finally, Chapter 6 address the main conclusion and discuss further directions of research.

Chapter 2

A social cognitive theory for reputation and gossip

The goal of this chapter is to present a cognitive theory of reputation relevant to the issue of communication, which informs the research questions explored in the chapters that follow.

2.1 Reputation reporting systems

Several attempts have been made to model and use reputation in artificial societies, especially in two sub-fields of information technologies, i.e. computerized interaction, with a special reference to electronic marketplaces, and agent-mediated interaction (Mui et al., 2002). The continuously growing volume of transactions on the World Wide Web and the potential for frauds

entailed led scholars from different disciplines to develop new online reputation reporting systems. These systems should provide a reliable way to deal with reputation scores or feedbacks, allowing agents to find out cooperative partners and avoid cheaters.

The existing systems can be roughly divided into two sub-sets: Agent-oriented individual approaches and agent-oriented social approaches, depending on how agents acquire reputational information about other agents.

The **agent-oriented individual approach** has been especially dominated by Marshs ideas (Marsh, 1992, 1994b,a), which have provided the bases for many additional algorithms and further study. This kind of approach is generally characterized by two attributes: (1) potential cooperation partners are found out by any one agent and (2) the agent only relies on its experiences from earlier transactions. When a potential partner proposes a transaction, the recipient calculates the "situational reputation" by weighing the reputation of his potential trading partner for further factors, such as potential output and the importance of transaction. If the resulting value overcomes a certain cooperation threshold, transaction takes place and the agent updates the reputation value to the current trade's outcomes. Otherwise, if the threshold is not reached, transaction will not be accomplished, and rejecting the transaction can be punished by a "reputation decline". These individual models differ with regard to the duration of memory. Agents may forget their experiences slowly, fast, or never (Marsh, 1994a).

In **agent-oriented social approaches**, the assessing of reputation is done by the agents themselves and not by an external entity. However, agents do not only rely on their direct experience, but are also allowed to consider third-party information (e.g., Yu and Singh, 2000; Regan and Cohen, 2005). Although these approaches share the same basic idea, i.e. that experiences of

other agents in the network can be used when searching for the right transaction partner, they rely upon different solutions when it comes to weigh the third-party information and to deal with "friends of friends". Thus the question arises as to how react to information from agents that seem to be little trustworthy. These differences point to the inter-subjective reputation values (Sabater and Paolucci, 2007).

Another problem lies in the storage and distribution of information. To form a whole picture of its potential trading partners, each agent needs both direct and reported evaluations in order to be able to estimate the validity and the informational content of the former. If said information resides in a public, accessible place, mutual ratings among agents may bring about collusion, blackmailing and retaliation (Regan and Cohen, 2005; Dellarocas, 2003).

Models of reputation for multi agent systems applications (Sabater and Sierra, 2002; Schillo et al., 2000; Huynh et al., 2006) clearly present interesting new ideas and advances over conventional online reputation systems, and more generally over the notion of reputation as a global, centrally controlled entity. Indeed, models of trust and reputation abound in this field (see Ramchurn et al. 2004; Sabater and Sierra 2004, for reviews).

One interesting contribution comes from Yu and Singh (2000), who proposed an agent-oriented model for social reputation and trust management which focused especially on electronic societies and MAS. The novelty of their contribution relies on the introduction of a gossip-mechanism ("If an agent A encounters a bad partner B during some exchange, A will penalize B by decreasing its rating of B [...] and informing its neighbours", Yu and Singh, 2000) in which ratings are transferred incrementally through the network of agents. And it arranges for a mechanism to include other agents' testimonies ("witness information") in one's own reputation calculation. For direct evalu-

ations, agents only rely on their own direct experience: they store information about the outcome of every transaction they had and recall this information in case they are planning to bargain with the same partner a second time. In case the agent meets another whit whom he has never traded, the second part of Yu and Singhs model come into play: the reputation mechanism. In this mechanism, so-called referral chains are generated that can make available witness information across several intermediate stations. An agent is thus able to gain reputation information with the help of other agents in the network. This reputation information is not as global as in classical online reporting systems (where every user can see all profiles of all other members and every evaluation a user is given accounts for his reputation value), but it is depending on the referral chain the requesting agent is using. As this chain represents only a small extract of the whole network, the information delivered by the chains can be partial.

It is worth emphasizing that in these domains trust and reputation are actually treated as the same phenomenon, and often the fundamentals of reputation mechanisms are derived from trust algorithms (Moukas et al., 1999; Zacharia, 1999; Zacharia et al., 1999). Moreover, several authors explain reputation in terms of trust and vice versa, continuously mixing up these two phenomena.

On the other hand, the research carried on with multi-agent systems' methodology explores the issue of reputation along several dimensions; however it characterizes for a tendency to consider reputation as an external attribute of the agents, without taking into account the process of creation and transmission of that reputation. These deficiencies prove that a more theory-driven approach is needed, and in the following sections we will suggest how a social cognitive approach to multi-agent systems can contribute to tackle the

complexity of reputation's dynamic.

2.2 Modelling reputation and gossip

In the present work I follow Conte and Paolucci (2002) when they consider reputation as a complex artifact, whose function is social control, and they regard gossip as the tool to build and alter the reputation of a social actor.

The theory developed is aimed at modelling the variety of mental states (including social goals, motivations, obligations) and operations (such as social reasoning and decision-making) necessary for an intelligent social system to act in some domain and influence other agents (social learning, influence, and control).

I will illustrate here the relevant features of the proposed social cognitive model of reputation, I will introduce the difference between reputation and image, the roles different agents play when evaluating someone and transmitting this evaluation and, finally, I will explain the decision processes underlying reputation.

2.2.1 Image and reputation

A cognitive process involves symbolic mental representations (such as goals and beliefs) and is based on the mental operations that agents perform upon these representations (reasoning, decision-making, etc.). A social cognitive process is a process that involves social beliefs and goals, and that is based on the operations that agents perform upon social beliefs and goals (e.g., social reasoning). A belief or a goal is social when it mentions another agent and possibly one or more of his or her mental states (Conte and Castelfranchi, 1995; Conte, 1999).

In their account of reputation, Conte and Paolucci (2002) focus on two different kinds of cognitive representations, namely: *image* and *reputation*. Image consists of a set of evaluative beliefs (Miceli and Castelfranchi, 2000) about the characteristics of a given agent, i.e. it is an assessment of her positive or negative qualities with regard to a norm, a competence, and so on.

According to Miceli and Castelfranchi (2000), an evaluation is a hybrid representation. An agent has an evaluation when he or she believes that a given entity is good for, or can achieve, a given goal. An agent has a social evaluation when his or her belief concerns another agent as a means for achieving this goal.

A social evaluation implies three sets of agents:

- a nonempty set E of agents who share the evaluation (evaluators);
- a nonempty set T of evaluation targets;
- a nonempty set B of beneficiaries, i.e., the agents sharing the goal with regard to which the elements of T are evaluated.

Often, evaluators and beneficiaries coincide, or at least have nonempty intersection but this is not necessarily the case. A given agent t is a target of a social evaluation when t is believed to be a good/bad means for a given goal of the set of agents B, which may include or not the evaluator. Social evaluations may concern physical, mental, and social properties of targets; e.g. agents may evaluate a target as to both his or her capacity and willingness to achieve a shared goal. The interest/goal with regard to which t is evaluated may be a distributed or collective advantage. It is an advantage for the individual members who are included in the set B, or it may favour a supra-individual entity, that results from the interactions among the members of B (for example, if B's members form a team).

An agent's reputation is argued to be distinct from, although strictly interrelated with, its image. More precisely, while image is defined as a set of evaluative beliefs about a given target, reputation is defined as the process and the effect of transmission of image.

More precisely, reputation comes into existence from the interaction of three distinct but interrelated objects:

1. a believed evaluation, i.e. a second-order belief;
2. a population object, i.e. a propagating believed evaluation;
3. an objective emergent property at the agent level, i.e. what the agent is believed to be.

Consequently, communication about reputation is a communication about a second-order belief, i.e., about others mental attitudes. To spread news about someone's reputation does not bind the speaker to commit himself to the truth value of the evaluation conveyed but only to the existence of rumours about it. Therefore, unlike ordinary sincere communication, only the acceptance of a second-order belief is required in communication about reputation. And unlike ordinary deception, communication about reputation implies:

- *no personal commitment* of the speaker with regard to the main content of the information delivered. If the speaker reports on t's bad reputation, he is not necessarily implying that t deserved it;
- *no responsibility* with regard to the credibility of the source of information: in fact, evaluations conveyed as rumors do not usually disclose the source (i.e. "I was told that t is a bad guy").

To assume that a target $t \epsilon T$ is assigned a given reputation, then, implies assuming that t is believed to be "good" or "bad," but it does not imply

sharing either evaluation. To account for this important characteristic, a fourth set of agents is added to the three previously listed:

- a nonempty set G of gossiping agents who share the 2nd-order belief that members of E share the evaluation; this is the set of all agents aware of the effect of reputation.

Often, E can be taken as a subset of G; the evaluators are aware of the effect of evaluation. In most situations, the intersection between the two sets is at least nonempty, but exceptions exist. G in substance is the set of reputation transmitters, or third parties. Third parties share a second-order belief about a given target, whether they share the concerned belief or not.

In real matters, agents may play more than one role simultaneously: Evaluator, Beneficiary, Target, and Third Party. In the following, I will examine the characteristics of the four roles in more detail.

2.2.2 Reputational roles

Evaluator Any autonomous agent is a potential evaluator. Social agents are likely to form evaluative beliefs about one another as an effect of interaction and social perception (see Castelfranchi, 1988). Agents may interfere positively and negatively with one another. On the one hand, one agent may be a fundamental resource to achieve another's goals, and a compensation for his or hers limited autonomy. On the other, each agent may be a source of social conflict or concurrence, a predator, an enemy, a rival in the acquisition of a number of resources, etc.

Social evaluations are brought about when agents evaluate one another with regard to their individual goals. In this case, evaluations serve to identify friends and partners and to avoid enemies. Furthermore, agents evaluate one

another also with regard to goals or interests shared by a given set of agents, which the evaluators may belong to or not.

A bad evaluation may be formed about violators of others' rights, about agents behaving in an (apparently) malevolent and hostile manner, whether or not the evaluators consider themselves as potential victims of such doings. Information thus obtained may be used, lacking more detailed data, to infer that the target could violate other rights in the future, namely, those of the evaluator. In addition, evaluators may be concerned with one another's power to achieve the goals or interests of abstract social entities or institutions, as when we judge others' attitudes towards the norms, the democracy, the government, the religion, the state, etc. Agents evaluate one another with regard to the goals of those they adopt, be the latter other individual agents (i.e., one's offspring) or supra-individual agents, such as groups, organisations, or abstract social entities.

Beneficiary A beneficiary is the entity that benefits from execution of the behaviour with regard to which targets are evaluated. Beneficiaries can either be individual agents or groups and organisations or even abstract social entities like social values and institutions. Beneficiaries may be aware of their goals and interests, and of the evaluations, but this is not necessarily the case. In principle, their goals might simply be adopted by evaluators —as happens, for example, when people who belong to the majority support norms protecting minorities. In fact, evaluators often are a subset of beneficiaries.

Target The target is the evaluated entity. In general, image may also concern inert targets, objects, or artefacts to be used by others, while in reputation the mental and moral components are necessarily involved. Holders of reputation are endowed with (or attributed) a number of important implicit

and/or explicit characteristics:

- Agency: more specifically, autonomous agency and sociality. The target is evaluated with regard to a given behaviour, autonomously executed.
- Mental states: specifically willingness to perform the above behaviour.
- Decision-making: deliberative capacity, i.e., the capacity to choose a desirable behaviour from a set of alternatives.
- Social responsibility: the power to prevent social harms and, possibly, to respond for their occurrence.

Targets of image and reputation may also be individual or supra-individual. In the latter case, they coincide with a set, a group, a collective, an abstract entity, or a social artefact, such as an institution, provided this (is attributed the capacity to make decisions, achieve goals, and perform actions.

Gossiper or Third Party A gossiper is the agent in the position to perform an act of reputation transmission. An agent is a (potential) third party if she transmits (is in position to transmit) reputation information about a target to another agent or set of agents. Although sharing awareness of a given target reputation, third parties do not necessarily share the corresponding image of the target. That is, they do not necessarily believe it to be true.

Agents may also spread a false reputation, i.e., pretend that a target has a given reputation when this is not the case. Agents do this in order to achieve the aforementioned benefits without taking responsibility for spreading a given social evaluation.

Third parties may have no personal experience and familiarity with the target, which is one reason why they may not have formed an image of it. In

such a case, they might have received information from other third parties or from agents who have had a direct contact with the target.

2.2.3 Reputation-based decisions

To better define the difference between image and reputation, the decision-making processes based upon them must be analysed at the following three levels:

- **Epistemic**: *accept* the beliefs that form either a given image or *acknowledge* a given reputation. To accept a given image implies coming to share it. This acceptance may be based, for example, upon: supporting evidence and first-hand experience with the image target; consistent pre-existing evaluations (concerning, for example, the class of objects to which the target belongs); or trust in the source of the evaluation. Conversely, to acknowledge a given reputation does not lead to sharing others' evaluations but rather to the belief that these evaluations are held or simply circulated by others. To assess such second-order belief is a rather straightforward operation. For the recipient to be relatively confident about it, it is probably sufficient for him or her to hear some rumours.

- **Pragmatic-Strategic**: use evaluative beliefs in order to decide whether and how to interact with the evaluation's target. Such belief may be the result of the agents' personal experience; but it may also follow from the acceptance of others' evaluations, or the acknowledgement of the target's reputation. Reputation has a high strategic value, because it may provide hints for social interactions when agents have no experiences or past interactions with the target, hence no image about him or her. In

other cases the social image and the reputation can have contradicting values. If this situation occurs, agents' decisions may be contingent to the social settings: e.g. usually inputs from reputation are superseded by the social image of the target; however, if the social setting requires it, agents may avoid interaction with the target in spite of positive personal evaluations about him/her.

- **Memetic**: transmit my (or others') evaluative beliefs about a given target to others. Whether or not I act in conformity with a propagating evaluation, I may decide to spread the news to others. Again, this decision —and the following behaviour— is completely autonomous from the previous two. A third party may be bluffing; he or she may pretend to be benevolent with regard to beneficiaries, in order to: enjoy the advantages of sharing information about reputation; be considered an in-group by the other evaluators; gain a good reputation without sustaining the costs of its acquisition (as would be implied by performing the socially desirable behaviour); avoid the consequences of a bad reputation.

2.2.4 Aims of the present approach

Besides communication and gossip, there may be other factors that can influence social evaluations; for example: similarity, social proximity or affiliation. Members may inherit the reputation of their social categories and groups, as offspring may inherit their parents' reputation. Affiliation may imply inheritance of the institution's reputation, and an employee may suffer from the bad reputation of the firm he or she works in. This dynamic corresponds to what is usually called prejudice, but it is not intrinsic to reputation, although it may empirically co-occur with reputation spreading. In the following chap-

ters, however, I will focus only on communication processes when considering possible sources for reputation formation and spreading. And I will use the concepts defined in the sections above to justify the implementation choices made in the development of the simulations.

The type of social cognitive approach advocated by Conte and Paolucci (2002) is receiving growing attention within the so-called Sciences of the Artificial (Simon, 1996), in particular Multi-Agent Systems and Social Simulation. It is oriented towards the processes, rather than the contents, of the phenomena analysed and it aims at modelling and possibly implementing systems acting in a social —whether natural or artificial— environment. In the next section I will consider the relevant methodological issues.

