J Archaeol Method Theory (2014) 21:325–342
DOI 10.1007/s10816-013-9186-3

Modelling Cooperation Mechanisms: Some Conceptual Issues

Mauricio Salgado · José A. Noguera · Francisco J. Miguel

Published online: 26 October 2013

Abstract Several evolutionary mechanisms have been proposed to explain how natural selection leads to cooperation between competing individuals. Social dilemmas modelled with the aid of game theory capture the essence of this problem, and computer simulation is usually the technique used to test and formalise those explanatory mechanisms. However, scarce attention has been paid to what the notion of 'mechanisms' means and involves in the literature. Also, the key issue about when a computer simulation provides a good explanation tends to be ignored. In this article, we tackle these two drawbacks in the literature by calling attention to the implications of the notion of 'social mechanism' along different conceptual dimensions, such as ontological status, regularity, transparency, intelligibility, and reduction. We also claim that computer simulation, and specially agent-based modelling, provides a plausible explanation to social cooperation only if it satisfies some criteria of empirical adequacy instead of just being capable of generating cooperation in a virtual system. Finally, we relate these issues to five evolutionary mechanisms that explain the evolution of cooperation. We review and briefly describe the literature on these mechanisms, and we explain their most important features, how they are to be considered along the conceptual dimensions used to describe the notion of mechanism, what is the empirical and computational evidence to support them, and which are the shortcomings that each of them has as explanatory hypotheses for the evolution of cooperation.

M. Salgado (✉) · J. A. Noguera · F. J. Miguel
GSADI: Analytical Sociology and Institutional Design Group, Department of Sociology, Universitat Autònoma de Barcelona, B Building Campus UAB, 08193 Bellaterra, Spain
e-mail: m.salgado@unab.cl

J. A. Noguera
e-mail: Jose.Noguera@uab.cat

F. J. Miguel
e-mail: Miguel.Quesada@uab.cat

M. Salgado
Escuela de Sociología, Facultad de Ciencias Sociales, Universidad Andres Bello, Avenida Quillota 980, Viña del Mar, Chile

Keywords Agent-based models · Causality · Evolutionary game theory · Explanation · Reciprocity

Introduction

Some of the most fundamental questions concerning our evolutionary origins, our interactions, and the organisation of society are centred on the issues of cooperation (i.e. when two or more individuals engage in joint actions that result in mutual benefit) and altruism (i.e. acts that benefit others at a personal cost). A wide range of phenomena can be related to the human-specific tendency to cooperate with others, from human morality and language to the emergence of social institutions (Boyd 2006; Nowak 2006; Warneken *et al.* 2007). Since the classical observations made by Mauss (2001) about 'the gift economy', social scientists have known that social behaviour is permeated by networks of cooperation and reciprocity. More recently, computational modelling has allowed researchers to explore, by experimental means, the conditions under which norms of cooperation have evolved (Axelrod and Hamilton 1981; Axelrod 1986, 1997; Nowak and Sigmund 1998).

Since human behaviour often leaves material traces that can be monitored archaeologically, it can be assumed that the social mechanisms that explain aggregated human behaviour can be of paramount interest for archaeologists. If cooperation does indeed play an important role in the evolution of human behaviour and social arrangements, we might be able to generate expectations about its past form and variation in specific ecological settings and assess those expectations archaeologically (Bird and O'Connell 2006, Cioffi-Revilla et al 2007). This is the reason why theories about the evolution of cooperation and computational models based on those theories are useful for archaeological research, and increasingly in use.

However, there are substantive aspects that are not tackled in the mainstream literature (i.e. evolutionary biology, anthropology and archaeology). On the one hand, the very concept of 'mechanism' is not usually defined or even questioned, so the epistemic and ontological status of the explanations of social behaviour based on this concept is ambiguous. On the other hand, since computer simulation is so relevant in testing and defining social mechanisms, the methodological drawbacks with computational methods also need to be addressed. Is the *generative sufficiency* enough to propose an evolutionary mechanism, or it also has to be evaluated in terms of plausibility? The way in which we solve these two questions will frame the proposed explanations.

In this paper, our objective is threefold. Firstly, we shall give an overview of the main literature on the concept of *social mechanism*, emphasizing its epistemological and ontological implications. Secondly, we shall explain when a computer simulation provides a good explanatory mechanism, in terms of generative sufficiency *and* plausibility. Thirdly, we shall relate this conceptualisation to the different mechanisms that have been proposed to explain the evolution of cooperation in humans. To do so, we describe the literature about the evolution of cooperation, revising five different explanatory mechanisms. We finish this article with some concluding remarks.

Social Mechanisms

Mechanisms are often mentioned as fundamental objects of scientific study, but only recently the implications of the concept have been thoroughly analysed. In this section, we review the current debate on the scientific role of mechanisms, which may be of interest for agent-based modelling (from here on, ABM) of evolutionary social dynamics. Before proceeding, it is worth noting that the concept of 'social mechanism' is defined by some social scientists as including only *social interaction* mechanisms (how an agent's beliefs, desires and actions are influenced by those of other agents with whom she interacts; see Hedström 2005); however, when studying evolutionary dynamics, this strict definition should probably be widened to include also *environmental mechanisms* (those that generate material constrictions to social behaviour) and *selection mechanisms* (how particular contextual conditions select types of agents).

Two main advantages of the social mechanisms approach have been claimed: knowledge about mechanisms increases the possibility of causal analysis in the absence of nomological laws (something important in disciplines that study complex systems with high contextual variability), and helps to open the 'black box' of social dynamics in order to provide the microfoundations of the observed phenomena. However, there is no agreement on the implications of the concept of social mechanism along different conceptual dimensions of the idea. Here, we will briefly identify these issues and present what we think is a suitable answer to each of them for disciplines that study evolutionary social dynamics.

