ARTÍCULO

WHY DO PEOPLE KEEP THEIR PROMISES? AN OVERVIEW OF STRATEGIC COMMITMENT

Miranda del Corral

Del Corral, M. (2015). Why do people keep their promises? An overview of strategic commitment. ***Cuadernos de Economía, 34*****(65), 237-259.**

Strategic commitments, such as promises and threats, pose several problems to the standard model of economic rationality: first, they can only arise when there is an incentive to free-ride; second, they need to be credible in order to manipulate the others' behaviour; third, once the commitment has succeeded, it is no longer in the agent's self-interest to fulfil her commitment. Why, then, do people keep their promises (and threats)? This paper reviews the literature concerning the problem of commitment within the scope of pro-sociality and cooperation, and examines two mechanisms that enable credibility and trust: reputation and social emotions.

Keywords: Strategic commitment, economic rationality, pro-sociality, cooperation, social dilemmas.
JEL: A12, C70, D03.

M. del Corral
Investigadora posdoctoral en CONICET (Argentina). Instituto de Filosofía "Alejandro Korn", Universidad de Buenos Aires, Argentina.
E-mail: miranda.delcorral@filo.uba.ar.

Sugerencia de citación: Del Corral, M. (2015). Why do people keep their promises? An overview of strategic commitment. *Cuadernos de Economía*, *34*(65), 237-260. doi:10.15446/cuad.econ.v34n65.40511.

Este artículo fue recibido el 28 de octubre de 2013, ajustado el 27 de julio de 2014 y su publicación aprobada el 8 de agosto de 2014.

Del Corral, M. (2015). ¿Por qué cumplimos nuestras promesas? Un panorama del compromiso estratégico. *Cuadernos de Economía*, *34*(65), 237-259.

Los compromisos estratégicos, tales como las promesas y las amenazas, son problemáticos para el modelo clásico de racionalidad económica: solo surgen cuando existe un incentivo para gorronear, necesitan ser creíbles para manipular el comportamiento de los demás, y una vez que esto sucede, su cumplimiento no es la mejor opción para el agente. ¿Por qué, entonces, cumple la gente sus promesas y amenazas? Este artículo ofrece una revisión bibliográfica del problema del compromiso desde la perspectiva de la prosocialidad y la cooperación, relacionándolo con la capacidad de la reputación y de las emociones sociales de generar credibilidad y confianza.

Palabras clave: compromiso estratégico, racionalidad económica, prosocialidad, cooperación, dilemas sociales.
JEL: A12, C70, D03.

Del Corral, M. (2015). Pourquoi nous tenons nos promesses ? Une vue d'ensemble de l'engagement stratégique. *Cuadernos de Economía*, *34*(65), 237-259.

Les engagements stratégiques, comme les promesses et les menaces, sont problématiques pour le modèle classique de rationalité économique : ils ne se produisent que lorsqu'il existe un stimulant pour vivre en parasite ; ils ont besoin d'être crédibles pour manipuler le comportement des autres, et une fois que cela se produit, il ne s'agit pas de la meilleure option pour l'agent. Par conséquent, pourquoi les personnes tiennent-elles leurs promesses ou exécutent-elles leurs menaces ? Cet article propose un examen bibliographique de la question de l'engagement sous l'aspect de la prosocialité et de la coopération, la reliant à la capacité de la réputation et des émotions sociales de générer crédibilité et confiance.

Mots-clés : engagement stratégique, rationalité économique, prosocialité, coopération, dilemmes sociaux.
JEL: A12, C70, D03.

Del Corral, M. (2015). Por que cumprimos as nossas promessas? Uma visão do compromisso estratégico. *Cuadernos de Economía*, *34*(65), 237-259.

Os compromissos estratégicos, tais como as promessas e as ameaças, são problemáticos para o modelo clássico de racionalidade econômica: somente surgem quando existe um incentivo para se aproveitar e filar; precisam ser críveis para manipular o comportamento dos outros e, uma vez que isto acontece, o seu cumprimento não é a melhor opção para o agente. Por que, então, as pessoas cumprem as suas promessas e ameaças? Este artigo oferece uma revisão bibliográfica do problema do compromisso da perspectiva da prossocialidade e a cooperação, relacionando-o com a capacidade da reputação e das emoções sociais de gerar credibilidade e confiança.

Palavras-chave: Compromisso estratégico, racionalidade econômica, prossocialidade, cooperação, dilemas sociais.
JEL: A12, C70, D03.

INTRODUCTION

In his famous article "An Essay on Bargaining" (1956), Thomas Schelling introduced the concept of commitment into the game-theoretical framework. Committed behaviour, following Schelling, is a kind of strategic action, the goal of which is to modify the other players' strategies, through the manipulation of their expectations (Schelling, 1960). The two paradigmatic cases of social commitment would be promises and threats. For a commitment to be effective, it has to be credible; Schelling argues that credibility can be attained by the agent by means of different mechanisms, such as the voluntary exclusion of one or more of the available options (either by making some choices impossible or by raising its cost), the power of reputation, and the ability to bargain. Nevertheless, as Schelling points out, credible promises and threats seem to lead to a paradoxical situation: once the commitment has been successfully created (*i.e.* it has managed to manipulate the other agent's choices and actions), what reason is there to keep it?

Appealing to social norms prescribing to keep one's commitment has not proved to be explanatorily successful. Social norms are one of the most invoked concepts in the social sciences; however, there is not a unified account on their formation process, or on the mechanisms that promote their enforcement (Fehr & Fischbacher, 2004a). Indeed, explanations in terms of social norms are not usual in behavioural economics, partly because their conceptualisation is vague and therefore it is problematic to include them in formal models (Bicchieri & Chavez, 2010).

The problem, then, is that the incentives to cooperate are exactly the same as those to keep one's promise or threat. Nonetheless, experimental results show a significant tendency to act according to one's commitments, as well as a tendency to consider the others' commitments credible (Boadway, Song & Tremblay, 2007; see, for instance, the experimental results in Kurzban, McCabe, Smith & Wilson, 2001). These results are puzzling because keeping a commitment is either irrational, for it does not lead to a maximisation of one's own benefit, or unnecessary, if the committed course of action is the agent's preferred option.