Chapter 3

Communication and cooperation

The work of this chapter is based on Di Tosto et al. (2007). It presents an exploratory work framed within a research project aimed at the study of the emergence and evolution of prosocial behaviour (including altruism, cooperation, and compliance with norms) among autonomous agents. It will focus on the effect of two kinds of communication mechanisms on the strategies of artificial agents in social settings inspired by the Prisoners' Dilemma Game, and compare their results.

3.1 Enforcing prosocial behaviour

As shown within a huge literature on the iterated prisoners dilemma (IPD), cooperation among non-kin needs to be sustained by enforcing mechanisms, the most frequent of which is punishment, i.e. a propensity to shift to defection with defectors (i.e. tit-for-tat). However, results obtained by this means are found to be sensitive to errors in strategy execution as well as invasions by free riders. Furthermore, some authors (Back and Flache, 2006) convincingly argued that reactive strategies are not so frequently followed in human societies and, together with other authors (Hruschka and Henrich, 2006), insist on the importance of social networks in the emergence of cooperation. Hence, variants of reactive strategies that fit with long-term relationships have been proposed, like cliquishness, i.e. a propensity to defect with strangers (Hruschka and Henrich, 2006), or commitment (Back and Flache, 2006). The latter, in particular was found to 'benefit more from being unconditionally cooperative' although unconditional cooperation makes strategies vulnerable to exploitation.

An enforcing mechanism that rapidly gained popularity in the literature of reference concerns partners' assortation. Generally speaking, in evolutionary game theory, play is forced and attention is preferably given to the use of partner selection for retaliation. Some game theory studies have allowed players to avoid unwanted interactions, or more precisely to affect the probability of interaction with other players through their own actions (see de Vos et al., 2001; Hirshleifer and Rasmusen, 1989; Orbell and Dawes, 1993; Stanley et al., 1994; Ashlock et al., 1996; Hauk, 2001). Thanks to partner choice and refusal, payoff scores are found to increase since players can protect themselves from defections without having to defect themselves, and defectors get ostracized. On the other hand, choice and refusal also permit opportunistic

players to home in quickly on exploitable players and form parasitic relationships. In particular, cooperators seem to take advantage of choice and refusal over nonreciprocators.

Finally, a fundamental mechanism supporting both punishment and partner choice is communication. As it requires agent-based rather than equation-based simulation, communication has none or poor tradition in the study of the emergence of altruism and cooperation. Nonetheless, its role in promoting informational cooperation as a means for material cooperation, at least in human societies, can hardly be denied.

Enforcing mechanisms have usually been observed in a fragmentary, non-systematic way, often starting from unclear concepts, not uniquely defined. Apparently, neither the interplay of punishment and partner selection and refusal nor the role of specific modalities of communication in supporting them has been addressed explicitly in the study of prosocial behaviour.

Last but not least, the variety of prosocial behaviour has not been dutifully considered. Again, altruism and cooperation are often treated as interchangeable notions, or at least as if potential differences among them did not affect the conditions under which they emerge. On the contrary, we believe that the peculiar features of these various forms of prosocial behaviour ought to be more carefully analysed and distinguished.

In this chapter, we will turn the reader's attention on the following issues:

1. Which is the most efficient mechanism of enforcement of prosocial action? Usually, this question is raised in the context of IPD. What about altruism?

2. To what extent do agents contribute with their social intelligence to the efficiency of these mechanisms? In particular,

a) to what extent does communication contribute, and how does it interact with partner selection and punishment?

b) which modalities of communication can be envisaged, and which one is more beneficial?

3.2 Previous notions

The question is which enforcing mechanisms are needed to promote prosocial action? We explored the impact of the following mechanisms:

3.2.1 Conditional cooperation

Punishment (P), defined as a change of strategy (namely from cooperation to defection in presence of a known defector). To note that punishment here is something different from what is usually intended in the literature about altruistic punishment, where it denotes a costly behaviour useful to prevent free-riding of public goods (Boyd et al., 2003; Fehr and Gachter, 2002). In the present setting, however, benefits are not shared among the population, and we call punishment *every defection towards a known noncooperator* (based on personal experience, if repeated interaction are allowed; or on reported social evaluations, if communication is allowed—see sections 3.3.2, 3.4 for further details).

3.2.2 Games with choice and refusal

Partner selection (PS) and refusal (PR): this is a non-random, rule-based assortation, such that each agent tries to associate with the partner from which they expect the highest payoff *given their own prosocial attitude*, and

reject any other. Both cooperators and defectors expect to obtain the highest payoff by playing with a cooperator.

3.2.3 Image-score game

With regard to altruism, we used an asymmetric game, modelled after the *image scoring game* (Nowak and Sigmund, 1998). In each period, players are partnered and one is given the chance to take a costly action that helps the other. Cooperating in this manner is socially efficient, but the only way to monitor free riding is through the image score which, in this context comes to an accounting of a player's past altruistic actions. The image score is a property of agents immediately and universally accessible to everybody, with no actual communication process involved. Modelling reputation as an explicit feature of the agent is a solution that has been proved to be a good answer to the problem of indirect reciprocity. Image score is not the only one; tag-based systems (Riolo et al., 2001; Hales and Edmonds, 2003) are based on a similar idea and generalize on the notion of image and image score to develop a system in which interactions take place between agents with similar tags. Tags are empty labels and, unlike image score, do not have to possess a semantic. But, like image score, they are public information which are compared by the agents before performing cooperative interactions. Tag-based systems, thanks to their level of abstraction, are now applied to different problems in different domains (Hales and Edmonds, 2005).

These techniques seems to share a common intuition which can be found elsewhere in the literature about the evolution of cooperation. Dawkins (1976) discusses the possible evolutionary effects of a phenotype trait capable of signaling the presence of its gene(s) to other organisms. Dawkins named this phenotype trait as green beard, and the following effect consists in an altru-

istic behaviour towards the organism with the green beard, independent to the degree of relatedness of the other organisms. Despite the fact that an example of the green beard effect was actually found in nature (Grafen, 1998), this theory, with the relative mechanisms, paves the way to a series of consideration. In nature, phenotype traits are common indicators of status and identities and are involved, e.g., in defensive strategies and conflicts; in some cases they are used to bluff. But, in all these cases, they are used to avoid dangerous interactions or to find solutions to the problems of the social life that minimizes costs for all the organisms.

To these, we add the communication mechanism in two distinct modalities:

- Private: one-to-one or one-to-few exchange of messages. Cooperators spread social evaluations they have acquired both directly through own experience, or by means of communication, to known cooperators. Social evaluations concern the prosocial attitudes of potential partners.
- Public: one-to-many or blackboard-like exchange of messages about social evaluations from cooperators to the whole population.

The intuition behind is that the latter mechanism is faster and therefore more efficient, but more dangerous as it exposes cooperators to belief-based exploitation from defectors.

3.3 Simulation model: How and when communication can enforce cooperation

Starting from the assumption that prosocial behaviours among autonomous agents require special conditions to emerge and proliferate, we draw a simulation model both to test the effects, alone and in combination, of three enforcing

mechanisms and to evaluate which of them is more suitable for making cooperation and altruism emerge. Traditionally, the main mechanisms to empower cooperators to protect themselves against cheaters are: punishment, partner selection, and communication. These three tools work differently and produce diverse effects depending upon the kind of prosocial situation considered. In what follows we will describe how each of these mechanisms works in the two games, and what effects they are expected to determine.

Prosocial action varies on many dimensions and in many ways. One main source of variability is mental: agents pursue different goals and are guided by different beliefs while executing different types of prosocial action (for a discussion of these aspects the reader is turned to Conte and Castelfranchi, 1995; Conte, 1999). However, types of prosocial action, like altruism and cooperation, differ also in directly observable features:

- Symmetry *vs* asymmetry: this consists of two specific components, role complementarity and direction of benefit. Altruism is asymmetric in both senses, as for each episode the roles of donor and recipient are played by different agents and the direction of benefit is from donor to recipient; cooperation is symmetric as it allows for role identity and bidirectional benefit. It is of some interest to notice that in other types of prosocial action, for example social exchange, roles are complementary but benefit is bidirectional. Conversely, role identity and unidirectional benefit occurs only in antisocial action, i.e. exploitation.

- Individual *vs* shared benefit: this consists of the recipient of benefit, which in altruism can only be the beneficiary, while cooperation, at least in principle, allows for a shared benefit. Again, for the sake of analysis it may be interesting to observe that in social exchange by definition no

shared benefit is allowed.

The interplay between these features leads to a third fundamental distinction:

- One-shot exploitation: in altruism, the donor is exploited only if she is not reciprocated later on. In cooperation, instead—think of the classic one-shot Prisoner's Dilemma—a cooperator may be exploited online by a partner playing defection.

Hence, in our terms, altruism is an asymmetric form of prosocial interaction characterized by individual benefit and no immediate exploitation. On the contrary, cooperation is a symmetric interaction, where benefit may be shared but immediate exploitation is also possible.

3.3.1 Games of cooperation and games of altruism

The PD game has been extensively used by game theorists to address the issue of cooperation. In this game, both players have the choice to cooperate (C) or to defect (D), and the equilibrium outcome is defection for both players. This outcome is deficient, whereas cooperation is the Pareto-optimal outcome for both players. The PD game structure is well-known and it does not deserves further explanations. The simplicity of the PD game has led many scholars to use it to model several social and biological phenomena (Doebeli and Hauert, 2005).

Altruists and cooperators in both games are doomed to extinction in a population where half of the agents are neither reciprocators nor cooperators, i.e. cheaters, as it is in our model. In fact, in the basic version, cooperators always play C in the PD game, and they always donate in the altruistic game, without differentiating between cooperative and non cooperative partners. This behaviour easily exposed themselves to exploitation to death by

cheaters. This extreme situation allows us to put to the test 4 mechanisms that are supposed to enforce prosocial behaviours such as altruism and cooperation. Given the two games, we explored the effects of these variables, once per time or in combination, in terms of the average payoffs of agents.

3.3.2 Communication: Private and public information

In our terms, punishment is the possibility to react to a defection playing Defect in the PD game, or playing keep in the game of altruism. When punishment is not active, agents simply play their built-in strategy, without the possibility of shifting it when facing dishonest partners. Partner Selection can be twofold: active and passive Partner selection. The former means that agents can select their partners in the interaction. In the PD game, this should lead cooperators to home in, avoiding defectors. Anyway, this same mechanism also permits cheaters to choose cooperators and exploit them. In the altruistic game, altruists should choose altruist to play give, expecting a reciprocal donation in the following of the game. On the other hand, we call Refusal the possibility to avoid an interaction, i.e. to escape from a cheater. Every time an agent refuses to interact, both agents of the couple loose an opportunity to interact. When Partner Selection was not available, agents were randomly paired. Finally, we explored three communication conditions:

1. No messages: no communication allowed.

2. Private messages: the access to this kind of communication is limited to cooperators. Cooperative agents send messages about their partners to other cooperative agents previously met. Cheaters can neither send nor receive messages in this condition.

3. Public messages: this modality works as the image score does. Once an interaction is over, the receiver sends a message about the nature of the mover, i.e. a cheater or a cooperator. The message is posted on a blackboard accessible to both cooperative and non cooperative agents.

3.4 Experimental Settings and Results

A population of 500 agents, randomly assigned to one of two groups, cooperators and defectors, is created to play the game of altruism, and the game of cooperation. At each turn of the simulation, agents are coupled and face the option to perform or not a prosocial action, which will confer a benefit $b = 1.0$ to another agent, at a cost $c = 0.1$ to himself. After 200 interactions the gain of each agent are collected and analysed.

Performances of the two groups of agents are tested in several experimental conditions, under which we observed the effects of the mechanisms proposed. In figure 1 the average payoffs for the groups of cooperators and defectors are reported. Results are divided by game type and communication type. Furthermore the effects of PS and P are presented, with and without the possibility of refusal.

Table 3.1 shows the results of a regression analysis conducted with agents' payoffs as independent variable in order to asses the relative importance of PS and P in the interpretation of simulations data. Generally, both PS and P have the effect to lower the average payoffs, in both the groups of agents. PS is found to be the most important factor for the explanation of the variation of simulation outcomes, with few exceptions: in the cooperative game, without the possibility of Partner Refusal, Punishment is the most important factor to interpret the outcome of the defector strategy, independently of the communication mechanism. The cooperative game without PR is the condition where

Games of altruism with refusal

Communication	No Messages				Private Msgs				Public Msgs			
Agent Type	Cooperators		Defectors		Cooperators		Defectors		Cooperators		Defectors	
	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	59.9390***	0.5303	102.91870***	0.21474	59.9047***	0.5350	104.86628***	0.20758	62.4438***	0.4332	102.5312***	0.2473
punish	1.6025***	0.2431	-1.61888***	0.09863	-0.2126	0.2460	-3.23283***	0.09506	0.4736*	0.1993	-7.0825***	0.1132
partner	-11.7385***	0.2431	-43.37943***	0.09863	-9.4042***	0.2460	-43.85994***	0.09505	-13.1087***	0.1993	-41.8904***	0.1132
R-Squared	0.1904		0.9514		0.1278		0.9553		0.3026		0.9335	
Adj. R-squared	0.1902		0.9514		0.1276		0.9553		0.3025		0.9334	

Games of cooperation with refusal

Communication	No Messages				Private Msgs				Public Msgs			
Agent Type	Cooperators		Defectors		Cooperators		Defectors		Cooperators		Defectors	
	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	139.7175***	0.6398	196.6988***	0.2692	138.24267***	0.60848	190.7708***	0.2821	141.9865***	0.4203	150.0152***	0.3093
punish	-0.5302 .	0.2939	-4.9999***	0.1233	-0.02781	0.28113	-6.9913***	0.1285	2.4354***	0.1927	-12.0476***	0.1420
partner	-37.7808***	0.2939	-84.3273***	0.1233	-37.20654***	0.28113	-80.5589***	0.1285	-45.0963***	0.1927	-56.3656***	0.1420
R-Squared	0.621		0.9793		0.6372		0.9753		0.8466		0.9425	
Adj. R-squared	0.621		0.9793		0.6371		0.9753		0.8466		0.9425	

Games of altruism without refusal

Communication	No Messages				Private Msgs				Public Msgs			
Agent Type	Cooperators		Defectors		Cooperators		Defectors		Cooperators		Defectors	
	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	61.8171***	0.5299	95.61725***	0.20895	59.6090***	0.5369	95.18848***	0.20811	62.5209***	0.4647	98.8703***	0.2614
punish	-0.1048	0.2425	-2.57903***	0.09613	0.6980**	0.2457	-2.43151***	0.09573	0.6623**	0.2136	-8.2626***	0.1197
partner	-11.4568***	0.2425	-34.05266***	0.09613	-10.4758***	0.2457	-34.44588***	0.09573	-12.1315***	0.2136	-34.0189***	0.1197
R-Squared	0.1824		0.9268		0.155		0.9284		0.2431		0.8959	
Adj. R-squared	0.1822		0.9267		0.1548		0.9284		0.2429		0.8959	

Games of cooperation without refusal

Communication	No Messages				Private Msgs				Public Msgs			
Agent Type	Cooperators		Defectors		Cooperators		Defectors		Cooperators		Defectors	
	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error	B	Std. Error
(Intercept)	121.3757***	0.6174	231.260	1.257	121.9467***	0.6177	233.8512	1.1616	114.0966***	0.4852	230.427	0.910
punish	7.4729***	0.2817	-70.247	0.580	4.7609***	0.2832	-76.7944	0.5334	14.1121***	0.2229	-96.307	0.417
partner	-31.7445***	0.2817	-15.082	0.580	-29.2487***	0.2832	-11.4454	0.5334	-34.3346***	0.2229	4.613	0.417
R-Squared	0.573		0.6071		0.5238		0.6789		0.7374		0.8406	
Adj. R-squared	0.573		0.607		0.5237		0.6788		0.7374		0.8406	

Table 3.1: OLS regression model with dependent variable *gain*; regressors (PS and P) are dummy variables.

defectors obtain their higher average payoffs (see figures 3.1 and 3.2), and the punishment algorithm is the only mechanism through which cooperators can compete and avoid exploitation. In all the other settings PS is found to be responsible for the results, and appear to be the most effective tool for the enforcement of prosocial behaviour.