Definition

There is no consensus on a unique definition of what a 'mechanism' is (see, for example, Hedström, 2005, p. 25, for a list of different definitions given in the literature). To our view, Machamer et al (2000, p. 3) provide a definition which is particularly well-suited for the social sciences: a mechanism consists of "entities and activities, organized such that they are productive of regular changes from start or set-up to finish or termination conditions". The dualistic nature of this definition makes it adequate for social sciences that deal with individual agents ('entities') and their actions ('activities'). According to Machamer et al (ibid.), entities (and their properties) "are the things that engage in activities", and activities "are the producers of change". It is then not surprising that analytical sociology has adopted this definition as canonical (Hedström 2005; Hedström and Bearman 2009a; Hedström and Ylikoski 2010). Besides, note that the definition has no 'mechanistic' implications in terms of 'contact forces' or 'spatiotemporal contiguity' like in classical 'mechanical' philosophers such as Galileo or Descartes (Woodward 2011).

Ontology

Are mechanisms 'real' entities and activities or are they just theoretical constructs (Hernes 1998; Gross 2009) or as-if stories (Friedman 1966)? Under the proposed definition, it

seems obvious that the underlying philosophy of the social mechanisms approach is *realist* and therefore differs from constructivist or as-if explanations: the emphasis is in providing knowledge about the 'cogs and wheels' (Elster 2007) that are responsible in the real world for the generation of the phenomena under study. It is important then not to confuse mechanisms with what Machamer *et al.* (2000, pp. 16–18) call 'mechanism schemata', which are abstract descriptions of types of mechanisms. These schemata are often implemented in ABM's algorithms and codes, and applied to the study of concrete real-world mechanisms (see several examples in Gilbert (2008), Hedström and Bearman (2009b), or Squazzoni (2012)).

Regularity

Mechanisms are often defined as regular patterns that do not take the form of covering laws, that is, of general nomological laws that correlate two types of events (Elster 2007; Hedström 2005; Hedström and Ylikoski 2010; Opp 2005). However, the degree of 'regularity' mechanisms have to show (how 'regular' they need to be) is also a contentious issue (Andersen 2012). There is no doubt that mechanisms as defined by Machamer et al. (ibid.) are regularities, for there is 'productive continuity' between their stages. This feature allows generalising a mechanism to some degree from one case to another. Although a mechanism' regularity typically lacks the certainty offered by a nomological covering law, it allows to avoid mere 'storytelling' in social science (Ylikoski 2011): a narrative account or historical description of how a given phenomena came about is not a mechanism; for a mechanism to exist, it has to be a pattern more general and more fundamental than the particular empirical series of events that generated a particular case or phenomenon (Barbera 2006).

Of course, this poses the problem of *indeterminacy* when trying to predict the presence of a mechanism or what results from its operation in a particular context (Elster 2007). As argued by some computational social scientists, ABM is precisely a "powerful virtual laboratory in which to design triggering conditions and to determine the resulting microscopic and macroscopic effects of concatenations of mechanisms" (Manzo 2012, p. 57). Although the mechanisms approach cannot render so much predictive power as the traditional covering law approach, it is much more realistic in complex systems where covering laws are scarce or non-existent; in fact, the study of mechanisms' triggering conditions—that is, of which are the contextual conditions that typically start the operation of a given mechanism—may improve our predictive capacities in fields where we cannot rely on any covering law.

Transparency

Knowledge of mechanisms may increase transparency by revealing adequate microfoundations for the phenomena under study, but how can we 'see' mechanisms in operation in a given system? Mechanisms are not always directly observable structures or processes (Hedström and Swedberg 1996), and there is also the possibility that two mechanisms cancel each other's effect, so we might wrongly infer that they are not present. ABM is again a useful tool for this, since any given mechanism

may be implemented virtually in the specifications of the computational code (Manzo 2012), thus providing with a virtual space where the mechanism's operation and results can be observed. Although ABM is not the only way to see mechanisms operating (see other methods applied to 'natural' systems in McAdam *et al*. (2008)), it is the more promising one (Hedström and Ylikoski 2010; Manzo 2012). However, the transparency of a mechanism's operation when implemented in an ABM is not always granted. An empirical calibration of ABM's parameters may help in observing which is the exact operation of a mechanism and to identify its presence isolating it from confounding factors.

Explanatory Power and Intelligibility

Although there is a general consensus that knowledge of mechanisms improves explanatory power, it has been also acknowledged that the identification of a mechanism may not be enough to explain a given phenomenon. Here, we face the problem of the appropriate selection of a mechanism or of the elements of a mechanism that have real explanatory relevance (Ylikoski 2011; Ylikoski and Kuorikoski 2010). Which among the many elements in the causal history of an event are relevant for its explanation? That is to say, how detailed should be our description of mechanisms, and which entities and activities should we include in them? Since mechanisms are often complex sets of activities and entities, how should we distinguish between the causally relevant and irrelevant parts of a mechanism's operation? (Couch 2011).

The answers to all these questions partially depend on a theory of causality and explanation. An increasingly consensual approach among philosophers of science (Woodward 2003; Ylikoski and Kuorikoski 2010) defines causality as counterfactual dependence between a cause variable c and an explanandum variable e, so that c causes e if we could bring about e by manipulating or bringing about c. Two apparently counterintuitive implications of this account are that identifying causes is not necessarily the same as providing microfoundations (see next section), and that causal claims can be made without any reference to mechanisms (Williamson 2011). Mechanisms are nonetheless essential for the theoretical *understanding* of that invariance and counterfactual dependence between c and e, and so for improving the explanatory understanding of the phenomena, if not for explanation as such. In other words, knowledge of mechanisms helps explanations to integrate causal claims with each other, to answer *what if* questions, and to generalise. Machamer *et al.* (2000, p. 21) emphasise that the key issue with mechanisms is intelligibility (perhaps more than explanation). To say it clearer, mechanisms make explanations intelligible, showing how explanans work to produce the explanandum. As Bunge claims, "no knowledge of mechanism, neither understanding nor control" (Bunge 2004, p. 206).