The purpose of this article is to review the main approaches to the problem of strategic commitment, and the challenge that experimental results pose to the model of self-interested rationality. This article is divided in two sections. First, three theoretical[1] approaches to the problem of commitment will be presented: the theory of pro-social behaviour, that analyses commitment and cooperation within the broader category of "pro-social behaviour"; Sen's view on altruistic motivation, which suggests that commitment discloses altruistic preferences that cannot be accommodated within rational choice theory; and the theory of the socially-mediated preferences, which focuses on the relation between preferences and social norms. The second section examines two mechanisms that aim

[1] An overview of how commitments work in the field is provided by Bryan, Karlan and Scott (2010).

to explain the tendency to cooperate, against the temptation to free-ride: communication and reputation, on the one hand, and social emotions, on the other.

THEORETICAL APPROACHES TO THE PROBLEM OF STRATEGIC COMMITMENT

A strategic commitment is a social interaction that has as its goal the modification of others' behaviour through the manipulation of their expectations. From a strategic point of view, then, commitment can only arise in situations in which there is an incentive to free-ride, that is, to cheat. Otherwise, a promise or a threat would not be a commitment, but a mere declaration of intentions, a forecast of one's own behaviour (Hirshleifer, 2001, p. 309). Furthermore, commitment serves as a control mechanism in order to incentivise cooperation and to overcome the temptation of free-riding: "commitment is a means by which players can assure one another that they are not going to free ride on the other's contributions, so that group members can contribute without fearing that they will be free ridden" (Kurzban, McCabe, Smith & Wilson, 2001, p. 1663).

In this sense, the possibility of creating a successful commitment (that is, a promise which is likely to be fulfilled) depends on the particular social interaction in which the commitment is created (Hardin, 1995). A social interaction is a situation in which individuals are affected by the choices of other agents (Bramoullé, 2007). The classification of different interactions depends on the context and the agent's interests; that is, on the structure of the payoffs for each player. In the game-theoretical framework, games are classified following the degrees of conflict and coordination expected from the players, depending on the structure of the payoffs. It is necessary to point out that the concept of coordination used in game theory differs from the concept of coordination that I have analysed above. A game of coordination is a game in which it is possible to reach an agreement about individual choices, because players are interested in knowing what the other players will do, and also in letting the others know what the agent is going to choose: "Coordination problems are often viewed as simple to solve. In large part this is because actors have similar interests, and, although they may not care about which solution is imposed, they all agree that some solution is needed" (Wilson & Rhodes, 1997, p. 767). There are four basic kinds of games, which correspond to four kinds of social interactions: cases of pure common-interest (in which pure coordination is expected), Battle of the Sexes games, Prisoner's Dilemma games, and inessential games (which are pure conflict situations) (Parisi, 2000).

Thus, how does a commitment modify the payoff structure? What other incentives, besides the explicit payoffs, are taken into account by the agents? Is the temptation to free-ride in a game the same kind of temptation as the one that leads to deception? In other words: If an agent is afraid of cooperating because she believes that the other agents are going to free-ride, then she would also have rea-

sons for believing that, in the case that all the other players commit themselves to cooperate, these agents plan to cheat, thus breaking their commitments (Frank, 2003). Following this argument, Sánchez-Cuenca (1998, p. 86) claims that there is a trade-off between the need of commitment and its credibility: "A commitment is credible when everyone expects that the person who makes it cannot renege on it. But it happens that the conditions that make it difficult to renege are the same conditions that make it difficult to commit. Thus, the more credible a commitment is, the more unlikely that the commitment can be made".

Hence, the more a commitment is needed, the less likely this commitment will help in solving the problem. Commitments are needed in cases in which the payoff structure incentivises the player to free-ride; and making a commitment is not going to change this situation. The stronger the temptation to free-ride, the more needed a commitment is, but the less credible it will be, so the less likely it will solve your problem. The problem of commitment can be summarised as follows: How does a commitment modify the agent's incentives, constraining their choices? And why are commitments fulfilled in the absence of an external enforcing mechanism?

Pro-sociality and Altruistic Behaviour

The capacity for making credible commitments can be understood as a mechanism that enables and promotes pro-social behaviour: knowing that others will cooperate enhances the rate of cooperation within the group. The concept of pro-social behaviour covers a broad category of interactions, which include cooperation, helping others, sharing resources, and altruistic actions (Dovidio & Penner, 2004; West, Griffin & Gardner, 2007). Its analysis takes into account cognitive, biological, motivational and social processes (Penner, Dovidio, Piliavin & Schroeder, 2005). Despite being quite a common phenomenon, pro-social behaviour challenges some central assumptions of the evolutionary and economic theories of strategic interaction, because these models claim that the goal of social interactions should be fitness, or utility, maximisation. Thus, cooperation is expected only in those cases in which both agents are better off through cooperation. In other situations, an agent would increase her fitness or her welfare by free-ridding on the others: without assuming any cost, she can benefit from the others' actions.

Promises, threats, agreements and contracts are commitment technologies that enforce cooperative behaviour. However, they inherit the problems concerning pro sociality: once the commitment is effective (*i.e*. it has manipulated the other's behaviour) there is no reason to fulfil it if doing so is not in the agent's self-interest—and, if it is, the commitment is not needed in the first place. When the other agents have cooperated believing in the honesty of the commitment, the temptation to free-ride does not disappear, unless other mechanisms intervene. Furthermore, knowing that free-ridding is the best strategy for the agent, her commitment should not be credible at all. However, despite these theoretical problems, people

make and fulfil credible promises. This conflict between the behavioural predictions of theories of economic rationality and the observed behaviour in both experimental settings and in everyday situations is called the "puzzle of pro-sociality" (Gintis, 2003, p. 157).

To cooperate means to act in a way that benefits the recipient of the action, and choosing to do so precisely because of its beneficial consequences on the recipient (West *et al*., 2007, p. 416). Cooperation can be either beneficial for the actor, thus generating a situation of mutual benefit, or it can be costly, and therefore considered a case of altruism. Then, the paradox of pro-social behaviour is the following: in situations in which cooperation is costly, there is an incentive to free-ride; however, if every individual was a free-rider, then cooperation would not be possible and the final result would not be beneficial for any individual.