3.5 Discussion and concluding remarks

In this chapter, we presented a simulation study of the effects of different enforcement mechanisms, such as punishment (P), partner selection (PS), in interaction with each other and with one-to-many and one-to-few communication, on symmetric prosocial behaviour, namely cooperation, and asymmetric one, namely altruism. P consists of a shift in strategy (from prosocial to antisocial) in presence of cheaters, while PSR reduces the potential number of interactions to those occurring between known prosocial partners (partner refusal is a variant of PS in which chosen partners can escape interaction).

In particular, P and PS are found to be

- almost perfectly complementary, i.e. PS favors altruists rather than cooperators whereas P at the opposite enforces cooperators but has no effect on altruists.
- P in general is found to produce average payoffs higher than PS.

However, communication mitigates these results. In particular, although the modality with private messages allows higher payoffs for cooperators in partner selection than the modality with public messages (as it does not expose them to exploiting defectors), the latter modality

- favors altruists even without partner selection;

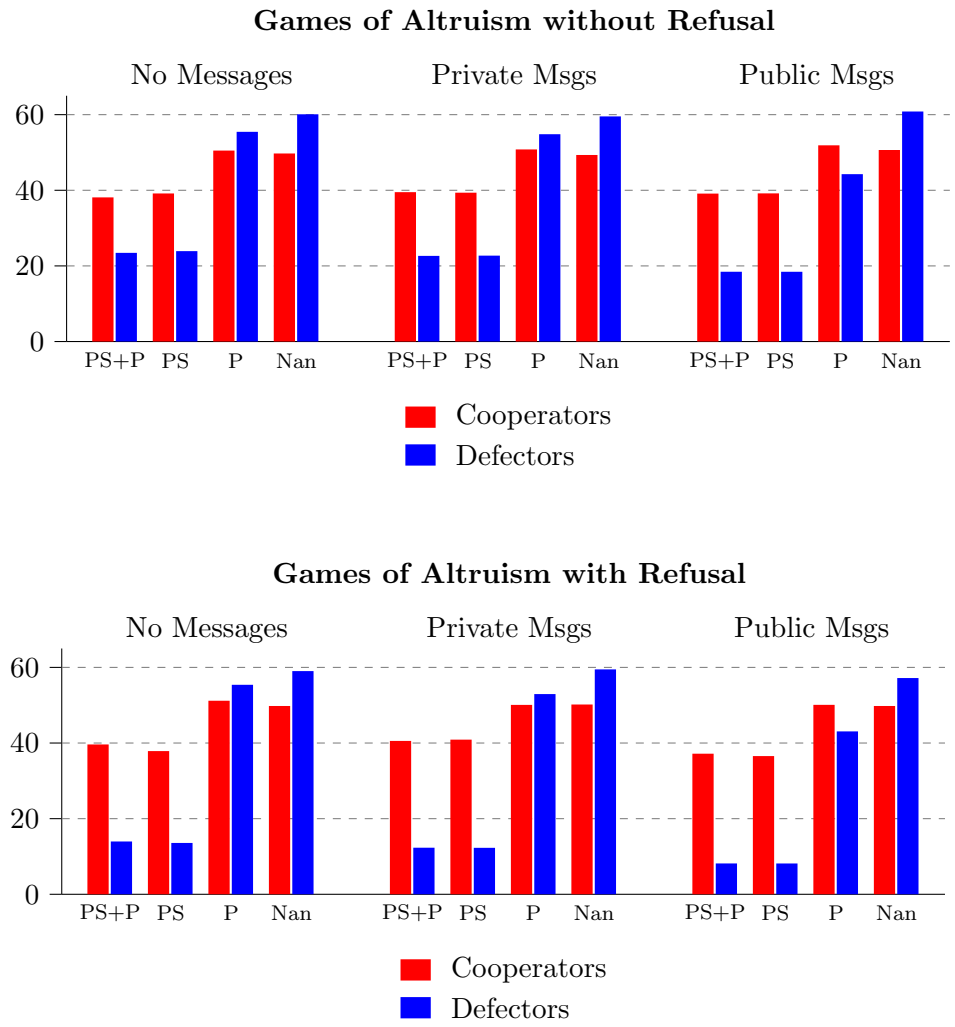

Figure 3.1: Average payoffs for the Game of Altruism, with and without Refusal, after 200 iterations. The two subpopulation of agents are divided according to the first move of their strategy: Cooperators if they start the game playing cooperation; Defectors, if they start cheating. Cooperators have then the possibility to either: punish cheaters for their defections in the following rounds (P), seek cooperating partners for future interactions (PS), combine both mechanisms ($PS + P$), or stick to the cooperative strategy in case the two mechanisms of strategy's change are not allowed ($None$). Results are compared in the two types of Communication conditions: $PrivateMessages$ and $PublicMessages$; and the Control condition, $NoMessages$, where no information is exchanged between agents.

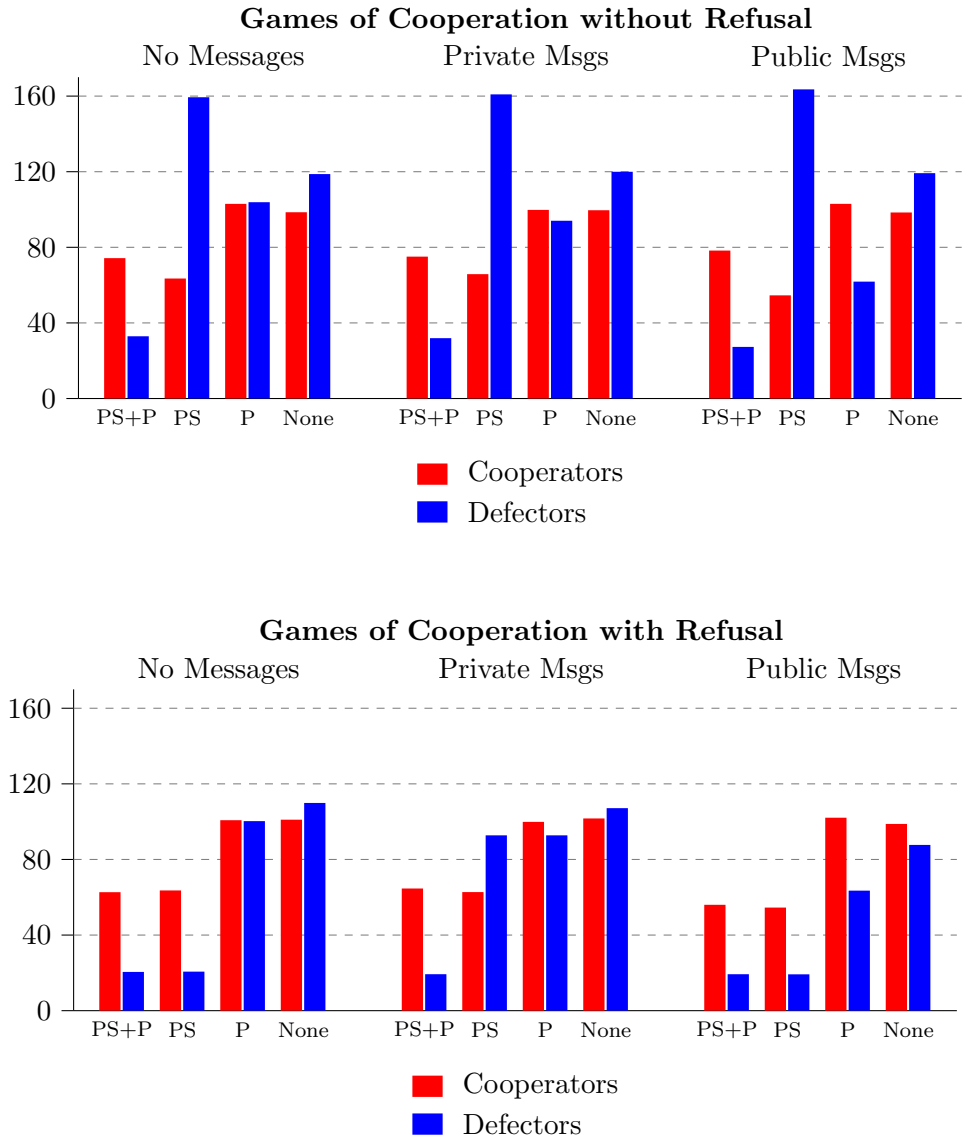

Figure 3.2: Average payoffs for the Game of Cooperation, with and without Refusal, after 200 iterations. The two subpopulation of agents are divided according to the first move of their strategy: Cooperators if they start the game playing cooperation; Defectors, if they start cheating. Cooperators have then the possibility to either: punish cheaters for their defections in the following rounds (P), seek cooperating partners for future interactions (PS), combine both mechanisms ($PS + P$), or stick to the cooperative strategy in case the two mechanisms of strategy's change are not allowed ($None$). Results are compared in the two types of Communication conditions: $PrivateMessages$ and $PublicMessages$; and the Control condition, $NoMessages$, where no information is exchanged between agents.

- favors cooperators even without punishment.

It is worth noting that in cooperation the coupling of punishment and partner selection yields worse results than punishment alone.

Furthermore, our findings allow for a comparison of PS and partner refusal (PR). In particular, PR is extremely competitive with punishment even in cases of cooperation. In this situation, this is the only experimental condition beside punishment in which cooperators are better off than noncooperators: this is rather obvious, since it is the only condition in which interactions between a cooperator and a known defector are impossible.

Finally, communication helps both altruists and cooperators. With the former, it works even in absence of PS, at least in the modality with public messages, and gives better outcomes to altruists and cooperators in the modality private messages. As expected, public messages are faster and can be decisive, but private ones are less dangerous and can give a stronger advantage to the good guys.

Not surprisingly, all the mechanisms examined contribute, although in different ways and to different degrees, to enforcing prosocial behaviour. However, unlike what one would expect, they are not hierarchically ordered in terms of efficiency. Rather, their efficiency depends on the type of interaction in which they are observed: in particular, PS is required to promote altruism and is irrelevant in cooperation, whereas P is needed for enforcing cooperation but is useless in altruism. Why is this the case?

The rationale of these results is incorporated in the very nature of the two forms of prosocial action considered. Thanks to agent-based computational modelling, requiring these two forms of prosocial action to be formally modelled before and in order to be compared, it was possible to observe that features such as symmetry/asymmetry not only allow altruism and coopera-

tion to be kept distinct but also call for different enforcing mechanisms. In particular, punishment reduces exploitation when immediate exploitation is possible, i.e. in cooperation but not in altruism. Conversely, partner selection promotes altruism thanks to an increased proportion of donations to the benefit of altruists on the total number of donations. But since this is obtained by reducing the total number of donations, the outcomes obtained by altruists are lower than those they would obtained by means of punishment. Still, in the latter condition, altruists are worse-off than nonaltruists. Otherwise stated, partner selection is more efficient when the promoted benefit is individual rather than shared, whereas punishment performs in the opposite way: as it excludes no one from interaction, neither the good nor the bad guys, it allows higher payoffs to be obtained, which than benefit cooperators more than defectors when the game is symmetric and each has something to gain from interaction, i.e. the benefit is shared.

Interestingly, partner refusal sometimes makes it on its own: this happens in cooperation. Whereas with PS only, cooperators that are not known and therefore not chosen by their fellows are fully exposed to exploitation, with PR they can find an escape, and end up with being even better-off than defectors, even without the help of punishment.

In future studies other forms of prosocial action will be investigated and compared with those examined in this work, also including other modalities of communication. In addition, we will compare the results obtained by cheaters and honest agents in non-homogeneous populations with varying percentages of the two behaviours in both conditions, in order to test how robust are altruism and cooperation to cheaters' invasion. Finally, the present findings will be re-analysed for different values of the individual attitude to cheating and with different payoffs structures.

Chapter 4

Simulating the spreading of social evaluations

Social evaluations are pieces of information regarding other agents whose attitudes, behaviours and actions are assessed with respect to some specific dimensions or aspects. Individuals use these evaluations as guidance to predict others' behaviours and to choose the most appropriate response when first-hand experience is not available or when it is too costly, in terms of risk, time and energy, to be acquired.

We conducted experiments on industrial districts because they represent an ideal candidate to study the effects of the spreading of social evaluations on dynamics of partner selection among self-interested agents. Industrial districts (Ids) can be conceived as complex systems made of heterogeneous but strictly interrelated and complementary firms that interact in a non-linear

way. One of the distinctive features of industrial districts is the tight connection existing between the social community and the firms: in this context, economic exchanges are mainly informed by social relationships and holding good reputation is an asset that may actually foster potential relations. We designed a simulation to model the effects of social evaluations on firms in an artificial cluster through Multi-Agent Simulation (MAS) techniques, in order to investigate whether and how different kinds of social evaluations have an impact on firms' quality and on their profits. Likewise, we then compared the effects of sincere and insincere information on the economic performances of the single firms and of the cluster as a whole.

The work of this chapter is based on Giardini et al. (2008a) and Giardini et al. (2008b).

4.1 Introduction

Transmitting social evaluations, i.e. gossiping, is crucial in human societies, in which gossip facilitates the formation of groups (Gluckman, 1963): gossipers share and transmit relevant social information about group members within the group, at the same time isolating out-groups. Besides, gossip contributes to stratification and social control, since it works as a tool for sanctioning deviant behaviours and for promoting, even through learning, those behaviours that are functional with respect to the group's goals and objectives, mainly norms and institutions (Wilson et al., 2000). Sommerfeld et al. (2007) consider gossip as a way to transfer social information within groups, alternative to direct observation. *This flow of information maintains cooperation by indirect reciprocity.*

Furthermore, reputation is considered pivotal in creating and sustaining prosocial behaviours in large human groups. Theories of indirect reciprocity

show how cooperation in large groups can emerge when the agents are endowed with, or can build, a reputation (Alexander, 1987; Nowak and Sigmund, 1998; Gintis et al., 2001). According to this theory, large scale human cooperation can be explained in terms of conditional helping by individuals who want to uphold a reputation and then to be included in future cooperation (Panchanathan and Boyd, 2004), as demonstrated by several experimental studies (for an introduction, see Fehr and Gachter, 2000). Reputational information can also solve the *tragedy of the commons*, a social dilemma referring to the fact that a public good will be overused if everybody is allowed to do so (Wedekind and Milinski, 2000). Allowing people to build up a reputation, prevented the public resource from being overused.

Although influential, these theories suffer the flaw of what Granovetter (1985) calls an *undersocialized* notion of reputation.

> Economists have pointed out that one incentive not to cheat is the cost of damage to one's reputation; but this is an undersocialized conception of reputation as a generalized commodity, a ratio of cheating to opportunities for doing so. In practice, we settle for such generalized information when nothing better is available. (Granovetter, 1985, p. 490)

Granovetter points out the relevance of information coming from one's own past dealings with someone, highlighting the benefits of this second kind of information that is cheap, more detailed, and, of course, accurate. This kind of information can be easily acquired thanks to *embeddedness*, i.e. the fact that human actions are motivated and explained by their being embedded in a network of social relationships that foster cooperation and guarantee against cheaters.