Reduction

As said in the previous section, a mechanism may improve explanatory power and intelligibility. It usually does so by providing microfoundations of the phenomenon to

be explained, that is to say, by ontologically reducing it to lower-order processes and elements (Stinchcombe 1998). In the social sciences, the reduction base for macrosocial phenomena and patterns is often that of individuals' properties and interaction: the idea that individual agents and their actions are the building blocks of social reality has the corollary that no intelligible mechanisms exist on the macro-level as such, so that "all macro-level change should be conceptualised in terms of three separate transitions (macro–micro, micro–micro, and micro–macro)" (Hedström and Swedberg 1996, p 299; see also Coleman 1990). If individual actions are to be explained by action–formation mechanisms that depart from agents' beliefs and desires, this implies accepting the extended Davidsonian view that some form of folk psychology (belief–desire psychology) is an explanatory legitimate ontology in the social sciences (Davidson 1963).

A usual concern is why to stop the search for social mechanisms at the individual or psychological level. The methodological adoption of stopping rules seems to be a sound answer: "Nested hierarchical descriptions of mechanisms typically bottom out in lowest level mechanisms. (…) Bottoming out is relative: Different types of entities and activities are where a given field stops when constructing mechanisms. The explanation comes to an end, and description of lower-level mechanisms would be irrelevant to their interests" (Machamer *et al*. 2000, p. 13). As Max Weber repeatedly claimed, scientists can only isolate objects for study and explain them on the basis of such stopping rules. For a mechanism to improve our explanatory power, it would be absurd to demand that all its integrating elements are themselves fully explained in terms of lower-level mechanisms, and so forth. As far as these elements really exist, nothing more is required for the explanation of a given phenomenon at the relevant ontological level (Hedström and Ylikoski 2010, p. 52).

Generative Sufficiency

In ABM, mechanisms are translated as the model *microspecifications*—the set of simple behavioural rules that specify how the agents behave and react to their local environment (Epstein and Axtell 1995; Epstein 2007). Once the population of agents and the environment are defined, the researcher can implement the microspecifications and run the computer simulation in order to evaluate whether these rules bring about or 'generate' the macro phenomenon of interest, over the simulated time. The motto of ABM is then: 'if you did not grow it, you did not explain it'. A mechanism's generative sufficiency is a condition for it to have explanatory power.

Whether to conceive the explanations formalised in an agent-based model as based on generative principles or, besides, in plausible or even empirical grounds is one of the current debates on ABM (Epstein and Axtell 1995; Hedström 2005). On the one hand, when the agent-based model cannot generate the outcome to be explained, the microspecification is not a candidate explanation of the phenomenon and the researcher has demonstrated the hypothesised mechanism to be false. On the other hand, when the agent-based model can generate the type of outcome to be explained, then the researcher has provided a computational demonstration that a given microspecification (or mechanism) is in fact sufficient to generate the macrostructure

of interest. Generative sufficiency, thus, provides a candidate mechanism-based explanation of the macro-phenomenon.

Empirical Adequacy

However, the fact that a given mechanism (or combination of mechanisms) is a good candidate explanation is not sufficient to provide an explanatory mechanism. The fact that a hypothesised mechanism can generate, in a computer program, an observed outcome does not mean that it *actually* explains it. As Gilbert demonstrated with the case of Schelling's segregation model (Gilbert 2002), many different mechanisms can generate a similar outcome.

Having different alternative mechanisms for the same outcome, we must be somehow able to identify the mechanism that most likely does generate it. This is where *plausibility* enters the picture. The foundation of any explanation of a social phenomenon needs a plausible behavioural model (either psychological or sociological) of individual action assumed at the micro level, because, otherwise, "we would simply be telling an as-if story, not detailing the actual mechanisms at work" (Hedström 2005, p. 35). Therefore, the agent-based modeller whose aims are explanatory should, first, be generally consistent with well-known facts and accepted scientific evidence and, second, use relevant and reliable data in order to empirically calibrate the microspecifications of his model.

Of course, the need for empirical adequacy raises several important issues (which, for space reasons, we cannot address here). Let us mention one as an example: where is the empirical calibration to be done, that is, where do empirical data enter in the model? Leaving aside validation, empirical data may be used to calibrate the initializing conditions of the model, the value of exogenous parameters, the behavioural rules governing the agents, or the environmental conditions in which they act. Generally speaking, it could be thought that the more aspects of the model that are empirically calibrated the better; but of course this may depend on the aims of the model (in particular, whether its purpose is empirical explanation or not; see Hedström 2007), on the costs of obtaining the relevant empirical data, or on their quality and reliability; for instance, when data calibration is not feasible, reliance on sensitivity analysis and/or accepted and well-tested particular theories might be advisable (Gilbert 2008; Salgado and Gilbert 2013).

The Evolution of Cooperation: A Formal Approach

One of the most important applications of ABM has been on the evolution of cooperation. In computer simulations, the social interactions among agents are often modelled using game theory. The prisoner's dilemma (Axelrod and Hamilton 1981; Axelrod 1997, 2006) has become one of the leading paradigms to explain cooperative behaviour in the biological and the social sciences (Colman 1995): it is a canonical example of a game that shows why two players might not cooperate, even if it appears that it is in their best interest to do so. The word *dilemma* is key: players are in a situation in which they have to decide on one action among a set of possible

actions taking into account what the other player *should* decide, which is uncertain. Let us define formally this game as an example (other well-known dilemmas used in evolutionary game theory are the security game or the stag-hunt game; see Gintis 2000; Skyrms 2003).