There are two approaches to this paradox. The first of them consists in considering that cooperation between individuals within a group enhances the group's fitness, and that evolutionary selection occurs at the group level (Penner *et al.*, 2005; Wilson, 1975; Wilson & Sober, 1994). However, the theory of group selection has been criticised for different reasons. For instance, Nesse (1994) argues that the works on group selection do not deal with a central conceptual problem: the existence of traits that are adaptively beneficial at the group level, but that are nonetheless prejudicial at the individual level. On the other hand, it is argued that other competing theories are more broadly applicable (West, Griffin & Gardner, 2008). Despite its critics, the theory of group selection has not been completely abandoned, but its application is limited to human groups, in which survival does not merely depend on natural selection, but also in cultural selection. There is empirical support to the claim that social norms and institutions could be the result of cultural selection mechanisms (Van den Bergh & Gowdy, 2009). However, as Fehr and Fishbacher (2003) point out, it is needed to introduce additional mechanisms, such as altruistic punishment, for these norms and institutions to arise in the first place.

The second type of approach to the paradox of altruistic behaviour focuses on the individual benefits of cooperation and altruism (for a game-theoretical approach to cooperation in non-cooperative scenarios, see Sigmund, 2010). Cooperation, rather than being costly, has long-term benefits. There are two main theories within this perspective. On the one hand, the "theory of inclusive fitness", or "kin selection", proposed by Hamilton (1964), focuses on the tendency of genetically related individuals to mutually benefit each other, thus facilitating their reproductive success, and raising the probability to pass their genes on to the next generation. However, this theory cannot explain why genetically unrelated individuals engage in cooperation, unless additional mechanisms are presupposed. On the other hand, the theory of reciprocal altruism, introduced by Trivers (1971), applies a game-theoretical framework to scenarios in which there is an incentive to free-ride, such as the Prisoner's Dilemma. Following Trivers, there are many cases of behaviour which were previously considered altruistic that are not fully disinterested. Recip-

rocal altruism is a kind of behaviour in social interactions in which an agent makes sacrifices (*i.e.* reduces her fitness) for another individual, with the expectation that the other agent will act in a similar way in future interactions. Cooperating can be seen as an investment for the agent, especially if punishment mechanisms are introduced.

The accounts focusing on reciprocal altruism pay attention both to the general tendency of cooperation at the group level, and to the specific individual interactions: the latter are supposed to explain the former. However, large groups are problematic, because the larger the group, the less advantageous it is to set up a control mechanism to avoid the temptation of free-riding (Fehr & Gächter, 2002). In small groups, individuals tend to interact repeatedly with other members. Thus, the outcome of previous interactions can be recorded and used to decide whether to engage in a new interaction with the same individual – there are no one-shot encounters, but repeated interactions. The fact that the group is small allows for a situation of perfect information. On the contrary, within large groups, the probability of repeating an encounter diminishes, and because it is no longer a situation of perfect information, problems of adverse selection arise. Furthermore, in large groups, the monitoring process is more costly, and hence the enforcement costs rise as well (Carpenter, 2007). Thus, additional mechanisms to keep the rate of cooperation are needed.

Inspired by Trivers' theory of reciprocal altruism, Axelrod and Hamilton (1981) tackled the problem of what strategies are evolutionarily stable—that is, a strategy, such as a spontaneous apparition of a mutation of that strategy, which does not alter its initial predominance—in the Prisoner's Dilemma game. They showed that the "tit-for-tat" strategy turned out to be stable, robust, and plausible to appear for the first time in a randomised system. In an iterated Prisoner's Dilemma, the "tit-for-tat" strategy consists in starting the game by cooperating, and then copying the other player's last movement: it is a strategy based on reciprocity, and has become a paradigmatic explanation of reciprocal altruism (Nowak & Sigmund, 1993). There are other similar strategies that are also able to punish defection, such as "always cooperate and punish your partner after each round in which it failed to cooperate as well", or "play tit-for-tat and in addition punish the partner for each defection", or "start cooperatively, punish your partner the first time it fails to cooperate and switch to defection if the punishment does not alter the partner's behaviour" (Bshary & Bergmuller, 2008; see also Hammerstein, 2003; Nowak, 2006). These strategies have in common that the player is sensitive to the other player's previous choices, and thus choice is not exclusively based on the immediate payoffs of the encounter, but includes external considerations.

The concept of strong reciprocity is central to this theoretical framework. It refers to the predisposition to cooperate with others, and to punish defective agents, even in cases in which this behaviour cannot be justified through self-interest, kin or reciprocal altruism (Gintis, 2000b). This strategy is a combination of various control mechanisms that incentivise cooperation:

> Strong reciprocity is a combination of altruistic rewarding, which is a predisposition to reward others for cooperative, norm-abiding behaviours, and altruistic punishment, which is a propensity to impose sanctions on others for norm violations. Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts. In contrast, reciprocal altruists, as they have been defined in the biological literature, reward and punish only if this is in their long-term self-interest (Fehr & Fischbacher, 2003, p. 785).

The difference between reciprocal altruism and strong reciprocity lies in that a reciprocal altruist will only cooperate if she expects future returns for cooperation, while a strong reciprocator will respond to the kindness perceived in the other player, rather than in the immediate or future payoffs of the game. Strong reciprocity is observed to take place both in real interactions and in laboratory experiments (Fehr, Fischbacher & Gächter, 2002), and plays a central role in the enforcement and content of social norms (Fehr & Fischbacher, 2004b), especially when third-party agents are allowed to reward or punish the agents involved in an interaction.

In brief, control mechanisms favour, in repeated encounters, agents that increase their tendency to cooperate; nonetheless, many of these control mechanisms are costly. However, what is the incentive to set up a control mechanism in a one-shot encounter? Why punish a cheater, incurring into costs, if the agent will not interact again with the cheater in the future? Experimental evidence shows that, in larger groups, the level of cooperation decreases (Fehr & Fischbacher, 2003). Altruistic punishment and strong reciprocity have been proven to be effective mechanisms to maintain the rate of cooperation within larger groups (Boyd, Bowles, Richerson & Gintis, 2003; Fehr & Rockenbach, 2004), which enhances the group fitness (Gintis, 2000a). It is thus necessary to specify the motivational mechanisms underlying this kind of behaviour because, despite being adaptive at the group level, they do not offer immediate advantages for cooperative agents.

A different way of dealing with the problem of commitment consists in challenging the relation between commitment and self-interest. The rationality of committed behaviour is problematic because the benefits of this behaviour are not immediate, or even non-existent, such as in the case of altruistic punishment to an agent in a one-shot interaction. Thus, it has been argued that preferences are not exclusively guided by self-interest or welfare maximisation, but also by other factors such as moral considerations.