Embeddedness applies to several contexts, but it becomes crucial in closed environments, in which the web of relationships among agents determines their behaviours, actions and results. This occurs, for instance, in industrial districts[1], in which the interplay between economic dimensions and social relationships is very close. Industrial clusters are usually defined as networks of interactions among heterogeneous and complementary firms embedded into a specific geographic area. In the district, the form of production requires a high degree of cooperation between firms and the lack of formal agreements could lead actors to behave in an opportunistic manner, but the merging between social community and firms (Becattini, 1990) helps preventing this result.

Farrell (2005) refer to an investment in *social capital* (Putnam, 2001; Putnam et al., 1993) as a key feature of industrial districts that inspire informal agreements, mutual trust and generalized reciprocity between the cluster's actors.

In this study we aimed to couple the model of an artificial cluster with a cognitive account of social evaluations. Reputation is a cognitive and social artifact rooted in individual minds but acting at the supra-individual level and evolved to solve collective problems. We adopted the socio-cognitive framework developed by Conte and Paolucci (2002), who describe how people create, manipulate and transmit social evaluations, and how these evaluations affect individuals' beliefs and behaviours. This approach is a dynamic one that considers reputation as the output of a social process that starts in agents' minds. Notably, this theory applies not only to humans, but also to artificial agents in a variety of distinct environments (Sabater et al., 2006). According to the theory, input to this process is evaluation that agents (Evaluators) directly

[1]In this work, *cluster* and *district* are used as synonyms.

form about a given Target during interaction or observation. This evaluation can be transmitted to Beneficiaries that share the goal with regard to which targets are evaluated and thus may use this information as a guide for their behaviour: knowing in advance others' behaviours and attitudes may thwart cheaters. The social and cognitive account of reputation proposed here allows one to:

1. distinguish between *image*, the output of a process of evaluation in which is made explicit who made that evaluation, and *reputation*, in which the source is impersonal. Image and reputation are both social evaluations regarding a Target, but they differ with regard to explicitness of the source. Evaluations from a nameless source can be less reliable, but they do not expose the gossiper to retaliation, as it happens with image.

2. account for the cognitive determinants of reputation and for its dynamic effects, both at the individual and at the collective level.

3. predict the agents' behaviours and resulting actions at the macro-level.

In what follows, I present a simulation model of reputation and its transmission in an artificial cluster of firms. The relevance of social evaluations in this context makes it suitable to verify the socio-cognitive theory of reputation, and to test whether and in what way the exchange of social information can be related to the quality of products delivered by artificial firms. The application of an agent-based computational approach to the study of industrial districts is not new (see Karlsson et al., 2005, for a review), but this study adds to this literature by using cognitive agents that manipulate and circulate two different kinds of social evaluations, i.e. image and reputation.

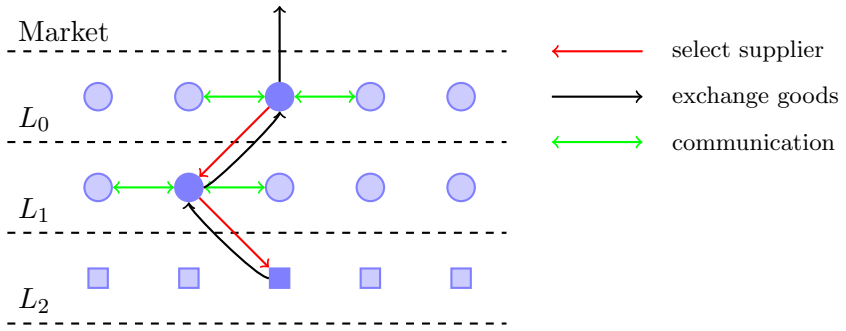

Figure 4.1: Agents interactions: firms select suppliers from the lower level, which in turn provide them components for their products, and communicate with firms belonging to the same level, exchanging evaluations about suppliers. L_2 firms are producers of raw materials and they do not communicate with each other; they are only chosen as suppliers by the firms of the layer above them.

4.2 Simulation model: gossiping about partners

The agents of the model are firms. They all produce components that are assembled into the only kind of final product sold in the market. Their goal is to select the best available supplier (in terms of goods' quality) in order to maximize their profits, and similar firms may collaborate with each other exchanging evaluations about known, tested suppliers (see Figure 4.1).

4.2.1 Partner selection and economic exchange

Firms are organised into different layers, with each layer containing agents that act as suppliers for the firms of the layer above. The number of layers can vary according to the characteristics of the cluster, but a minimum of two layers is required. Here, we have three layers (L_0 - L_1 - L_2), but n possible layers can be added, in order to develop a more complex production process. Final firms (L_0 agents) need one supplier from L_1, and the latter needs his own supplier from L_2, to assemble and sell the final product on the market. The market demand of the cluster's products is assumed to be fixed.

Firms differ in the quality of the goods they produce: $0.5 \leq Q < 1.0$; where $Q = 0.5$ indicates a very bad partner for interaction inside the cluster, and $Q = 1.0$ indicates a very good one. The average quality value, $Q = 0.75$, is the threshold the agents use to discriminate between a good and a bad supplier.

Firms buy components from suppliers at a fixed cost, $K = 0.75$ (thousands of euros), but the profits, U, they can make depend on supplier's quality: $U_{L_i} = f(Q_{L_{i+1}}) = F * Q_{L_{i+1}} - K$. Profits loss may be explained as if the bad quality components needs more work to be assembled and prepared for the final product.

Both L_0 and L_1 agents evaluate their suppliers, comparing the quality of the product they bought with the threshold value set at 0.75, and store these evaluations. If the product's quality exceeds that threshold, the supplier is considered good, otherwise it is labelled as a bad supplier. In an attempt to maximise their profits, firms always avoid interactions with bad suppliers, while trying to interact with the best known ones.

4.2.2 Evaluations exchange

The material exchange described above is paired in the model with an exchange of social evaluations: when the transmission of social evaluations is allowed, both leader firms and suppliers exchange information with their fellows regarding their suppliers from the level below, thus creating and taking part in a social network.

This process works only horizontally. There is no communication between agents that inhabit different levels of the cluster. Since agents exchange goods only with a specific set of suppliers, the information they acquire are only relevant inside their own level. Inside a layer, agents can play two possible

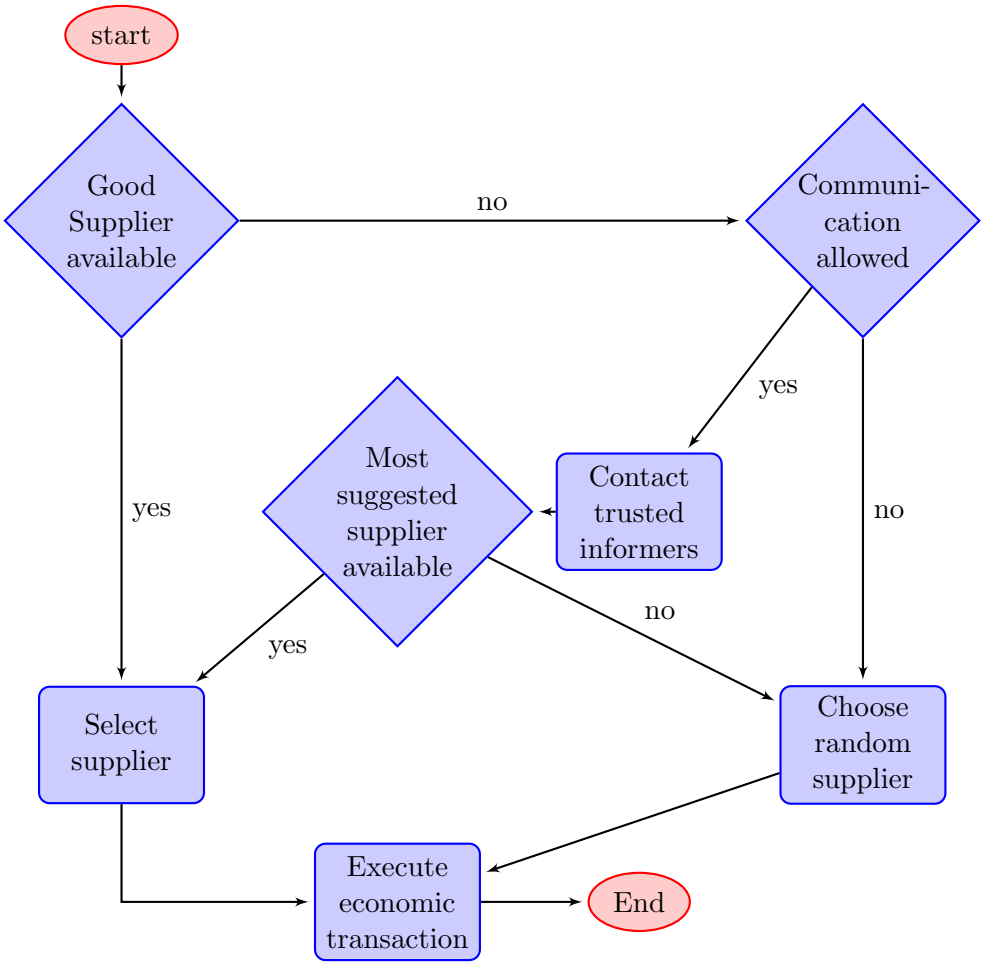

Figure 4.2: Flowchart of the simulation's cycle: at the beginning of every turn of the simulation, the agents of the first two level, L_0 and L_1, look for and select a partner for the economic interaction. The decision is first based on the availability of the best known partner. If unavailable, and the experimental condition allows the agents to exchange messages, they will consider information communicated by their peers, in order to establish a partnership with the supplier suggested by the highest number of peers. As an escape procedure, agents will select an unknown or random partner, regardless of his/her quality in the economic performances.

roles: (1) the *Questioner* asks an Informer, i.e. another firm of the same layer, to suggest a good supplier; (2) the *Informer* provides the ID of a good supplier. Honest informers suggest their best rated supplier, whereas cheaters transmit an evaluation concerning the worse supplier, as if it was a good one.

Partner selection can then be performed in three ways, listed in their priority order:

1. Experience-based Selection: the best rated supplier among those already tested is chosen.

2. Communication-based Selection: the most frequently suggested supplier by trusted Informers is selected.

3. Random Selection: among the unfamiliar suppliers (as an escape procedure).

All the relevant social information is stored by the agents in three different internal repositories, which are updated and checked at run-time to take decisions regarding both suppliers and informers.

Image Table As previously stated, firms store here the memories of their economic transactions (i.e. the actual quality value of each known supplier).

Candidates Table Identities of potential good suppliers —without reference to his quality— suggested by a fixed percentage of Informers among the same level (10% in the current implementation of the model) are aggregated here, either in the form of a direct evaluation (*image*) or a reported evaluation, i.e. in which the source is impersonal (*reputation*). To most frequently suggested agent is selected for economic transaction,

and after the transaction the information acquired are updated in the Image Table.

Informers Table Once the information about suggested suppliers are tested and moved to the Image Table, the Informers Table is updated with the ratings of the sources of the information. If a suggested supplier is found to be bad ($Q < 0.75$), the credibility of the Informers is compromised, the agents are categorised as cheaters and further communications from them are discarded.

The presence of cheaters in the cluster set up a social dilemma. Agents acting as suppliers are able to fulfill just one economic transaction at each simulation turn. Hence, giving away the identity of a good supplier, agents reduce the probability to interact with him in the future. False evaluations, on the other hand, have two main effects: they enhance the chances of advantageous economic transactions for the cheaters, keeping away the other firms from the good suppliers; and, at the same time, let the cheaters take advantages of the information received from truthful firms —information acquired without the costs of a possible bad economic transaction.

After a cheater is detected, cooperative firms adopt a retaliatory strategy: known cheating Questioners are provided with false evaluation even by cooperative Informers. Obviously, this behaviour depends on the type of evaluation circulating in the cluster. In the case of *reputation*, lacking an identifiable source, agents are not allowed to retaliate.

Hence, our main research question is whether the exchange of social evaluations —and what type of evaluations— can improve the economic performance of the cluster, when firms in the first two levels compete over high quality suppliers, and communication can be exploited by cheaters.

4.3 Simulation settings and results

We tested the agents' performance in terms of average quality of production, both for single layers and for the cluster as a whole. A cluster of 300 firms was so composed: 20% of agents for L_0, 40% of agents for L_1 and 40% for L_2. Quality was assigned randomly during the set up of each simulation's run, with values distributed normally between 0.5 and 1.0. During set up was also defined the behavioural trait of the agents relevant for the communication process: honest agents were the ones who cooperated with others with truthful information about available suppliers; a fraction of the population was instead composed by Informers who spread false information, agents that we called cheaters. In each experimental setting, a different number of agents in the population was be conferred the cheating trait.

We ran the experiments in three different conditions :

1. Control Condition (CC): no communication allowed. In this case, social information was not available and the suppliers' choice was exclusively experience-based. In this condition the behavioural trait relevant for communication are not considered: hence there are no cheaters in the population.

2. Image Condition (IC): agents exchanged true or false images. At the beginning of each simulation turn, agents in L_0 and L_1 were given the opportunity to collect information about available suppliers among their pairs. Contacted Informers replied according to their behavioural trait. The percentage of cheaters was set by a *cheating rate* parameter: the higher the value of the parameter the higher the number of cheaters in the cluster. Retaliation was possible: honest agents responded to cheaters, previously detected, with false information in order to punish

them.

3. Reputation Condition (RC): agents are still allowed to communicate at the beginning of each turn, however the messages exchanged by the Informers contain reputational information, i.e. evaluations without an explicit source. In this case, retaliation was not allowed, since the evaluator remains unknown.

4.3.1 Effects of communication

Comparing the Control Condition (CC) with the Image Condition (IC), we found that the possibility to communicate positively affects cluster's performance. In Figure 4.3, agents who are not allowed to communicate randomly explore the cluster and learn the identities of the best suppliers. Since the cluster is a closed environment, i.e. there is no change over time in its composition, this learning phase comes to an end once all the suppliers are being tested. There will still be competition for the good ones, but this will not affect the economic performance of the cluster.

A different pattern arising from agents interaction is observed in the IC: when agents of the same level are allowed to communicate with one another quality values increase more rapidly, because exploring the cluster in order to obtain higher profit requires less time. However, the presence of cheaters in the communication process can alter this effect, with a profit for the cheaters, but with a great damage to the cluster as a whole (see Figure 4.4).