Let N_t be a well-mixed population in time t, in which any two agents interact with the same probability. Social interactions occur in each simulation step t, when agents attempt to find a partner to play the prisoner's dilemma. The game is usually played between pairs of agents randomly selected, but this might depend on the mechanism that has been implemented in the model. In each interaction, agents might cooperate or defect, because cooperation and defection are the two possible actions that are defined by the prisoner's dilemma. Behaviour $a=1$ corresponds to cooperation and $a=2$ to defection. The interactions, therefore, might report individual benefits or losses. The success of the interactions for each individual—which is quantified in terms of 'payoffs'—depends on her own strategy i and the strategy j of the respective interaction partner. When the payoff for the individual i is P_{ij}, the payoff for the interaction partner is P_{ji}. The payoff matrix of the prisoner's dilemma that summarises the possible payoffs is given by:

$$P = \left(P_{ij}\right) = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{matrix} \\ C \\ D \end{matrix} \begin{matrix} \begin{matrix} C & D \end{matrix} \\ \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{matrix} \tag{1}$$

The entries of the payoff matrix (1) refer to the row player. If a cooperator, C, interacts with another cooperator, both get payoff R, which is the "reward" for mutual cooperation. If a cooperator, C, interacts with a defector, D, the cooperator gets the "sucker's payoff," S, whereas the defector gets the highest payoff of the game, T, which denotes the "temptation to defect." Two defectors obtain the payoff P, which stands for the "punishment" of mutual defection. Therefore, $P_{11}=R$, $P_{12}=S$, $P_{21}=T$, and $P_{22}=P$. The game is a prisoner's dilemma if for each player their action preferences are ordered by $T>R>P>S$. Finally, given that in the simulation, the game is repeatedly played by two players, the following condition should be added: $2\cdot R>T+S$, to prevent alternating cooperation and defection giving a greater reward than mutual cooperation. The payoff matrix (1), together with these two conditions, defines a useful artifact to model and simulate the evolution of cooperation and defection strategies among rational agents.

Five Mechanisms for the Evolution of Cooperation

The 'one-shot' prisoner's dilemma (that is, the game played only once) has only one result that is a *Nash equilibrium* (Nash 1950): mutual defection—an outcome that is Pareto inferior. In such a nonrepeated prisoner's dilemma, it is best to defect no matter which strategy is adopted by the other player (because $R<T$ and $S<P$). Consequently, the evolution of cooperation requires specific mechanisms that allow natural selection to favour cooperation over defection. There are at least five mechanisms that have been proposed to explain the emergence and evolution of cooperation and altruism (Choi and Bowles 2007; Nowak 2006), namely: kin selection, direct reciprocity, indirect reciprocity, group selection, and parochial altruism. Since these are proposed

as explanatory mechanisms, we can use the conceptual dimensions we discussed above (see "Social Mechanisms" section) to characterise them (the empirical adequacy of the mechanisms is usually assessed by reference to well-known evolutionary and behavioural evidence). Table 1 compares each mechanism for the evolution of cooperation according to these dimensions. Let us define and explain each of these mechanisms.

Kin Selection

Kin selection or *inclusive fitness* theory (Hamilton 1964; Smith 1964) explains the evolution of cooperation by taking the 'micro level' or genetic view, which accounts all fitness effects back to the individual gene and claims that just a gene can be favoured by natural selection, by increasing the reproductive success of its bearer and by increasing the reproductive success of other individuals that carry *the same* gene. According to this explanation, kin selection is biologically adaptive because the individual beneficiaries of this type of behaviour—the altruist actor's nondescendent relatives—share some genes with the individual who helps them; for this reason, the survival and reproduction of the beneficiaries contribute to the propagation of the altruist's genes. Although kin selection models were proposed in archaeological research as an explanatory mechanism to food sharing systems (cf. Isaac 1981), researchers have started to question whether kin altruism alone can maximise the indirect fitness of the recipients and explain cooperative behaviour in humans and other species (Clutton-Brock 2002; Griffin and West 2002).

Although this mechanism seems to be sufficiently generative, especially in small groups, there are reasons to believe that it is not plausible. For instance, several species of *cooperative breeders* (cooperative breeding refers to any species with alloparental assistance in both the care and provisioning of young) have shown that helpers can be unrelated to the young they are raising (they may even be from different species) and that unrelated helpers invest as heavily as close relatives (Clutton-Brock *et al*. 2000), a feature that is observed in cooperative breeding monkeys such as marmosets and tamarins (Digby and Ferrari 1994; Sussman and Garber 1987) and humans (Hrdy 2009). Consequently, although altruism is undoubtedly most common in familial groups (Emlen 1995), it is by no means restricted to groups of closely related members. Thus, kin selection provides a viable explanation for the evolution of cooperative breeding in many vertebrate and invertebrate species, but it does not constitute a strong mechanism to explain the evolution of altruism in general.

Direct Reciprocity

If there are repeated encounters between the same two individuals, then direct reciprocity can emerge and lead to the evolution of cooperation (Trivers 1971). Thus, whilst in the 'one-shot' prisoner's dilemma it is always best to defect, the repeated (or iterated) prisoner's dilemma opens a doorway for cooperation to emerge: the expectation of future interactions makes cooperation an attractive option. Direct reciprocity