Sen on Commitment as Altruistic Motivation

Amartya Sen's “Rational Fools” (1977) is nowadays one of the most cited and commented works in the field of rational choice theory (RCT from now on). From Sen's point of view, commitment cannot be accommodated in RCT explanations because it opens a wedge between welfare and choice. In “Rational Fools”, Sen argued that we must distinguish between two separate concepts: sympathy and

commitment. The former corresponds to the case in which the concern for others directly affects one's own welfare: "If the knowledge of tortures of others make you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment" (Sen, 1977, p. 319). Hence, the difference between sympathy and commitment lies in how the status of others affects one's welfare. Later, in "Goals, Commitment and Identity" Sen (1985), developed the theoretical distinction between self-centred welfare, self-welfare goal, and self-goal choice, and placed this distinction at the core of RCT models. The self-centred assumption states that an agent's welfare depends only on her own consumption. The self-welfare goal assumption states that an agent's only goal is to maximise her welfare. The third assumption, self-goal choice, states that the agent's choices must respond merely to her goals. Sen argues that sympathy only violates the self-centred welfare condition, because the welfare of others influences our own welfare[2]. RCT can easily explain this kind of "altruism", due to the fact that an agent's welfare increases by making other's welfare increase as well. Sen argues that commitment, however, involves making a choice, which violates either the RCT requirement of self-welfare goal, or self-goal choice. Sen claims that "commitment is concerned with breaking the tight link between individual welfare (with or without sympathy), and the choice of action (for example, being committed to help remove some misery even though one personally does not suffer from it)" (Sen, 1985, pp. 7-8).

Sen's critique focuses on the self-interested assumptions of classic RCT (Debreu, 1959). In spite of the attempts to broaden the concept of welfare in order to include altruistic preferences (Becker, 1974), Sen argues that broadening the concept of welfare is not a satisfactory solution, because the underlying problem is the connection between welfare and preferences: Sen claims that an agent is able to choose an option that violates her preferences, because the choice is not exclusively made on the basis of the agent's welfare, but on the agent's commitments.

The claim that agents are able to make counter-preferential choices is highly controversial, because it undermines the common understanding of preferences. Other alternative concepts of preference can broaden the motivational scope of the agent, and thus include committed action in the set of preferred actions. For example, Hausman (2005) argues that commitment does not entail counter-preferential choice if preferences are seen as all-things-considered rankings. Rather, Hausman argues, commitment should be invoked as one of the preference formation mechanisms; that is, as a kind of motivation, among other factors. From a different perspective, but also regarding broader concepts of preferences, it has been argued that individuals do not only deliberate in order to maximise their individual welfare, but that they also deliberate as participants of a group or a team (Sugden,

[2] Sen (2009) has continued the development of this idea in recent work, such as his "The Idea of Justice". See also the volume edited by Peter and Schmid (2007), "Rationality and Commitment", devoted to Sen's work.

1993, 2000). Individual agents would have team preferences that are not reducible to individual preferences, and they try to fulfil these preferences when acting as a group member.

On the other hand, Sen argues that commitment violates the self-goal choice assumption. Commitment consists thus in the adoption of some other agent's goals, and the willingness to promote this goal; threats cannot be considered commitments from this perspective (see Guerini & Castelfranchi, 2007 for an analysis of the asymmetry between promises and threats). However, adopting someone else's goal is different from acting to promote someone else's goal. Drawing a distinction between goal modifying and goal displacing, Pettit (2005) argues that, while the modification of an agent's goals in order to consider other people's goals is quite common, the possibility of acting in order to attain a goal that the agent does not have is highly implausible: the very notion of agency entails a relation between an agent's goals and actions. While goal displacing requires a departure from RCT, modelling the deliberation process that allows agents to include other agent's goals as their own can accommodate goal modifying.

It has been argued that Sen's critique has a normative dimension that cannot be accommodated within the strategic rationality framework, insofar as RCT is not a theory of action or rationality, but a framework to explore the formal restrictions on the structure of preferences (Brennan, 2007; Güth & Kliemt, 2004). While RCT includes some assumptions about the content of the agents' preferences, it also leaves room for motivational theories to explain the formation of preferences.

Socially-Mediated Preferences

Social norms prescribe a certain behaviour, which is in turn expected from the agents, and therefore enable the prediction of sanctions in case of non-compliance. External mechanisms, such as rewards and punishments, are used to enforce social norms; and often these behavioural mechanisms are guided by norms of fairness or reciprocity. Thus, norms tend to be self-enforcing: the violation of a norm of reciprocity may be responded to with a sanction; and this process of sanctioning is norm-mediated, and not a product of deliberation (Posner & Rasmusen, 1999). The enforcement of social norms is related to strong reciprocity (Fehr & Fischbacher, 2004b): not only do agents have a tendency to comply with the norm, but they are also willing to sacrifice part of their welfare to reward or punish other agents. The compliance to norms of cooperation and fairness increase when strong reciprocators are able to make credible commitments to punish deviant behaviour (Sethi & Somanathan, 2005).

Besides having a preference for acting consistently, agents also tend to follow informal rules of fairness and to avoid inequity. In fact, these two concepts are deeply interlinked:

> [W]e model fairness as self-centered inequity aversion. Inequity aversion means that people resist inequitable outcomes, *i.e*., they are willing to give up some material payoff to move in the direction of more equitable outcomes. Inequity aversion is self-centered if people do not care per se about inequity that exists among other people but are only interested in the fairness of their own material payoff relative to the payoff of others (Fehr & Schmidt, 1999, p. 819).

Sometimes norms of fairness conflict with self-interest, understood as a maximisation of self-welfare. This is why the tendency to comply with norms of fairness has been referred to in the literature as if the agent chose according to her "social preferences". Models of social preferences assume that people are self-interested, but are also concerned with the payoffs of the other players (Charness & Rabin, 2002). For example, in a Dictator's game,[3] a rational agent would give zero tokens to the other player; however, on average, people share 30% of their tokens (Croson & Konow, 2009). This 30% is understood as a measure of social preference. Reciprocal altruism, inequity aversion and strong reciprocity are usually modelled as social preferences. Nonetheless, the explanation of the formation of these kinds of preferences are usually left unattended in the economic literature, partially because the answer to the motivations underlying preferences may have an evolutionary (both biological and cultural) origin, leaving its analysis for evolutionary biologists and anthropologists. One possible approach to the formation of social preferences, including altruistic behaviour, is that they could be the result of human "docility", in Simon's terms (Simon, 1990, 1993). Docility is the "tendency to depend on suggestions, recommendations, persuasion, and information obtained through social channels as a major basis of choice" (Simon, 1993, p. 156). Human rationality is limited and not able to support optimisation, and it is therefore approximate and bounded (Gigerenzer & Selten, 2002). Thus, by learning how to appropriately respond to a social scenario (such as sharing resources), the agent employs heuristic mechanisms, which, rather than maximising the outcome, optimise the decision process.