4.3.2 Effects of social evaluations

Changing the type of evaluations exchanged among agents, cheating is still bad for the average product quality —especially for high levels of cheating rate. But in the Reputation Condition, as the Figure 4.5 shows, in the long

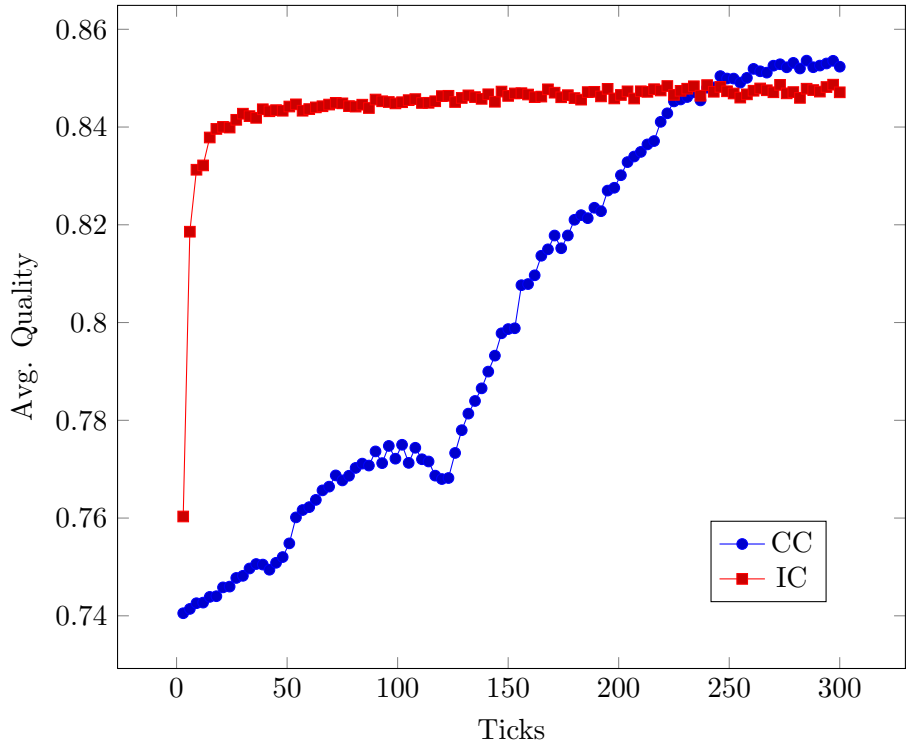

Figure 4.3: Average values of quality in the Image Condition (IC) compared to the Control Condition (CC). Simulations are performed ten times in each condition with a cluster of 300 firms.

run the cluster can absorb relatively high percentages of cheaters, without compromising its economic performance.

The communication algorithm proved to be robust with respect to the numbers of the agents in the cluster. What is really important is their distribution among the three levels, which in turn affects the availability of suppliers and the competition over them. When firms can chose among many suppliers (see Figure 4.6) we have no difference between IC and RC. When competition is hard, however, not only communication in RC performs better in the long run, but the all cluster obtain higher profits if compared to IC (see Figure 4.7).

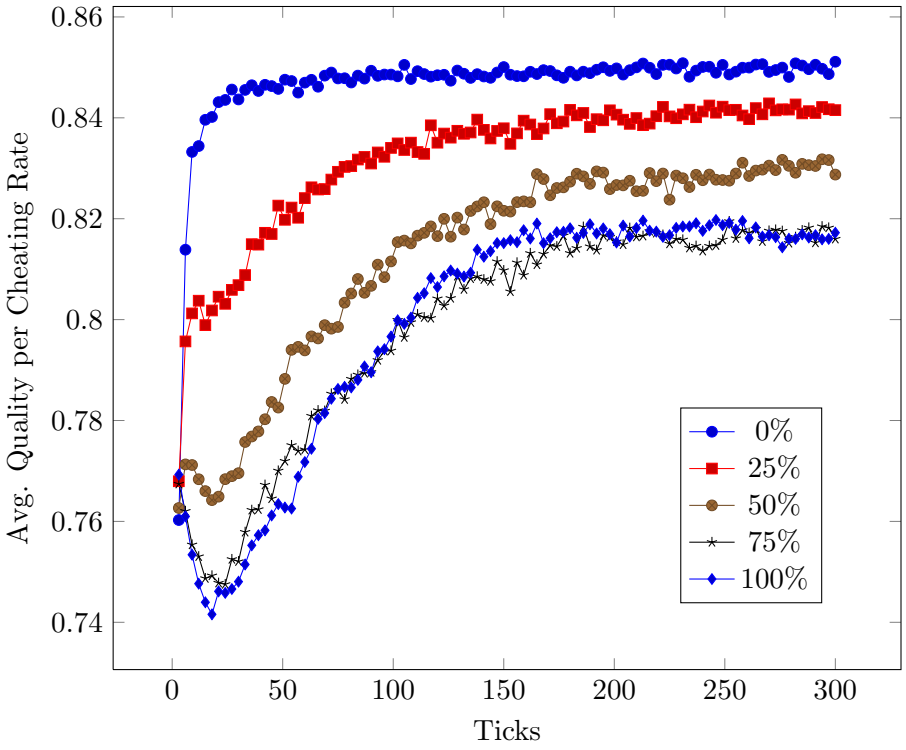

Figure 4.4: Average quality per cheating rate in the Image Condition (IC)

4.4 Conclusion

This study on partner selection sought to test the effects of two different kinds of social evaluations in an artificial cluster, adding to previous studies that applied the social and cognitive theory of reputation to other settings, both natural and artificial (Conte and Paolucci, 2002). We suggested that image and reputation, although closely related, are distinct objects, with different aims, functions and effects. We used the "small-world" of industrial clusters as a test bed of our theory, given the importance that reputational concerns have in this context. Material exchanges are usually supported and even improved by the social network of individuals and firms acting into a cluster: the merging of economic structure and social community makes the exchange

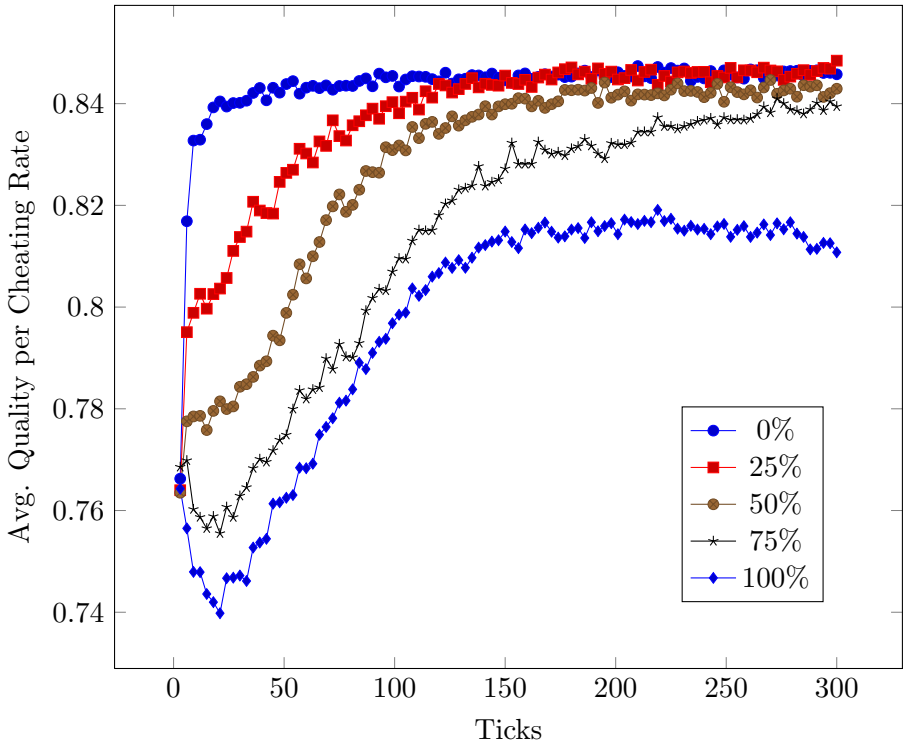

Figure 4.5: Average quality per cheating rate in the Reputation Condition (RC)

of social evaluations especially relevant to isolate cheaters, prevent frauds between cluster's actors and preserve quality of the single firms and of the entire cluster.

In order to test our predictions about the positive effects of communication on firms' economic performances, we designed an artificial cluster with companies grouped into three layers that trade products and exchange social evaluations. In this artificial cluster we tried to figure out how social information may affect the search for good partners and whether image and reputation make a difference to economic performances of both single firms and district as a whole. Our results showed that communication matters: compared to control condition, communication positively affected cluster's performance. Firms

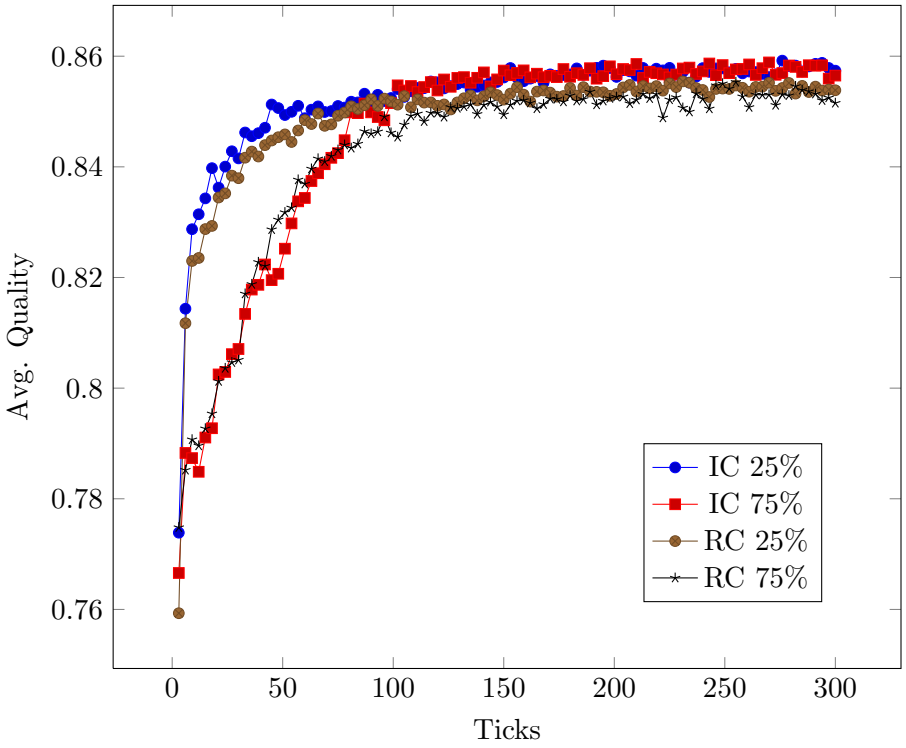

Figure 4.6: Average quality of 300 firms with the following distribution: $L_0 = 10\%$, $L_1 = 45\%$, $L_2 = 45\%$. Both IC and RC for 25% and 75% of cheating rate.

receiving reliable information about potential partners easily found good suppliers, compared to firms that were systematically cheated by their fellows. Furthermore, modelling reputation as rumors —i.e. evaluations where the source is unknown— we were able to preserve the benefits of communications for low level of cheating rate. In other words, reputation prevented retaliation, thus avoiding generalized punishment that would have lowered firms' profits.

We acknowledge that further improvements are needed, regarding both agents' refinement and cluster's structure. Although basic, our model allows one to verify theoretic predictions about the different effects of image and reputation and to relate them with the economic performance of an idealized industrial district. Future directions of this work will include introduction of

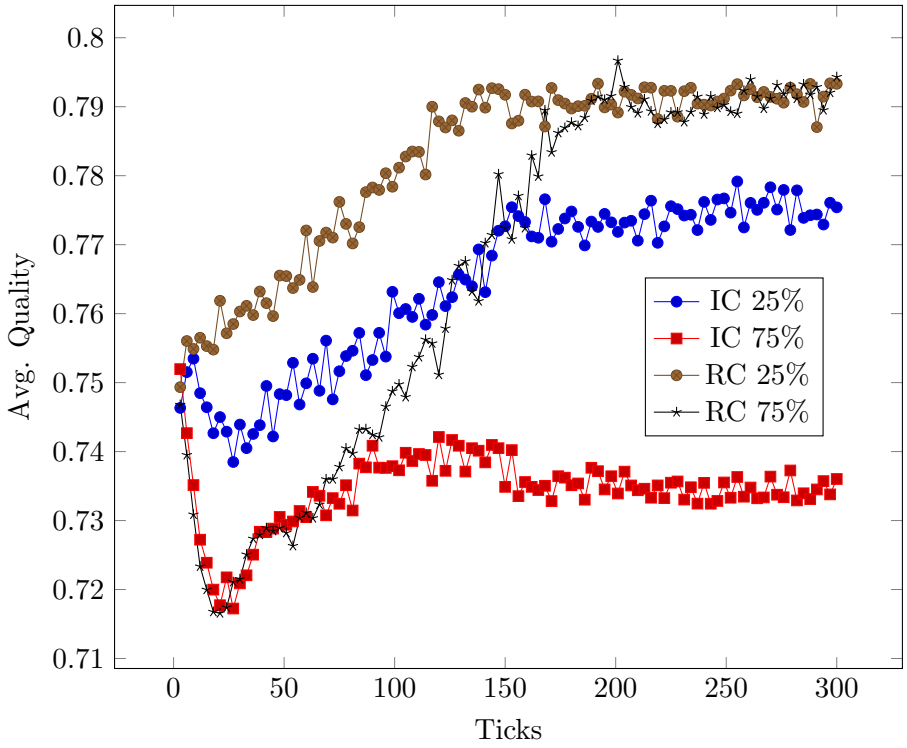

Figure 4.7: Average quality of 300 firms with the following distribution: $L_0 = 30\%$, $L_1 = 35\%$, $L_2 = 35\%$ (right), both IC and RC for 25% and 75% cheating rate.

communication flows between levels, refinement of firms' economic structure and testing for other social control mechanisms, as for instance ostracism. On the one hand, the lack of a true cognitive architecture prevented the possibility of exploring more interesting ways of implementing agents' decision making, since agents' strategies varied only according to the cheating rate. The hyper-simplified economic structure, however, allowed us to analyse what happens at the macro-level, linking it directly to agents' actions.

Chapter 5

Communication in the laboratory

In a different experimental setting —performed with groups of natural subjects interacting through a graphic computer interface— we analysed reciprocal forms of messages exchanges.

The positive effects of communication on rates of cooperation is a robust experimental finding. When individuals can talk to one other, cooperation increases significantly. Proposed explanations to this phenomenon consider the formation of group identity, as well as the chance to make explicit commitments —where reputational and moral factors come into play— as fundamental causes. This research, however, looks at communication as a means to establishing and enforcing cooperation among people; to our knowledge, no attempt has been made to analyse communication strategies, when commu-

nication processes are actually the place where cooperation evolve.

We developed a novel experimental setting in which participants playing a memory game (with numbers instead of images) could either play alone, or exchange messages containing the position and the value of the cards, so that those who received a truthful message could more easily get a match. This setting was conceptually modelled after the stag hunt game, a coordination game in which players do better if they coordinate their behaviour with the behaviour of others.

In our experiment, playing alone is faster than sending messages, but it leads to a quicker depletion of the available moves. On the contrary, sending messages is more time consuming, but it allows players to know the position of cards, provided the information is correct, without wasting moves in trying to guess.

Results show that the exchange of messages is a mutually beneficial activity, allowing participants to jointly discover the game board, to score higher and more efficiently. Cooperation through communication is conditional to receiving messages from other participants and is performed with this very expectation (as reported by the majority of the subjects afterward). This strategic behaviour could be explained according to two alternative frameworks: either a game-theoretical interpretation of reciprocity, analysed as an imitation strategy (Tit-For-Tat); or a cognitive view in which cooperative behaviour is regarded as a socially prescribed activity, and every deviation from the norm is punished according to the interpretation of the violation. The absence of retaliatory behaviour; and a tendency to exclude non-cooperative partners from further communication seems to exclude the possibility of an imitation strategy.

The work of this chapter is based on Giardini and Di Tosto (2007).

5.1 Introduction

Humans often adopt cooperative behaviours, and this tendency toward cooperation represents one of the most debated human features in a wide variety of disciplines. Cooperation is usually termed as a puzzling phenomenon (Boyd and Richerson, 2006; Noe, 2006), whose motives and mechanisms are still obscure. This is partially due to the lack of agreement on the definition of cooperation *per se*, which leads to a terminological and conceptual confusion between altruism, reciprocity and cooperation (Croson, 2008).