Table 1 Five mechanisms that may explain the evolution of cooperation

	Kin selection	Direct reciprocity	Indirect reciprocity	Group selection	Parochial altruism
Definition of the mechanism	The beneficiaries share some genes with the altruist; the survival and reproduction of the beneficiaries contribute to the propagation of the altruist's genes.	Both the beneficiaries and the cooperators face repeated encounters; the cooperator's decision to cooperate is based on what the beneficiary has done to her in previous encounters.	Both the beneficiaries and the altruists interact with each other only occasionally; the altruist's decision to cooperate is based on information (i.e. reputation) about what the beneficiary has done to others in previous encounters.	The altruists' genes can become fixed within certain groups because of the benefits they bestow on those groups as wholes, even when the effect of these alleles on individuals' fitness is negative. These groups reproduce faster and outcompete other groups made up of self-interested individuals.	The altruists cooperate with group members and are hostile to individuals from other groups. In contexts of intergroup conflict, internally cooperative groups prevail over less cooperative rival groups.
Ontology Is the mechanism really at operation or just an as-if story?	Realist	Realist	Realist	Realist	Realist
Regularity In which instances can the mechanism be generalised from one case to another?	It works in small familial groups.	It works in contexts of repeated encounters within small groups.	It works when information about the beneficiaries' reputation is available.	It works when the group benefit of the altruists' actions is higher than the altruists' individual costs.	It works when there is severe intergroup conflict.
Transparency Is the presence of the mechanism observable?	Yes	Yes	Yes	Yes	Yes
Intelligibility	Yes	Yes	Yes	Yes	Yes

Table 1 (continued)

	Kin selection	Direct reciprocity	Indirect reciprocity	Group selection	Parochial altruism
Does the presence of the mechanism improve the intelligibility of the explanation?					
Reduction base What are the mechanism's microfoundations?	Individuals' genes.	Individuals in repeated interactions.	Individuals in occasional interactions.	Individuals' genes within groups that compete with each other.	Individuals within groups that compete with each other.
Generative sufficiency Can the mechanism generate the observed phenomenon?	Yes	Yes	Yes	Yes	Yes
Empirical adequacy Is the mechanism consistent with well-known evidence and/or does it fit the available empirical data?	Although it might work in small and familial groups with little migration, it does not explain cooperation to unrelated individuals	Although it might work in small groups, it is a weak mechanism in larger groups.	Although it might work in larger groups, it requires dense connections or cultural artefacts that register individuals' reputation	Although it might work in small groups, it requires within-group homogeneity, between group heterogeneity and low migration rates	Although it might work in contexts of violent intergroup conflict, it does not apply in contexts in which groups cooperate among them

is based on the idea "I help you and you help me." Direct reciprocity favours are exchanged directly and repeatedly between individuals: A helps B and, in return, B helps A. In a nutshell, for direct reciprocity, B's decision to cooperate with A is based on how A has treated B in previous encounters. Axelrod and his simulated tournaments during the 1980s (Axelrod and Hamilton 1981) were the catalyst in discovering the best, most robust strategies for playing the iterated prisoner's dilemma. One of the simplest and most effective is called *tit-for-tat* where a player reciprocates the behaviour of the other player in their previous game so that co-operation is rewarded with cooperation and defection is punished with defection. Tit-for-tat cannot be selected unless the agents holding this strategy have a 'good chance' of competing against the same opponents more than once during their life. Good chance means that competition continues long enough for repeated punishment and forgiveness to generate a long-term payoff higher than the possible loss from cooperating initially.

For tit-for-tat to be a successful strategy in the long run, the probability of playing repeatedly with the same agent has to be high. This means that as population size increases, tit-for-tat produces lower payoffs and, over time, it is beaten by other more successful evolutionary strategies. In a nutshell, with direct reciprocity, the power of direct retaliation decreases as population size increases.

Even more important, ethnographic and archaeological research has shown that people in small-scale societies routinely have important (i.e. fitness relevant) interactions that are short-term or one-shot (Fehr and Henrich 2003). Therefore, it seems that direct reciprocity is rooted in a false, but widely believed anthropological myth about the nature of life in small-scale societies and, therefore, it is plausible just for some social contexts. Thus, although this mechanism is sufficiently generative, it should be applied with caution by archaeologists, especially in big groups of individuals, or in 'noisy environments' (e.g., social exchanges across many domains).

Indirect Reciprocity

Unlike other primate groups, hominid groups grew to sizes that could not function exclusively on the basis of kin selection (commitment falls off precipitously as genetic distance increases between individuals) or direct reciprocity (ability to directly monitor trustworthiness in reciprocation decreases rapidly as the number of transactions multiply). Under *indirect reciprocity*, individuals interact with each other only occasionally (sometimes only once), but individuals have access to information about the past behaviour of the individual with whom they are about to interact. In a nutshell, for indirect reciprocity, my decision is based on what you have done to others. As Taylor and Nowak (2007, p. 2284) remind us, "indirect reciprocity arises out of direct reciprocity in the presence of interested audiences." Indirect reciprocity crystallises the idea "I help you and somebody will help me." It is based on reputation (Nowak and Sigmund 1998, 2005). Each event can be seen as an interaction between two agents, according to a single prisoner's dilemma game given by Payoff Matrix (1). Others observe each game. Cooperation is costly, but leads to the reputation of being a helpful individual. Defection is more profitable in the short run, but leads to a bad reputation. Natural selection favours strategies that base their decision to cooperate or to defect on the reputation of oneself and of others. Experimental studies

confirm that helpful individuals are more likely to receive help in the future (Bolton *et al.* 2005; Rockenbach and Milinski 2006).

Extensive theoretical and computer simulation research has shown that the availability of *accurate* reputation information is the key to indirect reciprocity's ability to solve the puzzle of cooperation (Henrich and Henrich 2006; Panchanathan and Boyd 2003). This means that, *ceteris paribus*, variables such as the size of the cooperative group (the number of individuals in any given situation), the population size (the number of individuals in the pool of potential interactants), the density of the social connections between individuals in the population and people's beliefs about gossip will strongly influence the effectiveness of indirect reciprocity (Henrich and Henrich 2006). Indirect reciprocity is a plausible mechanism to sustain cooperation within dense, bounded social networks that are stable through time, which lead to the highest levels of indirect reciprocity-based cooperation.