Lastly, it is possible to model the compliance to social norms as a kind of preference to follow the norms under specific conditions, such as the belief that other players will do so, and the belief that other players think that the agent should comply with the norm (Bicchieri, 2006). Both empirical expectations, which are our beliefs about what other agents will do, and normative expectations, which are our beliefs about what other players believe we should do, seem to be determinant for conditional cooperation (Bicchieri & Xiao, 2008). Moreover, the salience of certain social norms in the social setting can affect the decision of following the norm (Bicchieri & Chavez, 2010). Thus, the conditional preference for rule-following would be, from this theoretical point of view, conditioned by the beliefs of

[3] The Dictator's game is not a game in the strict sense, for there is only one player (the dictator), who is given a certain amount of tokens, and has the possibility of offering a share of the tokens to the other player, who has no active role in the game (he can neither accept of reject the share).

the agent about what the other will do, except in the cases of moral norms, which demand an unconditional commitment (Bicchieri, Nida-Rümelin & Spohn, 2000).

MECHANISMS THAT ENABLE CREDIBILITY AND TRUST

Accounts of cooperation in terms of socially mediated preferences have been recently criticised. From the point of view of evolutionary psychology, the cognitive architecture required for cooperation includes several mechanisms that regulate our social lives, particularly at the level of interaction between individuals:

> A strong bias to trust; the use of cooperative reputation to initially decide which partners to trust; placing greater weight on how the partner treated you versus others in making decisions to trust, cooperate, and punish; the replacement of reputational cues by direct experience to regulate subsequent interactions; the use of punishment as a bargaining tool when you plan to continue the relationship–these features all fit together as an efficient architecture for small scale social exchange, rather than large scale norm maintenance. It is possible to argue that both psychologies coexist (Krasnow, Cosmides, Pedersen & Tooby, 2012, p. 8).

Thus, an explanation of pro-social behaviour ought to, following this perspective, include multiple cognitive mechanisms and capacities that would be the result of the evolution of the human species as fundamentally social beings.

Following Nesse (2001) there are four reasons to believe that a commitment is likely to be fulfilled. First, a commitment can be self-enforcing: after the creation of the commitment, the action involved becomes the best option for the agent, following her self-interest. In this case, the creation of the commitment implies a restriction of options, either by making them unavailable, or by raising its costs (Elster, 2000, 2003). Burning one's bridges or ships would be a paradigmatic way to create a self-enforcing commitment. Once the commitment technology is set up, there is no need for an additional mechanism to incentivise its fulfilment, because fulfilling it matches the agent's self-interest.

Second, a commitment can be reinforced by external incentives controlled by third parties. For example, a contract is a commitment that is enforced by legal punishment. These two mechanisms turn the fulfilment of a commitment into a self-interested action. Besides the sceptical argument, which states that commitment faces the same problems as cooperation, and thus is not effective without external constraints, there is larger experimental support to the claim that communication enhances cooperation in social dilemmas (Balliet, 2010; Kerr, Garst, Lewandowski & Harris, 1997; Kerr & Kaufman-Gilliland, 1994; Meleady, Hopthrow & Crisp, 2013). Thus, other internal mechanisms play a role in the explanation of the effectiveness of a commitment.

The third mechanism Nesse points out is reputation. Forthly, the binding force of a social commitment can be related to emotions. Nesse claims that, when the reputational and emotional mechanisms come into play, the commitment is "subjective". The only reason why these commitments are effective, Nesse argues, is because of their capacity to persuade others that the committed agent will act against his self-interest if he violates his commitment.

Communication and Reputation

The problems of pro-social behaviour stated in the previous section can be extrapolated to the analysis of communication. Truthful communication has the same problems as cooperative behaviour: additional mechanisms are needed to overcome the temptation of sending wrong signals to take advantage.

Communication is a necessary part of social interaction, in which the actions of an individual generate a signal that modifies the behaviour of the receiver (Wiley, 1983). There are two main theoretical approaches to animal communication. The first of them considers communication as a mechanism to transmit information. From the point of view of group selection, clear and non-ambiguous signals are evolutionarily advantageous, especially when they are meant to inform about states of affairs that have not been directly experienced. To overcome the temptation of emitting false signals, and free-ride on the honest signals of others, (Zahavi 1975; Zahavi, Zahavi, Balaban & Ely, 1999) proposes the existence of a "handicap principle". According to this principle, the communication of honest signals arises when the cost of sending the signal is elevated, and therefore cheating becomes too costly.

The second approach to communication comes from sociobiology, and it criticises the claim that the function of communication is to share information: communication would be better understood as the manipulation of the other's behaviour: "the evolution of many animal signals is best seen as an interplay between mind-reading, and manipulation" (Krebs & Dawkins, 1984, p. 380). A cheater would have adaptive advantage within a group in which individuals always send honest signals, thus complete honesty cannot be an evolutionarily stable strategy. Krebs and Dawkins suggest that the goal of sending a signal is to manipulate the receiver in a way that fulfils the sender's self-interest. The receiver then needs to predict what action the sender will perform.

Maynard, Smith and Harper (1995, 2003) suggest a combination between the informational and the manipulative approaches. They argue that it is not evolutionarily stable for the receiver to modify his behaviour (just as the sender pretends) unless the information contained in the message is credible and useful. For example, a threat signal will not have any effect on the receiver if he does not identify that signal as a credible threat. Game-theoretical models of signal credibility confirm the inverse correlation between the cost of a signal and the incentives

to free-ride by sending that signal (Gintis, Smith & Bowles, 2001). Besides the cost of producing a signal (intrinsic cost), reputation also raises the cost of sending false signals.