Noe (2006) identifies in the experimental literature three main uses of the term: cooperation as a certain type of interaction with a specific form or outcome, as a strategy used by members in an interaction, or even as a characteristic of a long-term relationship. Here, we focus on the meaning of cooperation as a strategy people adopt in a coordination game, and we will investigate how normative reasoning can pave the way for the emergence and the maintenance of cooperation.

The positive effects of communication on rates of cooperation is a robust experimental finding. When individuals can talk to one other, cooperation increases significantly. Proposed explanations to this phenomenon consider the formation of group identity (Kollock, 1998), as well as the chance to make explicit commitments (Kritikos and Meran, 1998) —where reputational and moral factors come into play (Milinski et al., 2002)— as fundamental causes. This research, however, looks at communication as a mean to establishing and enforcing cooperation among people; to our knowledge, no attempt has been made to analyse communication strategies, when communication processes are actually the place where cooperation evolve.

We developed a novel experimental setting in which participants playing a memory game (with numbers instead of images) could either play alone,

or exchange messages containing the position and the value of the cards, so that those who received a truthful message could more easily get a match. This setting was conceptually modelled after the *stag hunt game* (Skyrms, 2001), a coordination game in which players do better if they coordinate their behaviour with the behaviour of others. In our setting, playing alone is like hunting hare, a solitary activity that leads to small payoffs, whereas exchanging relevant information is analogous to hunting stag. In the latter case, players engage in a mutually beneficial activity whose results are faster exploration of the game board and the resulting higher probability of getting a match. In our experiment, playing alone is faster than sending messages, but it leads to a quicker depletion of the available moves (for further details see Section 5.2). On the contrary, sending messages is more time consuming, but it allows players to know the position of cards, provided the information is correct, without wasting moves in trying to guess. Moreover, participants can decide how many addressees their message can have, choosing from three alternatives: group message, private message and sub-group message. Whether one or many receivers are selected depends upon the preferred strategy, which is conditional to the expected contribution of other players.

In this framework, the exchange of valuable information is a costly action, analogous to other costly behaviours normally considered inside the game-theoretic literature. We are thus concerned with whether and how cooperative behaviour gives rise to the emergence of a norm which sustains and promotes cooperation, including the punishment of those who do not cooperate. Our hypothesis is that cooperative behaviour is mediated by normative reasoning that leads people to cooperate, in our case by sending group messages, and to expect cooperation in return. When this expectation is not fulfilled, subjects punish those who do not cooperate, by excluding them from communication.

More specifically:

- Communication through group messages is a mutually beneficial activity, allowing participants to jointly discover the game board, to score higher and more efficiently.

- Cooperation through communication is conditional to receiving messages from other participants. This strategic behaviour could be explained according to two alternative frameworks (Conte and Castelfranchi, 1995):

  - game-theoretical interpretation of reciprocity, analysed as an imitation strategy (Tit-For-Tat).

  - a cognitive view in which cooperative behaviour is regarded as a socially prescribed activity, and every deviation from the norm is punished according to the interpretation of the violation.

Following these two alternative hypotheses, we expect subjects to show different reactions to non-cooperative use of the communication process. If the Tit-For-Tat interpretation holds, then we should observe subjects reciprocating by means of imitation (i.e. returning false information to sender of false information, and refraining from cooperation with people who do not communicate). In the other case, the cognitive interpretation of the normative prescription predicts a different behaviour: non-cooperators are excluded from the benefits of cooperation.

The experiment was conducted using a customized computer graphical interface featuring a mechanism to send and receive messages to and from other participants (see Figure 5.1). Computer mediated communication is known to be less effective than face-to-face communication in the establishment of cooperation (Brosig et al., 2003), even though, in the present study, it offered

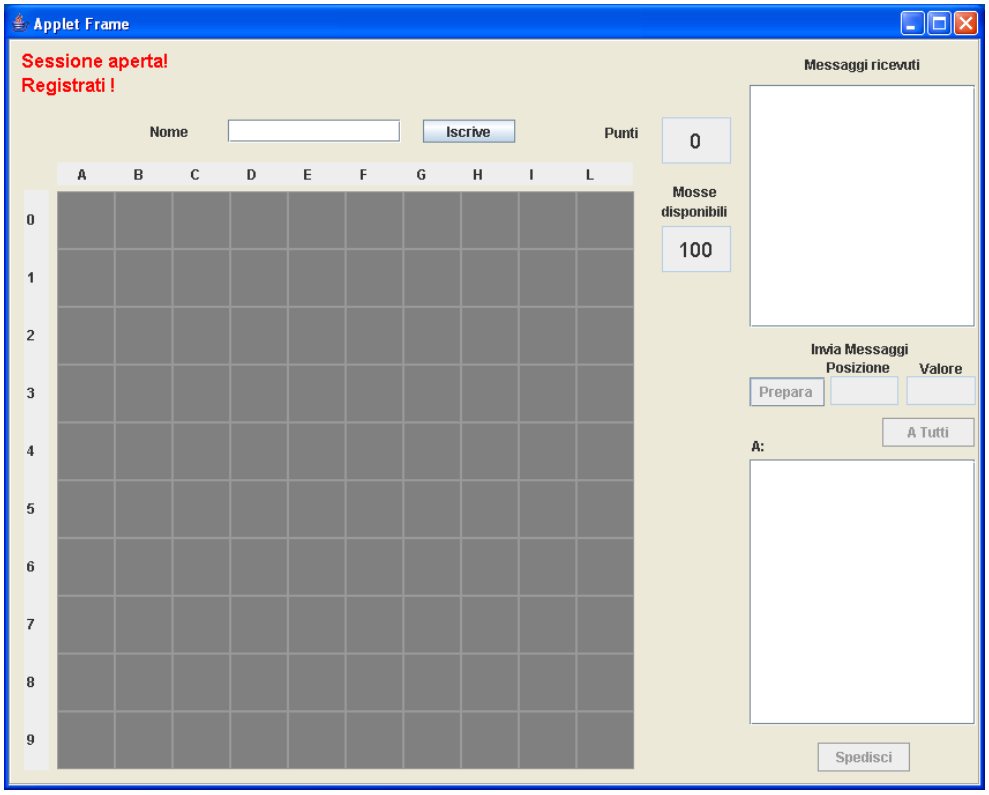

Figure 5.1: The customized graphical interface, with the game board on the left side, the window with the list of the received messages and the nickname of the sender (upper right side), the list of the subjects available for communication (lower right side), and the form for composing and sending messages (at the center of the right side). The interface also inform the subjects about their score and the number of moves left.

the possibility to control and monitor the information flow, to reduce the complexity and ambiguity of natural language and to clearly distinguish between false and truthful communication.

We acknowledge from the start that our study is exploratory and that further studies are needed to put forward firm conclusions about the effect of normative reasoning on cooperation in a coordination game setting. This is a critical issue which deserves a deeper theoretical analysis as well as more experimentally grounded results.

5.2 Experiment

To test whether and how people cooperate by exchanging relevant information, we designed a novel experimental setting using the well-known memory cards game. Players had to find the greatest number of matching pairs among all cards placed face-down on the game board. We modified the game, adding two constraints:

Moves limit the subjects were allowed 100 moves to play the game; two cards selected equaled one move. Once the available moves were exhausted, the game ended.

Time limit each session lasts a maximum of 10 minutes. Players are not informed about the exact length of the experiment.

Card selection and communication processes were concurrent activities; participants decided autonomously whether to communicate with other subjects or to play alone. The game and the communication platform were implemented on a client-server software architecture. On the client side, the user interface is composed of:

- A board of 50 pairs of cards, numbered form 0 to 49, placed in a 10 x 10 matrix and identified by their coordinates (LETTER; number).
- The message preparation form, composed of a list of the other participants,from whom the user could select the receiver(s); two boxes comprising the content of the communication (that is, the card's coordinates and value); and a button to send the messages.
- The incoming message frame, displaying a list of the messages received by the subject, each message being a pair of *card's coordinates* and *card's value*, plus the *nickname* of the sender.

- A display of the score —the number of matching cards discovered in the board.

- A display of the number of moves remaining. When the player(s) use up those 100 moves, the computer program automatically blocked that player's session, and the subject was prevented from continuing the game and also from sending messages.

5.2.1 Participants

Twenty-four students from the University of Siena, Italy (11 male, 13 female), mean age 22.71 ($SD = 2.19$), volunteered to participate in the experiment and were paid a 5 euro show-up fee, along with their earnings for participation in the experiment for one hour of experimental time (mean earnings= 7.90 Euros, plus the show-up fee). Subjects were assigned randomly to one of 4 groups of six subjects. Each subjects participated in three sessions of the game. They were naive with respect to the nature and aims of the experiment.

5.2.2 Procedure

5.2.2.1 The game

The subjects were assigned to one of four groups; each group was tested separately along three 10 minute sessions (with a 5 minutes interval between them and a short training session at the very beginning, lasting 3 minutes). The participants, registered in the system using nicknames to preserve anonymity, played a memory card game together, on different computers but with the same game board. At the beginning of the game, all cards were laid face-down, and subjects were provided with 100 moves to explore the board: one mouse-click revealed the value of the chosen card. Each time they uncovered

a pair of cards with the same value, one point was added to their score and the matching cards remained face-up on that player's board.

5.2.2.2 Communication

During the exploration of the board, subjects could send information about uncovered cards through the interface. No other forms of communication were allowed. The designed procedure required them to activate the message preparation phase, then select the coordinates of a card by mouse-clicking on it and entering its value. The system did not check the content of the communication, so participants had the possibility of sending false messages, as either an intentional lie or an unintentional typing error. To send the message, subjects should select one, more than one, or even all of the possible recipients from the list of active participants, and then press the send button, generating a *Private*, *Sub-group* or *Group* message, respectively. All messages then appeared in the incoming message frame of the receiver(s), in chronological order and until the end of the session.

5.2.3 Questionnaires

After the three experimental sessions, participants responded to a questionnaire designed to assess the utility, from the standpoint of participants' perception, of the messages and highlight the emergence of communicative strategies. The questionnaire comprised 4 questions (1. Did you send at least 5 messages to the other participants? Yes or No; 2. Provided you sent messages, they were: a. mostly true; b. mostly false; c. always true; d. always false; 3. Provided you sent messages, these were addressed: a. to everybody; b. to someone on purpose; c. to someone by chance; d. to those who already sent me a message.; 4. Do you believe that receiving messages affected your final

score? Yes or No), plus a blank space where subjects were required to briefly explain why they exchanged messages with other players.

5.3 Results

The results showed that subjects' mean score (see Figure 5.2) significantly increased between the first and second session for each group of subjects (Wilcoxon rank sum test, $Z = 402,\ p = 0.009$), and between first and third session (). We found no significant differences between the second and third session ($Z = 342,\ p = 0.134$).

The average number of sent messages (see Figure 5.3) increased all along the experiment. On the contrary, no significant difference was observed in the number of moves used by subjects over the different sections (J-T test[1], $trend\ p - value = 0.984,\ rank\ correlation = -0.002$).

Figure 5.4 shows a significant change in the communicative strategy of subjects. In the first session they almost equally used single ($M = 1.37$; $SD = 2.93$), group ($M = 1.41$; $SD = 2.63$) and subgroup messages ($M = 2.41$; $SD = 2.60$), whereas they decisively turned to the group message in the second (Single $M = 0.87$ with $SD = 1.96$; Subgroup $M = 0.91$ with $SD = 1.66$; Group $M = 5.83$ with $SD = 2.83$) and in the third session (Single $M = 0.91$ with $SD = 2.82$; SubGroup $M = 0.58$ with $SD = 1.10$; Group $M = 7.29$ with $SD = 6.8$).

The percentage of false messages was extremely low, as showed in Figure 5.5. The highest number of false messages was observed in Group 3. In the same group we found the greatest number of Sub-group messages (see Figure 5.6).

[1]Performed using the R package SAGx

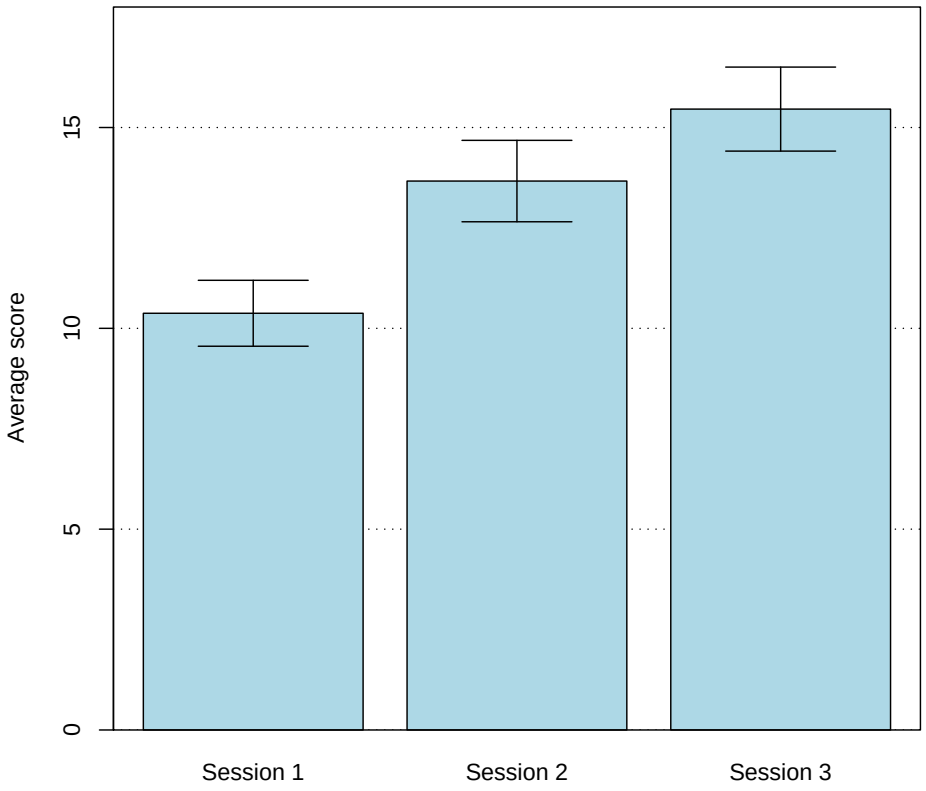

Figure 5.2: Average score of each session among all the groups of subjects ($N = 24$). Average scores in the second and third sessions are significantly higher than the one of the first session.