Group Selection

Kin selection, direct reciprocity or indirect reciprocity cannot explain altruism to non-kin when costs of cooperation are high, reciprocation unlikely and there is no information about reputation. Heroism in warfare is an example. Explaining such extravagant behaviour via indirect benefits to altruists and their kin has proved difficult. A growing body of work seeks instead to explain altruism with models that include selection on both individuals and groups. *Group selection* is based on the idea that competition occurs not only between individuals but also between groups (Wilson and Sober 1994). These models have been increasingly used in the last decades and have provided a 'Darwinian' ground for the co-evolutionary dynamics between genes and culture (Gintis *et al.* 2003; Richerson and Boyd 2005; Sober and Wilson 1999). According to this hypothesis, altruistic alleles can become fixed or spread within certain groups because of the benefits they bestow on those groups as *wholes*, even when the effect of these alleles on the individuals' fitness is negative. The groups in which, by chance, altruistic strategies (i.e. unconditional cooperators) are selected and spread will be fitter, *as groups*, and they will outperform other groups made up of self-interested individuals. A growing body of work explains altruism with models that include selection on both individuals and groups. In such 'multi-level' models, the evolutionary outcome depends on the relative impact of competing pushes and pulls at individual and group levels. Individual selection pushes counter-productive behaviours like altruism out of the gene pool. Group selection exerts a contrary pull, favouring groups with many altruists over groups of more selfish folk. In most species, individual selection wins out. For humans living in small groups (as our ancestors did), however, a strong group selection pull is plausible (Wright 2007).

However, as Dietz *et al.* (1990), group selection will prevail only when group extinction rates are high, within-group genetic homogeneity is high, between-group heterogeneity is also high, and migration rates are low. Most human populations do not meet these conditions. Thus, it does not seem that group selection can provide a genetic basis for altruism. For this reason, Richerson and Boyd (2005) state that the outcome of the group selection mechanism depends on the relative amount of variation within and between groups. If group members are closely related, most of

the variation will occur between groups. This is easiest to see if groups are composed of clones (as in colonial invertebrates such as corals). Then there is almost no genetic variation within groups; all the variation is between groups, and selection acts to maximise group benefit.

Parochial Altruism

Alternative explanatory mechanisms have focused on *warfare* and *group competition* as a force for robust group selection. Among those explanatory models, *parochial altruism* has gained recent popularity among scholars. This model highlights the idea that individuals are altruistic with group members and are hostile to individuals not of one's own group (Arrow 2007; Bowles 2006; Choi and Bowles 2007). Since internally cooperative groups prevail over less cooperative rival groups, parochial altruism rests on the evolutionary belief that violent intergroup conflict played a key role in the dawn of human cooperation. Recent evidence seems to be consistent with this hypothesis. Gneezy and colleagues demonstrated in controlled cooperative dilemmas with real players that violent intergroup conflict led individuals to inflict costs on free-riders and bestow benefits on cooperators (Gneezy and Fessler 2011). Additionally, by empirically calibrating his model of between-group competition, Bowles (2009) established that intergroup conflict can promote the evolution of altruism. More recently, Mathew and Boyd (2011) showed that in warfare among nomadic Turkana pastoralists in East Africa, costly cooperation in combat is sustained through punishment of free-riders. From an evolutionary viewpoint, the hypothesis that warfare provided a selective pressure that favoured internally cooperative groups assumes that, during earlier stages of human evolution, exploitation, widespread carnage and intergroup competition for mating opportunities, access to resources and status was the norm (Keeley 2001).

However, such generalisations are unlikely to correctly describe the conditions in which our Pleistocene ancestors lived, so parochial altruism seems to be implausible. Small bands of hunter–gatherers, numbering 25 or so individuals, under chronic climate fluctuation, widely dispersed over large areas and unable to fall back on staple foods, would have suffered from high mortality rates, particularly child mortality, due to starvation as well as predation and disease, so they hardly would have been able to sustain warfare against competing groups (Johnson and Earle 2001). All these socio-ecological features rule out the idea that conflict and warfare among groups was a widespread phenomenon. Evidently, to claim the opposite, that is, to question the idea that persistent intergroup conflict—sufficiently common to produce the selection of genetic or even cultural advantageous traits—during the Pleistocene era does not lead one to endorse the myth of the 'noble savage', as some authors have claimed (among others, Pinker 2003); it means to take into account the socio-ecological conditions of our foraging ancestors seriously. Those conditions would have precluded foraging individuals from engaging in intergroup mayhem. Thus, far from increasing their fitness by competition and conflict, these individuals would have enhanced their chances through improved access to resources, mating opportunities or safety by "eschewing efforts to achieve inter-community dominance in favour of egalitarian relations of friendship, mutuality, and sharing" (Kelly 2005, p.

15297), so nearby members of their own species would have been more valuable as potential sharing partners.

Concluding Remarks

A major goal of biology and economic theory has been to explain how wide-scale cooperation among self-regarding individuals occurs in a decentralised setting. To explain this, we need to propose plausible mechanisms that must be later tested in formal ways. We have started this paper by addressing the notion of 'social mechanisms', clarifying its epistemological and ontological status, and we have also identified some methodological drawbacks with computational modelling, particularly with ABM.