Reputation affects the willingness of individuals to engage in a repeated social interaction with another individual. Triver's reciprocal altruism consists in cooperating with those who have been cooperative in earlier interactions; Alexander's (1987) strategy of indirect reciprocity consists in cooperating with those who have either been cooperative in earlier rounds, or that are known by the agents to have been cooperative in earlier rounds. Strategies based on reputation, such as reciprocal altruism or indirect reciprocity, need a mechanism to register the past behaviour of individuals, and a set of rules to assess how to behave, depending on the information available about the other individual (Nowak & Sigmund, 2005). It is possible to distinguish different levels of complexity in the mechanisms that generate a reputation system. The first of them is based on the emotions of fear and submission (Henrich & Gil-White, 2001). The mechanism that promotes (or restrains) cooperation is the set of emotions that the other individual causes on the agent, and these emotions can be prompted by previous interactions, or by the observation of interactions. A more complex level would involve more complex cognitive processes, such as the possibility of making predictions about the behaviour of others. At this level, instead of manipulating directly the behaviour of others, the agent tries to manipulate their expectations, thus generating trust relations.

Regarding positive commitments (that is, leaving threats aside), trust is a necessary mechanism for assigning credibility. Trust is a complex motivational and cognitive state that enables the generation of empirical and normative expectations about the other's behaviour under risk circumstances. The effectiveness of a social commitment depends on the successful manipulation of the other agent's choices, and a minimum level of trust is required between the two agents for a commitment to be credible, and hence effective (Hardin, 2003; Simpson, 2007). A relation of trust entails a disposition to rely on the other agent for the fulfilment of one's own goals (Castelfranchi & Falcone, 2002). Thus, trust leads to the credibility of commitments when there is a situation of dependence and uncertainty (Barbalet, 2009). By trusting the other, the agent incurs in costs: she chooses according to a future payoff, rather than a present one. Without uncertainty, trust is not necessary, because the agent expects the choices of the other player independently of the commitments in which that player has incurred.

Reputation is able both to incentivise the fulfilment of a threat and its credibility. In cultures of honour, the defence of one's own reputation is achieved through violent and disproportionate threats, which are usually carried out when the agent is challenged (Cohen & Vandello, 2001). Not carrying out a threat can, as a consequence, mean that in future encounters the agent is free-ridden. Furthermore, the defence of honour is related to social emotions such as rage or anger, and shame and humiliation, which are relevant to explain the fulfilment of

threats in cases in which it is more advantageous not to do so (Mosquera, Manstead & Fischer, 2002).

This kind of reputation is exclusively found in human societies. The problem of the credibility of signals and its relation to reputation switches the focus: instead of asking, why do agents send honest signals in contexts in which not doing so enhances the individual's fitness? The question would be, why do agents send honest signals in contexts in which not doing so is strategically advantageous?

However, recent experimental results show that communication is an effective mechanism for enhancing cooperation even when it is anonymously performed; therefore, it is necessary to appeal to other mechanisms, such as social emotions, or to cultural norms of sincerity and trust (Baum, Paciotti, Richerson, Lubell & McElreath, 2012), in order to explain the statistical correlation between communication and cooperation.

The Role of Emotions

Social emotions play an important role in the explanation of the effectiveness of commitments, because they are understood to attribute a positive impact on the promotion of credibility and on the motivation for their fulfilment. Explanations of social behaviour appealing to emotions usually take the form of mechanistic explanations (Elster, 2005; Muramatsu & Hanoch, 2005). Mechanisms differ from simplistic causal (chemical or physical) devices in that they may trigger different responses when facing the same situation. Hence, although emotions serve as enablers of social commitments, they may also prevent the fulfilment of a commitment. Particularly, in complex situations in which the agent has conflicting interests (derived from the outcome of keeping her commitment, on the one hand, and violating it, on the other), emotional mechanisms may make the commitment more or less effective.

Evolutionary analysis shows that emotions have survival and reproductive functions, which are manifested at four different levels: intra-individual, dyadic, group, and cultural (Keltner, Haidt & Shiota, 2006). While the functions of emotions at the first level tend to enhance individual fitness, the same functions at the other three levels usually favour the creation of social bounds and cooperation.

From the point of view of strategic rationality, as it has been pointed out above, one-shot encounters are essentially different from repeated encounters. In a one-shot game, control mechanisms such as long-term investments or the building of a reputation cannot arise. Social emotions play this role: They motivate cooperation, serve as a guide to choose a partner for interaction, and enable the creation and perdurance of long-term relationships (Back & Flache, 2008; Frank, 2001; Gonzaga, Keltner, Londahl & Smith, 2001). Furthermore, detecting the other's emotions also serves as a mechanism for evaluating the interaction partners and to avoid cheaters or free-riders (Cosmides & Tooby, 2004; Frank, 2001). The

feeling of anger or frustration after being cheated dis-incentivises future interactions with the same individual, and can motivate altruistic punishment (Lerner, Goldberg & Tetlock, 1999; Petersen, Sell, Tooby & Cosmides, 2012), whilst guilt-aversion dis-incentivises the breach of a commitment (Battigalli & Dufwenberg, 2007; Charness & Dufwenberg, 2010; Ellingsen, Johannesson, Tjotta & Torsvik, 2010). Concerning the relation between emotions and beliefs, Vanberg (2008) argues that emotions create an indirect bond between preferences and promises, because they are able to modify the second-order beliefs of the agent.

Fehr and Gächter (2002) argue that in a Public Goods game[4], cooperation only arises when agents have the possibility of punishing free-riders. Their study shows that there is a correlation between the intensity of the emotion felt and the punishment executed. In the Public Goods game, those players who have invested more tokens report the most intense negative emotions, and this intensity also increases when the amount of tokens invested by the other player is lower. Other studies point out the necessity of including the role of social norms to understand the relation between the emotions and expectations of the agents: expectations are based on the fulfilment or violation of the agent's expectations, and these, in turn, are generated following social standards (Bosman & Van Winden, 2002; Hoffman, McCabe, Shachat & Smith, 1994; Wu *et al.*, 2009). Lastly, emotions do not only play a role in the motivation of punishment, but the expression of a negative emotion serves as a punishment mechanism through the generation of feelings of guilt or shame (Xiao & Houser, 2005).

CONCLUSION

The problem of strategic commitment, thus, derives from a broader puzzle that challenges the assumption of natural or rational selfish behaviour. Of course, it may be argued, in keeping with Sen's view, that there are moral reasons to keep our promises—we ought not manipulate others to obtain a strategic advantage over them. In fact, as Tomasello and Vaish (2013) have recently argued, the mechanisms that enable morality and cooperation share a common evolutionary origin. The last decades of research on pro-social behaviour have proved that the individualistic, self-centred, and egoistic model of instrumental rationality offers a narrow understanding of human agency and motivation.