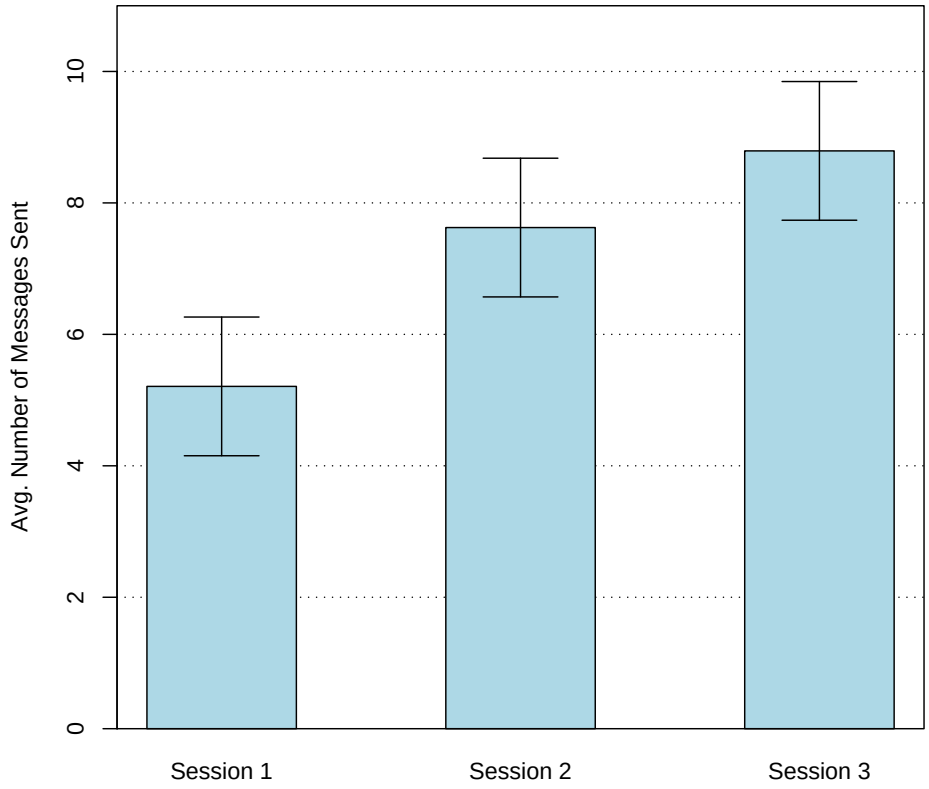

Figure 5.3: Average number of sent messages along the three sessions, $N = 24$.

A chi-squared test of independence was performed to examine the relationship between score and received messages. The relation between these variables was significant, $\chi^2(2,\ N = 24) = 18.95$, $p = 0.025$. There was a positive relationship between score and number of received messages. A second chi-squared test of independence was performed to examine the relationship between received messages and sent messages. The relation between these variables was significant, $\chi^2(2,\ N = 24) = 28.79$, $p < .001$.

The analysis of the responses to the questionnaire showed that all sub-

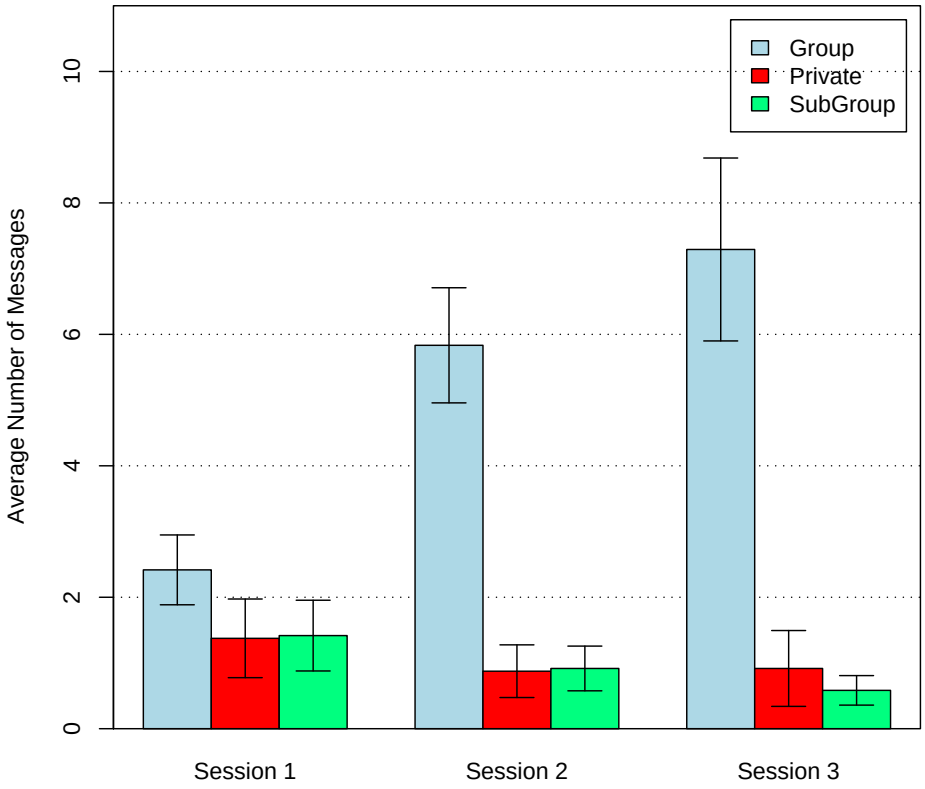

Figure 5.4: Group, Sub-group and Private messages' average number in each session.

jects sent at least 5 messages. The great majority (90%) responded they used always true messages, whereas some (10%) responded mostly true messages. Group messages prevailed (90%) and all subjects agreed about the existence of a correlation between received messages and their score. We grouped the responses about personal motives for communication into two classes: conditional and unconditional cooperation. The 70% of participants responded they sent messages in order to receive messages from other players, whereas the remaining 30% reported that they communicated with the aim to help

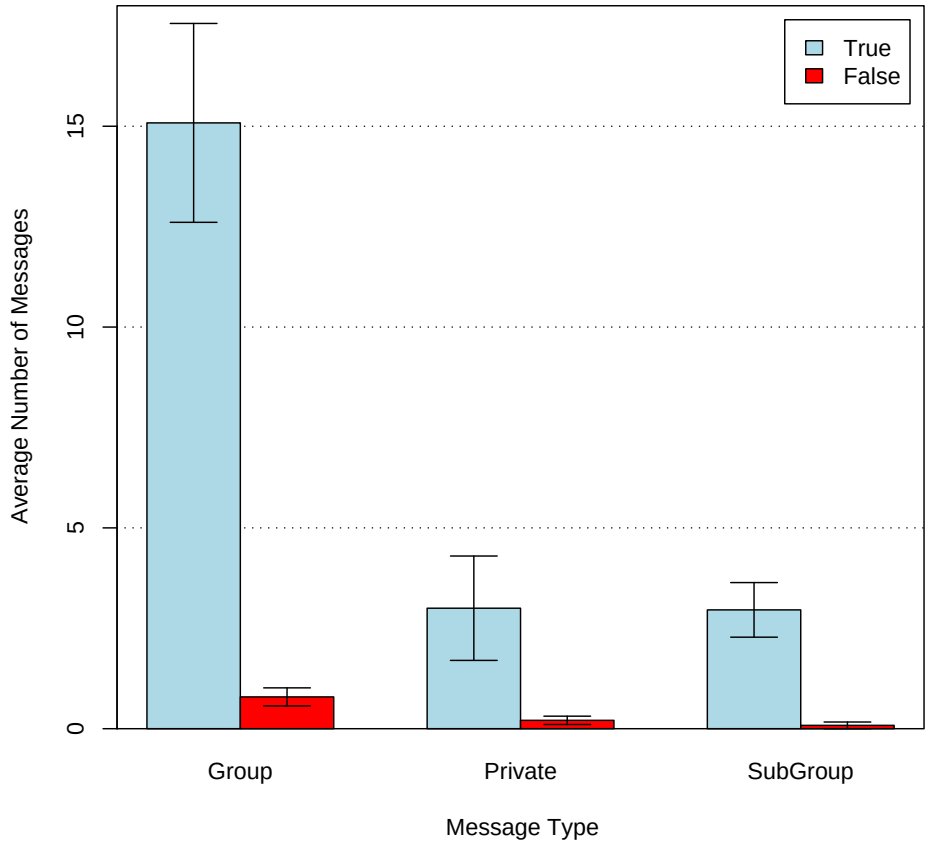

Figure 5.5: Distribution of true and false messages: Average number of true and false Group, Sub-Group and Private messages.

other people in getting matches.

5.4 Discussion

We showed how cooperation through communication emerged and evolved in a coordination game. In the memory game, the possibility to exchange relevant information with other players put the participants in front of a social dilemma, where both cooperating and playing alone were costly. Sending

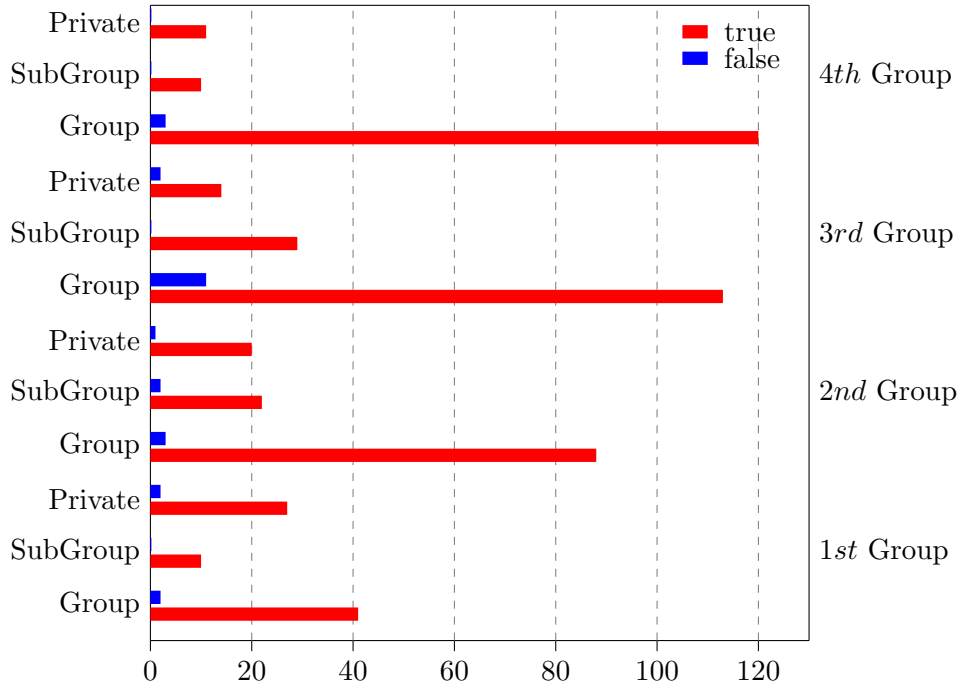

Figure 5.6: Distribution of true and false messages: Absolute number of true and false messages, divided by type (Group, SubGroup and Private Messages), among the four groups of subjects.

messages was a useful tool for saving moves and, at the same time, quickly uncovering the game board, provided all cooperate. On the contrary, playing alone was less time consuming, and it was a better strategy when others did not cooperate.

We are not able to rule out the possibility of an improvement in the subjects' performance due to a learning effect. It is plausible to postulate the existence of a hidden cause related to mnemonic skills. Our results, however, show that participants preferred cooperation, and that this choice was rewarding, as demonstrated by the positive dependence between average score and received messages; communication seems at least equally important in the interpretation of the observed behaviour.

Moreover, the prevalence of truthful communication can be interpreted as

further evidence of cooperative behaviour, since participants did not try to negatively interfere with others by sending false messages, but mostly sent true messages. In this setting, cooperation is an equilibrium whose emergence is favored by cooperative behaviour, and whose maintenance requires a cognitively mediated form of reciprocation.

Game-theoretic approaches predict a Tit-for-Tat strategy in these cases, but in the third group of subjects we found a different behaviour. Noncooperators (liars or people avoiding communication) were not directly retaliated against, but they were excluded from communication: subjects who spread false information were excluded form further communication —but were not the addressee of following false messages— as well as subjects who refused to share their information; that is the reason why we observed the formation of sub-group, where cooperative participants coordinate playing the game. This is further confirmation of the importance of cooperation, which was actively pursued but also withheld from noncooperators. Our findings support the hypothesis that, to preserve cooperation from the exploitation of non-cooperative people, strategic interaction does not rely upon a mere mirroring behaviour, but it requires a fully cognitive means-ends reasoning. Unfortunately, we observed this behaviour in only one group, thus new investigations and a greater number of subjects are required to fully validate this hypothesis.

In responding to the questionnaires, people recognized the importance of cooperation, and the great majority of them explicitly declared that they sent messages in order to receive back messages from other players (as reported by many subjects).

Because this is an exploratory study that utilizes a novel setting, some adjustments are needed. For instance, the computer interface might be a bit tricky. To send a message subjects had to activate the message preparation

phase in order to select coordinates and value of a card and, finally, to select one or more (even all) recipients from the list of active participants. Another potentially influential point was the difference between the time and the effort required to select one or more participants, compared to the ease of sending a message to all addressees. In fact, in the latter case, subjects had an *All* button that automatically selected all other players.

Nonetheless, this study adds to the discussion of cooperation dynamics and raises important questions, especially when considered together with the findings of the previous chapters.

Chapter 6

Conclusions

This dissertation revolved around the design of communication mechanisms for the study of reputation formation in artificial societies, with reputation defined as the effect of the spreading of social evaluations. The experiments addressed problems derived from the study of cooperative behaviour, where an autonomous agent helps other agents at a cost for him or herself, expecting cooperation in return, and refrains from helping others if they behave opportunistically. When the helping behaviour is returned by an agent different from the one who benefited from it in the first place, it generates a chain of cooperative behaviours, usually referred to as *indirect reciprocity*.

The first study, reported in Chapter 3, analysed if and how communication of evaluations about other members of a population can preserve this chain of cooperation. In a large population, non-cooperative agents can be difficult to identify, and they can prosper by exploiting the help of unaware individu-

als, while refusing to pay it back. Communication can make non-cooperative agents be preceded by their reputation, informing cooperators about their potential partners even when they have never before interacted. Thus, cooperators are able to avoid exploitation, or can choose to seek social interaction only with partners who are in good standing in the society.

With the use of agent-based methodology, the simulations were able to reproduce these analytical predictions about reputation by having agents exchange messages regarding the behaviour of their partners after each interaction. Two types of communication mechanisms were tested: one with public messages, in which information was shared among all agents of the population; the other with private messages, where messages were addressed only to known cooperating agents. And, although communication via private messages is a less efficient way to spread evaluations, it proved effective in preserving cooperation when agents had the ability to choose whom they are interacting with in the next round of the game, instead of being matched at random.

Chapter 4 then presented a second study that focused on the spreading of evaluations, which, as we have seen, leads to the formation of reputation and thus, eventually, to various levels of cooperation within a group. Here the social simulation modelled the spreading of social evaluations as gossip, or rumors, i.e. third-party evaluations about a given target, where the source of the evaluation is not identified. As explained in Chapter 2, this peculiarity of gossip is what makes reputation a difficult subject to study: it is a property of social agents with a dynamic that is partially beyond the control of the single individual. Social evaluations shared through gossip can be modelled as a second order belief: agents can assume the first-order evaluation to be true and plan their actions accordingly; or they can rely on second-order beliefs, trusting gossipers who, in reality, can pass on evaluations of this kind

regardless of their truth value. By comparing gossip about reputation with the communication of regular, direct evaluations (here referred to as image) in a task of partner selection —where agents compete with their peers to interact with partners of different quality— simulations results showed that, even if image exposes the source of the evaluation to possible future retaliation from the recipient of a false messages, evaluations spread as gossip, despite the presence of possible informational cheaters, were more effective in selecting valuable partners for the interaction.

In the third study, presented in Chapter 5, communication was analysed as a cooperative device, assuming reciprocatory behaviour to apply when subjects are exchanging relevant information during the performance of the assigned task. Here the messages do not contain social evaluations and thus are not intended to sanction a particular social agent; their sole purpose is to help other participants of the experiments score better results. When given the opportunity to choose between performing the task alone and cooperating with others by sharing information, the great majority of the subjects recurred to communication, and —as they reported— decided to do so with the expectation of receiving messages in return. In the case where subjects were confronted by dishonest individuals, they did not resort to retaliation using false massages; instead they excluded liars from further communication. This behaviour indicates that communication intended as a common good can be better preserved by means of ostracism, than by retaliation. Such a tendency could have great implications for further study of communication, as well as in fields involving commercial trade and other transactions.