We have also explained five mechanisms for the evolution of cooperation that may be implemented in agent-based models. These mechanisms have pros and cons, and they match specific socio-ecological conditions the modeller must identify empirically in the target system to be modelled; otherwise, the explanatory mechanism might be implausible. Kin selection offers a good explanatory mechanism for the evolution of cooperation in small and familial groups with little migration; when these conditions do not apply, modellers should not use it as a mechanism for the evolution of cooperation. Direct reciprocity overcomes some of the drawbacks that kin selection creates, since it can explain the evolution of cooperation among (genetically) unrelated individuals. However, when population size increases, the power of direct retaliation decreases—since reciprocation becomes unlikely—so this mechanism should not be used in big groups. Indirect reciprocity provides a good alternative for larger groups of agents, since punishment to defectors is based on information about what group members have done to others. This mechanism requires identifying the specific ways in which 'reputation' circulates within the group (e.g., gossips, symbolic artefacts). Group selection models are based on the idea that, even though the effect of cooperation and altruism on the individuals' fitness might be negative, the group effect of these behaviours can compensate the individual losses, so natural selection can favour groups with many altruists and cooperators. Group selection requires small group size and low migration rates between groups. Parochial altruism is a special case of group selection, based on the idea that intergroup conflict killed a nontrivial proportion of our ancestors. War is a strong candidate because people kill each other on the basis of group membership. However, the specific condition in which our Pleistocene ancestors lived rule out the idea that warfare among groups was a widespread phenomenon. Again, these are issues that have to be addressed empirically.

The reviewed mechanisms in this paper are all well-defined, and they all satisfy minimum requirements of ontological realism, transparency, intelligibility and generative sufficiency to produce social cooperation in the long run. They differ in how regular they are depending on circumstances such as group size, availability of information on agents' reputation, the costs of altruist actions or the existence of intergroup conflict. Their reduction base is always in the individual level, although in some cases they also refer to genetic selection. All of these mechanisms are sufficient to generate the emergence and evolution of cooperation under specific empirical

conditions. Finally, their empirical adequacy seems to be contextual and bounded to specific empirical pre-requisites in each case. Therefore, in addition to generative sufficiency, empirical adequacy is a condition for satisfactory mechanism-based explanations tested through ABM. This suggests that fine empirical calibration and enrichment of agent-based models aimed to explain the emergence of social cooperation is important and even necessary to adequately isolate the mechanism at work in each particular case of social evolution; it also invites to think of combinations of different mechanisms, instead of single-mechanism explanations, in order to generate evolutionary dynamics which can fit observed data and patterns (such as those contained in different archaeological records).

Acknowledgments The National Plan for R&D has supported this work through the CONSOLIDER-INGENIO 2010 Programme of the Spanish Ministry of Science and Innovation (MICINN, grant no. CSD2010-00034-SimulPast). The authors also thank additional financial support from MINECO, grant no. CSO2012-31401, and two anonymous referees for their comments and suggestions.

References

Andersen, H. (2012). The case for regularity in mechanistic causal explanation. *Synthese, 189*(3), 415–432.
Arrow, H. (2007). Evolution: the sharp end of altruism. *Science, 318*(5850), 581–582.
Axelrod, R. (1986). An evolutionary approach to norms. *The American Political Science Review, 80*(4), 1095–1111.
Axelrod, R. (1997). *The complexity of cooperation: agent-based models of competition and collaboration*. New Jersey: Princeton University Press.
Axelrod, R. (2006). *The evolution of cooperation* (Revth ed.). London: Penguin.
Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396.
Barbera, F. (2006). A star is born? The authors, principles and objectives of analytical sociology, *Papers. Revista de Sociologia, 80*, 31–50.
Bird, D. W., & O'Connell, J. F. (2006). Behavioral ecology and archaeology. *Journal of Archaeological Research, 14*(2), 143–188.
Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics, 89*(8), 1457–1468.
Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science, 314*(5805), 1569–1572.
Bowles, S. (2009). Did warfare among ancestral hunter–gatherers affect the evolution of human social behaviors? *Science, 324*(5932), 1293–1298.
Boyd, R. (2006). The puzzle of human sociality. *Science, 314*(5805), 1555–1556.
Bunge, M. (2004). How does it work? The search for explanatory mechanisms. *Philosophy of the Social Sciences, 34*(2), 182–210.
Choi, J.-K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science, 318*(5850), 636–640.
Cioffi-Revilla, C., Luke, S., Parker, D. C., Rogers, J. D., Fitzhugh, W. W., & Honeychurch, W. (2007). Agent-based modeling simulation of social adaptation and long-term change in inner Asia. In S. Takahashi, D. Sallach, & J. Rouchier (Eds.), *Advancing social simulation: the First World Congress*. Kyoto: Springer.
Clutton-Brock, T. (2002). Breeding together: kin selection and mutualism in cooperative vertebrates. *Science, 296*(5565), 69–72.
Clutton-Brock, T. H., Brotherton, P. N. M., O'Riain, M. J., Griffin, A. S., Gaynor, D., Sharpe, L., et al. (2000). Individual contributions to babysitting in a cooperative mongoose. *Suricata suricatta*. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 267*(1440), 301–305.
Coleman, J. S. (1990). *Foundations of social theory*. Cambridge: Belkmap.
Colman, A. M. (1995). *Game theory and its applications in the social and biological sciences*. London: Routledge.
Couch, M. B. (2011). Mechanisms and constitutive relevance. *Synthese, 183*(3), 375–388.
Davidson, D. (1963). *Actions, reasons and causes. Essays on actions and events*. Oxford: Oxford University Press.