[4] A Public Goods game is a social dilemma, in which the players have the chance to invest an amount of tokens in the production of a public good. The tokens invested in the public good are multiplied and distributed equally among the players, even among those who have invested zero tokens (free-riders).

REFERENCES

1. Alexander, R. D. (1987). *The biology of moral systems*. Camden, NJ: Transaction Publishers.
2. Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390-1396.
3. Back, I., & Flache, A. (2008). The adaptive rationality of interpersonal commitment. *Rationality and Society*, *20*(1), 65-83.
4. Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39-57.
5. Barbalet, J. (2009). A characterization of trust, and its consequences. *Theory and Society*, *38*(4), 367-382.
6. Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *The American Economic Review*, *97*(2), 170-176.
7. Baum, W. M., Paciotti, B., Richerson, P., Lubell, M., & McElreath, R. (2012). Cooperation due to cultural norms, not individual reputation. *Behavioural processes*, *91*(1), 90-93.
8. Becker, G. S. (1974). A theory of social interactions. *The Journal of Political Economy*, *82*(6), 1063-1093.
9. Bicchieri, C. (2006). *The grammar of society: The emergence and dynamics of social norms*. Cambridge: Cambridge University Press.
10. Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*(2), 161-178.
11. Bicchieri, C., Nida-Rümelin, J., & Spohn, W. (2000). Words and deeds: A focus theory of norms. In J. Nida-Rümelin & W. Spohn (Eds.), *Rationality, rules and structure*. Dordrecht: Kluwer Academic Publishers.
12. Bicchieri, C., & Xiao, E. (2008). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, *21*, 1-18.
13. Boadway, R., Song, Z., & Tremblay, J. F. (2007). Commitment and matching contributions to public goods. *Journal of Public Economics*, *91*(9), 1664-1683.
14. Bosman, R., & Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *Economic Journal*, *112*(476), 147-169.
15. Boyd, R., Bowles, S., Richerson, P. J., & Gintis, H. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3531-3535.
16. Bramoullé, Y. (2007). Anti-coordination and social interactions. *Games and Economic Behavior*, *58*(1), 30-49.
17. Brennan, G. (2007). The grammar of rationality. In F. Peter & H. B. Schmid (Eds.), *Rationality and commitment*. Oxford: Oxford University Press.

18. Bryan, G., Karlan, D., & Scott, N. (2010). Commitment devices. *Annual Review of Economics*, *2*(1), 671-698.
19. Bshary, R., & Bergmuller, R. (2008). Distinguishing four fundamental approaches to the evolution of helping. *Journal of evolutionary biology*, *21*(2), 405-420.
20. Carpenter, J. P. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, *60*(1), 31-51.
21. Castelfranchi, C., & Falcone, R. (2002). Social trust: A cognitive approach. In C. Castelfranchi & Y. H. Tan (Eds.), *Trust and deception in virtual societies*, Dordrecht: Kluwer Academic Publishers, 55-90.
22. Charness, G., & Dufwenberg, M. (2010). Bare promises: An experiment. *Economics Letters*, *107*(2), 281-283.
23. Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817.
24. Cohen, D., & Vandello, J. (2001). Honor and "faking" honorability. *Evolution and the capacity for commitment*, 163-185.
25. Cosmides, L., & Tooby, J. (2004). Evolutionary psychology and the emotions. *Handbook of emotions*, 91.
26. Croson, R., & Konow, J. (2009). Social preferences and moral biases. *Journal of Economic Behavior & Organization*, *69*(3), 201-212.
27. Debreu, G. (1959). *Theory of value: An axiomatic analysis of economic equilibrium*. New Haven, Londres: Yale University Press.
28. Dovidio, J. F., & Penner, L. A. (2004). Helping and altruism. In M. B. Brewer & M. Hewstone (Eds.), *Emotion and motivation* (p. 247). Londres: Blackwell Publishers.
29. Ellingsen, T., Johannesson, M., Tjotta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95-107.
30. Elster, J. (2000). *Ulysses unbound: Studies in rationality, precommitment, and constraints*. Cambridge: Cambridge University Press.
31. Elster, J. (2003). Don't burn your bridge before you come to it: Some ambiguities and complexities of precommitment. *Tex. L. Rev.*, *81*(7), 1751-1787.
32. Elster, J. (2005). En favor de los mecanismos - A plea for mechanisms. *Sociologica, 19*(57), 239-273.
33. Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785-791.
34. Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185-190.

35. Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87.
36. Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1-25.
37. Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140.
38. Fehr, E., & Rockenbach, B. (2004). Human altruism: Economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology*, *14*(6), 784-790.
39. Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, *114*(3), 817-868.
40. Frank, R. H. (2001). Cooperation through emotional commitment. *Evolution and the capacity for commitment*, *3*, 57-76.
41. Frank, R. H. (2003). Commitment problems in the theory of rational choice. *Tex. L. Rev.*, *81*(7), 1789-1804.
42. Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.
43. Gintis, H. (2000a). Group selection and human prosociality. *Journal of Consciousness Studies, 7*, *1*(2), 215-219.
44. Gintis, H. (2000b). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169-179.
45. Gintis, H. (2003). Solving the puzzle of prosociality. *Rationality and Society*, *15*(2), 155-187.
46. Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signalling and cooperation. *Journal of theoretical biology*, *21, 3*(1), 103-119.
47. Gonzaga, G. C., Keltner, D., Londahl, E. A., & Smith, M. D. (2001). Love and the commitment problem in romantic relations and friendship. *Journal of Personality and Social Psychology*, *81*(2), 247-262.
48. Guerini, M., & Castelfranchi, C. (2007). Promises and threats in persuasion. *Pragmatics and Cognition*, *15*(2), 277-311.
49. Güth, W., & Kliemt, H. (2004). The rationality of rational fools: The role of commitments, persons and agents in rational choice modelling. In F. Peter & H. B. Schmid (Eds.), *Rationality and commitment*. Oxford: Oxford University Press.
50. Hamilton, W. D. (1964). The genetical evolution of social behavior, parts 1 and 2. *Journal of Theoretical Biology*, *7*(1), 1-52.
51. Hammerstein, P. (2003). *Genetic and cultural evolution of cooperation*. Cambridge, MA: MIT Press.
52. Hardin, R. (1995). *One for all: The logic of group conflict*. Princeton, N.J.: Princeton University Press.