6.1 Further directions of research

The effective use of reputation and communication for strategic behaviour is a cognitively demanding task. A natural extension of the work presented in this dissertation consists in the adoption of a suitable framework for the implementation of cognitive agents, which would get us further in the exploration and testing of the theoretical analysis of reputation illustrated in Chapter 2. In order to incorporate all aspects of this theory into the simulations, a more complex framework for modelling them is required. The studies discussed here aimed to model agent interactions as they communicated beliefs, but decision-making processes (strategic, epistemic and memetic; see Section 2.2.3) require the implementation of agents endowed with goals and planning capabilities, in addition to beliefs, in order to design a model that accounts for more of the complexities of cognition.

One such possibility is represented by Belief-Desire-Intention (BDI) agents.

Inspired by the theory of practical reasoning (Bratman, 1987), BDI agents constitute a standard for the implementation of cognitive agents at the symbolic level (in the same way neural networks are a standard for implementing agents at the sub-symbolic level). BDI agents should be considered for an enhancement and further research because they offer a number of advantages over simpler, rule-based agents. BDI offers:

- a more complex structure for representing an agent's knowledge base.
- the possibility to model more accurately the epistemic state behind a particular social behaviour.
- as an implementation for practical reasoning, BDI support planning, which in turn would allow for an explicit analysis of the goal of agents' behaviour.

Among the possible directions for future research, I'm going to sketch below two of the most promising.

Agents' sensitivity to gossip Agents can differ either in their typology (more or less sensibility to social judgment) or in terms of their goals (goals themselves and the agents' ability to achieve them). The dependence of one agent on another (Castelfranchi et al., 1992; Sichman et al., 1994), can compromise the first agents ability to fulfill a goal at least two potential reasons: lack of capabilities, or lack of resource — the first agent requires something from the second. No matter the reason, dependent relationships and their implied power dynamic (Castelfranchi, 1990) carry a risk in terms of social evaluations: negative evaluations regarding the first agent could prevent the second agent from interacting with him or her. In other words, for the goals for which an agent depends more on others, he or she will be particularly sensitive to reputation (while for independently-achievable goals, reputation may be a less important factor for the agent).

Strategic use of communication The studies discussed in this dissertation raise two particularly intriguing questions regarding the strategic use of communication: first, how can we "convince" others to cooperate with us? And, conversely, how can deceit be used to persuade other agents into cooperation?

This second question leads us to the definition of two kinds of intentions, which we can identify in terms of the third-party to whom the signal is addressed:

1. Signal for beneficiaries/gossipers: the intention of being recognised as a loyal in-group member, accomplished by sharing relevant information (a

low cost action) while violating social norms or prescriptions when we don't risk being discovered.

2. Signal for the group of targets of a given evaluation: the use of gossip as a threat, that is, the preventative use of gossip, spreading rumors about the social category to which the target belongs with the intention of influencing his/her future behaviour.

Bibliography

Alexander, R. D. (1987). *The Biology of Moral Systems (Foundations of Human Behavior)*. Aldine.

Ashlock, D., Smucker, M., Stanley, E. A., and Tesfatsion, L. (1996). Preferential partner selection in an evolutionary study of prisoner's dilemma. *BioSystems*, 37(1-2):99–125.

Axtell, R. L. and Epstein, J. (2006). Coordination in transient social network: an agent-based model of the time of retirement. In Epstein, J., editor, *Generative Social Science*. Princeton University Press.

Axtell, R. L., Epstein, J. M., Dean, J. S., Gumerman, G. J., Swedlund, A. C., Harburger, J., Chakravarty, S., Hammond, R., Parker, J., and Parker, M. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 3:7275–7279.

Back, I. and Flache, A. (2006). The viability of cooperation based on inter-

personal commitment. *Journal of Artificial Societies and Social Simulation*, 9(1). http://jasss.soc.surrey.ac.uk/9/1/12.html.

Bateson, M., Nettle, D., and Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, 2(3):412–414.

Becattini, G. (1990). *The marshallian industrial district as socio-economic notion*. International Institute of Labour Studies.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *PNAS*, 100(6):3531–3535.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. The University of Chicago Press.

Boyd, R. and Richerson, P. J. (2006). Culture and the Evolution of the Human Social Instincts. In Enfield, N. J. and Levinson, S. C., editors, *Roots of Human Sociality. Culture, Cognition and Interaction*. Berg, Oxford.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, Mass.

Brosig, J., Weimann, J., and Ockenfels, A. (2003). The Effect of Communication Media on Cooperation. *German Economic Review*, 4(2):217–241.

Castelfranchi, C. (1988). *Che Figura!* Il Mulino, Bologna.

Castelfranchi, C. (1990). Social power. In Demazeau, Y. and Muller, J. P., editors, *Decentralized AI – Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 49–62. Elsevier.

Castelfranchi, C., Miceli, M., and Cesta, A. (1992). Dependence relations among autonomous agents. In Werner, E. and Demazeau, Y., editors, *Decentralized AI 3: Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 215–231. Elsevier.

Coleman, J. S. (1990). *Foundations of Social Theory*. The Belknap Press of Harvard University Press.

Conte, R. (1999). Social intelligence among autonomous agents. *Computational and Mathematical Organization Theory*, 5:202–228.

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

Conte, R. and Paolucci, M. (2002). *Reputation in Artificial Societies: Social Beliefs for Social Order*. Springer.

Croson, R. T. A. (2008). Differentiating altruism and reciprocity. In Plott, C. R. and Smith, V. L., editors, *Handbook of Experimental Economics Results*. Elsevier, Amsterdam.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

de Vos, H., Smaniotto, R., and Elsas, D. (2001). Reciprocal altruism under conditions of partner selection. *Rationality and Society*, 13(2):139–183.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424.

Di Tosto, G., Giardini, F., and Conte, R. (2007). Enforcing prosocial behaviour. In Amblard, F., editor, *Proceedings of Essa'07 the 4th Conference*

of the European Social Simulation Association, pages 597–607. Toulouse, IRIT Editions.

Doebeli, M. and Hauert, C. (2005). Models of Cooperation based on the Prisoner's Dilemma and the Snowdrift Game. *Ecology Letters*, 8:748–766.

Dunbar, R. (2004). Gossip in Evolutionary Perspective. *Review of General Psychology*, 8(2):100–110.

Elias, N. (1974). Towards a theory of communities. In *The Sociology of Communities. A Selection of Readings*. Frank Cass & Co, London.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12. http://jasss.soc.surrey.ac.uk/11/4/12.html.

Farrell, H. (2005). Trust and political economy: Institutions and the sources of interfirm cooperation. *Comparative Political Studies*, 38(5):459–483.

Fehr, E. and Gachter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives*, 14(3):159–181.

Fehr, E. and Gachter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Giardini, F. and Di Tosto, G. (2007). Cooperation through communication: Reciprocal exchange of relevant information among humans involved in strategic interactions. In Vosniadou, S., Kayser, D., and Protopapas, A., editors, *Proceedings of EuroCogSci07 - The European Cognitive Science Conference*, pages 190–195. Lawrence Erlbaum Associates.

Giardini, F., Di Tosto, G., and Conte, R. (2008a). A model for simulating reputation dynamics in industrial districts. *Simulation Modelling Practice and Theory*, 16(2):231–241.

Giardini, F., Di Tosto, G., and Conte, R. (2008b). Reputation and economic performance in industrial districts: Modellig social complexity through multi-agent systems. In Cioffi Revilla, C. and Deffuant, G., editors, *The Second World Congress on Social Simulation (WCSS08)*, George Mason University, Fairfax, VA, USA, July 14–17, 2008.

Gintis, H. (2007). A framework for the unification of the behavioral sciences. *Behavioral and Brain Sciences*, 30(1):1–61.

Gintis, H., Smith, E. A., and Bowles, S. (2001). Costly Signaling and Cooperation. *Journal of Theoretical Biology*, 213(1):103–119.

Gluckman, M. (1963). Papers in honor of Melville J. Herskovits: Gossip and scandal. *Current Anthropology*, 4(3).

Grafen, A. (1998). Evolutionary biology: Green beard as death warrant. *Nature*, 394(6693):521–522.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3).

Hales, D. and Edmonds, B. (2003). Evolving Social Rationality for MAS using "Tags". In Rosenchein, J. S., Wooldridge, M., Sandholm, T., and Yokoo, M., editors, *Proocedings of the 2nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 03)*, pages 497–503. ACM Press.

Hales, D. and Edmonds, B. (2005). Applying a socially-inspired technique (tags) to improve cooperation in p2p networks. *IEEE Transactions in Systems, Man and Cybernetics – Part A: Systems and Humans*, 35(3):385–395.

Haley, K. J. and Fessler, D. M. T. (2005). Nobody's watching? subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26:245–256.

Hauk, E. (2001). Leaving the prison: Permitting partner choice and refusal in prisoner's dilemma games. *Computational Economics*, 18(1):65–87.

Henirch, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53:3–35.

Henrich, J. and Gil-White, F. J. (2001). The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196.

Hirshleifer, D. and Rasmusen, E. (1989). Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior and Organization*, 12:87–106.

Hruschka, D. and Henrich, J. (2006). Friendship, cliqueness, and the emergence of cooperation. *Journal of Theoretical Biology*, 239:1–15.

Huynh, T. D., Jennings, N. R., and Shadbolt, N. R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154.

Kalick, S. M. and Hamilton, T. E. (1986). The matching hypothesis reexamined. *Journal of Personality and Social Psychology*, 51(4):673–682.

Karlsson, C., Johansson, B., and Stough, R. (2005). *Industrial Clusters And Inter-firm Networks (New Horizons in Regional Science Series)*. Edward Elgar Publishing.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Reviews of Sociology*, 24:183–214.

Kritikos, A. and Meran, G. (1998). Social norms, moral commitment, and cooperation. *Homo Oeconomicus*, 15:71–92.

Marsh, S. (1992). Trust and reliance in multi-agent systems: A preliminary report. In *MAAMAW '92, 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, Rome, Italy.

Marsh, S. (1994a). Formalising trust as a computational concept. Technical report, Dept. of Computing Science and Mathematics, University of Stirling.

Marsh, S. (1994b). Optimism and pessimism in trust. In *Proceedings of the Ibero-American Conference on Artificial Intelligence (IBERAMIA-94)*.

Miceli, M. and Castelfranchi, C. (2000). The role of evaluation in cognition and social interaction. In Dautenhahn, K., editor, *Human cognition and agent technology*. John Benjamins Publishing Company, Amsterdam, NL.

Milinski, M., Semmann, D., Bakker, T. C. M., and Krambeck, H.-J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? *Proceedings of the Royal Society of London B*, 268:2495–2501.

Milinski, M., Semmann, D., and Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, 415(6870):424–426.

Moukas, A., Zacharia, G., and Maes, M. (1999). Amalthaea and histos: Multi-agent systems for www sites and reputation recommendations. In Klusch, M., editor, *Intelligent Information Agents. Agent-Based Information Discovery and Management on the Internet*, pages 292–322. Springer, Berlin.

Mui, L., Halberstadt, A., and Mohtashemi, M. (2002). Notion of reputation in multi-agent systems: A review. In *Proceedings of the AAMAS Conference*, pages 280–287.

Noe, R. (2006). Cooperation experiments: coordination through communication versus acting apart together. *Animal Behaviour*, 71(1):1–18.

Nowak, A., Szamrej, J., and Latané, B. (1990). From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact. *Psychological Review*, 97(3):362–376.

Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577.

Orbell, J. M. and Dawes, R. M. (1993). Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review*, 58:787–800.

Panchanathan, K. and Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, 432(7016):499–502.

Putnam, R. D. (2001). *Bowling Alone : The Collapse and Revival of American Community.* Simon & Schuster.

Putnam, R. D., Leonardi, R., and Nanetti, R. Y. (1993). *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton University Press.

Ramchurn, S. D., Huynh, D., and Jennings, N. R. (2004). Trust in multiagent systems. *The Knowledge Engineering Review*, 19(1):1–25.

Regan, K. and Cohen, R. (2005). A model of indirect reputation assessment for adaptive buying agents in electronic markets. In *Proceedings of the Business Agents and Semantic Web (BASeWEB05)*, Victoria, Canada.

Riolo, R. L., Cohen, M. D., and Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, 414:441–443.

Sabater, J. and Paolucci, M. (2007). Representation and aggregation of social evaluations in computational trust and reputation models. *International Journal of Approximate Reasoning*, 46(3):458–483.

Sabater, J., Paolucci, M., and Conte, R. (2006). Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2).

Sabater, J. and Sierra, C. (2002). Social ReGreT, a reputation model based on social relations. *SIGecom Exch.*, 3(1):44–56.

Sabater, J. and Sierra, C. (2004). Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton.

Schillo, M., Fischer, K., and Klein, C. T. (2001). The micro-macro link in dai and sociology. In *MABS 2000: Proceedings of the second international workshop on Multi-agent based simulation*, pages 133–148, Secaucus, NJ, USA. Springer-Verlag New York, Inc.

Schillo, M., Funk, P., and Rovatsos, M. (2000). Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence*, 14(8):825–848.

Semmann, D., Krambeck, H.-J., and Milinski, M. (2004). Strategic investment in reputation. *Behav Ecol Sociobiol*, 56:248–252.

Shoham, Y. (1993). Agent-oriented programming. *Artif. Intell.*, 60(1):51–92.

Sichman, J. S., Conte, R., Castelfranchi, C., and Demazeau, Y. (1994). A social reasoning mechanism based on dependence networks. In *ECAI 94. 11th European Conference on Artificial Intelligence*, pages 188–192. John Wiley and Sons.

Simon, H. A. (1996). *The Sciences of the Artificial, 3rd Edition*. MIT Press, Cambridge.

Skyrms, B. (2001). The Stag Hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2):31–41.

Smith, E. R. and Conrey, F. R. (2007). Agent-Based Modeling: A New Approach for Theory Building in Social Psychology. *Pers Soc Psychol Rev*, 11(1):87–104.

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences*, 104(44):17435–17440.

Stanley, E. A., Ashlock, D., and Tesfatsion, L. (1994). Iterated prisoner's dilemma with choice and refusal of partners. In *Artificial Life III*, volume 17 of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 131–175, Reading, MA.

Stasser, G. (1988). Computer simulation as a research tool: The DISCUSS model of group decision making. *Journal of Experimental Social Psychology*, 24(5):393–422.

Tesfatsion, L. and Judd, K. L. (2006). *Handbook of Computational Economics, Volume 2, First Edition : Agent-Based Computational Economics (Handbook of Computational Economics)*. North Holland.

Wedekind, C. and Milinski, M. (2000). Cooperation Through Image Scoring in Humans. *Science*, 288(5467):850–852.

Wilson, D. S., Wilczynski, C., Wells, A., and Weiser, L. (2000). Gossip and other aspects of language as group-level adaptations. In Heyes, C. and Huber, L., editors, *The evolution of cognition*. MIT Press, Cambridge.

Yu, B. and Singh, M. P. (2000). A social mechanism of reputation management in electronic communities. In Klush, M. and Kerschberg, L., editors, *Cooperative Information Agents IV. The Future of Information Agents in Cyberspace, 4th International Workshop, CIA2000*, LNAI 1860, pages 154–165. Springer-Verlag.

Zacharia, G. (1999). Trust management through reputation mechanisms. In Castelfranchi, C., Falcone, R., and Firozabadi, B. S., editors, *Deception, Fraud and Trust in Agent Societies*, pages 163–167. Seattle.

Zacharia, G., Moukas, A., and Maes, P. (1999). Collaborative reputation mechanisms in electronic marketplaces. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Wailea Maui.