Dietz, T., Burns, T. R., & Buttel, F. H. (1990). Evolutionary theory in sociology: an examination of current thinking. *Sociological Forum, 5*(2), 155–171.
Digby, L. J., & Ferrari, S. F. (1994). Multiple breeding females in free-ranging groups of *Callithrix jacchus*. *International Journal of Primatology, 15*, 389–397.
Elster, J. (2007). *Explaining social behavior*. New York: Cambridge University Press.
Emlen, S. T. (1995). An evolutionary theory of the family. *Proceedings of the National Academy of Sciences, 92*(18), 8092–8099.
Epstein, J. M. (2007). *Generative social science: studies in agent-based computational modeling*. New Jersey: Princeton University Press.
Epstein, J. M., & Axtell, R. L. (1995). *Growing artificial societies: social science from the bottom up*. Washington, D.C.: Brookings Institution.
Fehr, E., & Henrich, J. (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation*. Cambridge, Mass: MIT Press.
Friedman, M. (1966). *The methodology of positive economics. Essays in positive economics*. Chicago: University of Chicago Press.
Gilbert, N. (2002). Varieties of emergence. Paper presented to Agent 2002 Conference: Social Agents: Ecology, Exchange, and Evolution, Chicago.
Gilbert, N. (2008). *Agent-based models*. London: Sage.
Gintis, H. (2000). *Game theory evolving: a problem-centered introduction to modeling strategic interaction: a problem-centered introduction to modeling strategic behaviour*. New Jersey: Princeton University Press.
Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*(3), 153–172.
Gneezy, A., & Fessler, D. M. T. (2011). *Conflict, sticks and carrots: war increases prosocial punishments and rewards*. Biological Sciences: Proceedings of the Royal Society B.
Griffin, A. S., & West, S. A. (2002). Kin selection: fact and fiction. *Trends in Ecology and Evolution, 17*(1), 15–21.
Gross, N. (2009). A pragmatist theory of social mechanisms. *American Sociological Review, 74*, 358–379.
Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology, 7*(1), 1–16.
Hedström, P. (2005). *Dissecting the social: on the principles of analytical sociology*. Cambridge: Cambridge University Press.
Hedström, P. (2007). Actions and networks: sociology that really matters (to me). *Sociologica, 1*(1), 1–18.
Hedström, P., & Bearman, P. (2009a). What is analytical sociology all about? An introductory essay. In P. Hedström & P. Bearman (Eds.), *The Oxford handbook of analytical sociology*. Oxford: Oxford University Press.
Hedström, P., & Bearman, P. (Eds.). (2009b). *The Oxford handbook of analytical sociology*. Oxford: Oxford University Press.
Hedström, P., & Swedberg, R. (1996). Social mechanisms. *Acta Sociologica, 39*(3), 281–308.
Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67.
Henrich, J., & Henrich, N. (2006). Culture, evolution and the puzzle of human cooperation. *Cognitive Systems Research, 7*(2–3), 220–245.
Hernes, G. (1998). Real virtuality. In R. Swedberg & P. Hedström (Eds.), *Social mechanisms. An analytical approach to social theory*. Nueva York: Cambridge University Press.
Hrdy, S. B. (2009). *Mothers and others: the evolutionary origins of mutual understanding: the origins of understanding*. Harvard: Harvard University Press.
Isaac, G. L. (1981). Archaeological tests of alternative models of early hominid behaviour: excavation and experiments. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 292*(1057), 177–188.
Johnson, A. W., & Earle, T. K. (2001). *The evolution of human societies: from foraging group to agrarian state*. Stanford: Stanford University Press.
Keeley, L. H. (2001). *War before civilization: the myth of the peaceful savage (reprint.)*. New York: Oxford University Press.
Kelly, R. C. (2005). The evolution of lethal intergroup violence. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15294–15298.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Manzo, G. (2012). Reason-based explanations and analytical sociology. *European Review of the Social Sciences, 50*(2), 35–65.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences, 108*(28), 11375–11380.
Mauss, M. (2001). *The gift: form and reason for exchange in archaic societies* (2nd ed.). New York: Routledge.
McAdam, D., Tarrow, S., & Tilly, C. (2008). Methods for measuring mechanisms of contention. *Qualitative Sociology, 31*(4), 307–331.
Nash, J. F. (1950). Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48–49.
Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560–1563.
Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*(6685), 573–577.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298.
Opp, K.-D. (2005). Explanations by mechanisms in the social sciences. Problems, advantages, and alternatives. *Mind and Society, 4*(2), 163–178.
Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology, 224*(1), 115–126.
Pinker, S. (2003). *The blank slate: the modern denial of human nature*. London: Penguin.
Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: how culture transformed human evolution*. Chicago: University of Chicago Press.
Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444*(7120), 718–723.
Salgado, M., & Gilbert, N. (2013). Agent-based modeling. In T. Teo (Ed.), *Handbook of quantitative methods for educational research*. New York: Sense Publisher.
Skyrms, B. (2003). *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.
Smith, J. M. (1964). Group selection and kin selection. *Nature, 201*, 1145–1147.
Sober, E., & Wilson, D. S. (1999). *Unto others: the evolution and psychology of unselfish behavior*. London: Harvard University Press.
Squazzoni, F. (2012). *Agent-based computational sociology*. Chichester: Wiley.
Stinchcombe, A. (1998). Monopolistic competition as a mechanism: corporations, universities, and nation-states in competitive fields. In P. Hedström & R. Swedberg (Eds.), *Social mechanisms: an analytical approach to social theory* (pp. 267–305). Cambridge: Cambridge University Press.
Sussman, R. W., & Garber, P. A. (1987). A new interpretation of the social organization and mating system of the Callitrichidae. *International Journal of Primatology, 8*, 73–92.
Taylor, C., & Nowak, M. A. (2007). Transforming the dilemma. *Evolution, 61*(10), 2281–2292.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology, 5*(7), 1414–1420.
Williamson, J. (2011). Mechanistic theories of causality part II. *Philosophy Compass, 6*(6), 433–444.
Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences, 17*(04), 585–608.
Woodward, J. (2003). *Making things happen. A theory of causal explanation*. Oxford: Oxford University Press.
Woodward, J. (2011). Mechanisms revisited. *Synthese, 183*(3), 409–427.
Wright, J. (2007). Cooperation theory meets cooperative breeding: exposing some ugly truths about social prestige, reciprocity and group augmentation. *Behavioural Processes, 76*(2), 142–148.
Ylikoski, P. (2011). Social mechanisms and explanatory relevance. In P. Demeulenaere (Ed.), *Analytical sociology and social mechanisms*. Cambridge: Cambridge University Press.
Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies, 148*(2), 201–219.