53. Hardin, R. (2003). Gaming trust. In E. Ostrom & J. Walker (Eds.), *Trust and reciprocity: Interdisciplinary lessons from experimental research* (pp. 80-101). Nueva York: Russell sage foundation publications.
54. Hausman, D. M. (2005). Sympathy, commitment, and preference. *Economics and Philosophy*, *21*(1), 33-50.
55. Henrich, J., & 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.
56. Hirshleifer, J. (2001). *The dark side of the force: Economic foundations of conflict theory*. Cambridge: Cambridge University Press.
57. Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*(3), 346-380.
58. Keltner, D., Haidt, J., & Shiota, M. N. (2006). Social functionalism and the evolution of emotions. In M. Schaller, J. A. Simpson & Kenrick, Douglas T. (Eds.), *Evolution and Social Psychology* (p. 115). Nueva York: Psychology press.
59. Kerr, N. L., Garst, J., Lewandowski, D. A., & Harris, S. E. (1997). That still, small voice: Commitment to cooperate as an internalized versus a social norm. *Personality and Social Psychology Bulletin*, *23*(12), 1300-1311.
60. Kerr, N. L., & Kaufman-Gilliland, C. M. (1994). Communication, commitment, and cooperation in social dilemma. *Journal of Personality and Social Psychology*, *66*(3), 513-529.
61. Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What are punishment and reputation for? *PloS one*, *7*(9), doi:10.1371/journal.pone.0045662.
62. Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation. *Behavioural ecology: An evolutionary approach*, *2*, 380-402.
63. Kurzban, R., McCabe, K., Smith, V. L., & Wilson, B. J. (2001). Incremental commitment and reciprocity in a real-time public goods game. *Personality and Social Psychology Bulletin*, *27*(12), 1662-1673.
64. Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology*, *29*(56), 781-795.
65. Meleady, R., Hopthrow, T., & Crisp, R. J. (2013). Simulating social dilemmas: Promoting cooperative behavior through imagined group discussion. *Journal of Personality and Social Psychology*, *104*(5), 839-853.
66. Mosquera, P. M., Manstead, A. S., & Fischer, A. H. (2002). The role of honour concerns in emotional reactions to offences. *Cognition & Emotion*, *16*(1), 143-163.

67. Muramatsu, R., & Hanoch, Y. (2005). Emotions as a mechanism for boundedly rational agents: The Fast and Frugal Way. *Journal of Economic Psychology 26*(2), 201-221.
68. Nesse, R. M. (1994). Why is group selection such a problem? *Behavioral and Brain Sciences*, *17*(4), 633-634.
69. Nesse, R. M. (2001). Natural selection and the capacity for subjective commitment. *Evolution and the Capacity for Commitment*, 1-44.
70. Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560-1563.
71. Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*(6432), 56-58.
72. Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291-1298.
73. Parisi, F. (2000). The cost of the game: A taxonomy of social interactions. *European Journal of Law and Economics*, *9*(2), 99-114.
74. Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: multilevel perspectives. *Annual Review of Psychology*, *56*(1), 365-392.
75. Peter, F., & Schmid, H. B. (2007). *Rationality and commitment*. Oxford: Oxford University Press.
76. Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*, *33*(6), 682-695.
77. Pettit, P. (2005). Construing Sen on commitment. *Economics and Philosophy*, *21*(1), 15-32.
78. Posner, R. A., & Rasmusen, E. B. (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics*, *19*, 369-382.
79. Sánchez-Cuenca, I. (1998). Institutional commitments and democracy. *European Journal of Sociology*, *39*(01), 78-109.
80. Schelling, T. C. (1956). An essay on bargaining. *The American Economic Review*, *46*(3), 281-306.
81. Schelling, T. C. (1960). *The strategy of conflict*. Harvard: Harvard University Press.
82. Sen, A. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, *6*(4), 317-344.
83. Sen, A. (1985). Goals, commitment, and identity. *Journal of Law, Economics, and Organization*, *1*(2), 341-355.
84. Sen, A. K. (2009). *The Idea of Justice*. Cambridge, MA: Belknap Press.

85. Sethi, R., & Somanathan, E. (2005). Norm compliance and strong reciprocity. In H. Gintis, S. Bowles, R. T. Boyd & E. Fehr (Eds.), *Moral sentiments and material interests: The foundations of cooperation in economic life* (pp. 229-250). Cambridge, MA: The MIT Press.
86. Sigmund, K. (2010). *The calculus of selfishness*. Princeton, N.J.: Princeton University Press.
87. Simon, H. A. (1990). A mechanism for social selection and successful altruism. *Science*, *250*(4988), 1665-1668.
88. Simon, H. A. (1993). Altruism and economics. *The American Economic Review*, *83*(2), 156-161.
89. Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, *16*(5), 264-268.
90. Smith, M. J., & Harper, D. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, *177*(3), 305-311.
91. Smith, J. M., & Harper, D. (2003). *Animal signals*. Oxford: Oxford University Press.
92. Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy*, *10*(01), 69-89.
93. Sugden, R. (2000). Team preferences. *Economics and Philosophy*, *16*(02), 175-204.
94. Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, *64*, 231-255.
95. Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35-57.
96. Van den Bergh, J. C. J., & Gowdy, J. M. (2009). A group selection perspective on economic behavior, institutions and organizations. *Journal of Economic Behavior and Organization*, *72*(1), 1-20.
97. Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, *76*(6), 1467-1480.
98. West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*(2), 415-432.
99. West, S. A., Griffin, A. S., & Gardner, A. (2008). Social semantics: How useful has group selection been? *Journal of Evolutionary Biology*, *21*(1), 374-385.
100. Wiley, R. H. (1983). The evolution of communication: Information and manipulation. *Animal Behaviour*, *2*, 156-189.
101. Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(1), 143-146.

102. Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences*, *17*(04), 585-608.

103. Wilson, R. K., & Rhodes, C. M. (1997). Leadership and credibility in n-person coordination games. *Journal of Conflict Resolution, 41*(6), 767-791.

104. Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, *106*(41), 17448-17451.

105. Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7398-7401.

106. Zahavi, A. (1975). Mate selection–a selection for a handicap. *Journal of theoretical Biology*, *53*(1), 205-214.

107. Zahavi, A., Zahavi, A., Balaban, A., & Ely, M. P. (1999). *The handicap principle: A missing piece of Darwin's puzzle*. Oxford, USA: Oxford University Press.