Simulation Models of the Evolution of Cooperation as Proofs of Logical Possibilities. How Useful Are They?

Published in: Etica & Politica / Ethics & Politics, XV, 2013, 2, pp. 101- 138, URL: hdl.handle.net/10077/9679.

by Eckhart Arnold
University of Düsseldorf (Germany)

July, 2nd 2013

Abstract

This paper discusses critically what simulation models of the evolution of cooperation can possibly prove by examining Axelrod's "Evolution of Cooperation" (1984) and the modeling tradition it has inspired. Hardly any of the many simulation models of the evolution of cooperation in this tradition have been applicable empirically. Axelrod's role model suggested a research design that seemingly allowed to draw general conclusions from simulation models even if the mechanisms that drive the simulation could not be identified empirically. But this research design was fundamentally flawed, because it is not possible to draw general empirical conclusions from theoretical simulations. At best such simulations can claim to prove logical possibilities, i.e. they prove that certain phenomena are possible as the consequence of the modeling assumptions built into the simulation, but not that they are possible or can be expected to occur in reality I suggest several requirements under which proofs of logical possibilities can nevertheless be considered useful. Sadly, most Axelrod-style simulations do not meet these requirements. I contrast this with Schelling's neighborhood segregation model, the core mechanism of which can be retraced empirically.

Keywords:
Evolution of Cooperation, Epistemology of Simulations, Critique of Modelling

Table of Contents

1. Introduction

In this article I am going to discuss the question what simulation models of the evolution of cooperation can possibly prove. I am going to discuss the question mainly in a historical context of models of the reiterated Prisoner's Dilemma that have become tremendously popular in the aftermath of Robert Axelrod's “Evolution of Cooperation” (1984). At the time of their publication Axelrod's simulation studies set the role model for a new approach to studying the evolution of cooperation. Despite its popularity this approach has never really been able to bridge the gap between the simulation models and empirical research. Moreover, it soon became apparent that far less can be deduced from pure simulation models about how the evolution of cooperation actually takes place than Axelrod had hoped for. In the first part of this article I am going to explain why this is the case and why the research design that Axelrod's study suggests is flawed: Simulations of the reiterated Prisoner's Dilemma do not allow us to draw general conclusions about the evolution of cooperation. At most they can prove logical possibilities that is, they can prove that certain phenomena are possible as the consequence of the modeling assumptions built into the simulation. But this does neither entail that the same phenomena can be expected to occur in reality nor that they are in reality the consequences of the same factors as those represented in the simulation model, *if* they occur in real-

ity. It might even still be the case that the phenomena produced by the simulation are impossible in reality. Thus, simulations of this kind can only prove logical possibilities, but not real possibilities.

Instead of speaking of "logical possibilities" one could also speak of simulations that prove "theoretical possibilities" or "proof of concept"-simulations. In the following, I use the terms "logical possibility", "theoretical possibility" and "proof of concept" synonymously. However, I prefer to avoid the term "theoretical possibility", because the wording suggests a deeper insight which – in my opinion – highly stylized simulation models do not confer. Notwithstanding the terminology the central question is how much, if any, insight we have gained into a natural phenomenon if we have been able to prove certain logical possibilities concerning this phenomenon.

The second section of this paper, therefore, is dedicated to the question under what circumstances the proof of logical possibilities via computer simulations may provide valuable insights. I describe three sets of circumstances under which this may be the case:

1. If the logical possibility demonstrates some thing that in virtue of our prior beliefs and background knowledge is highly surprising or totally unexpected to us, or which we would not even have considered possible at all. I call this the *novel discovery* condition.
2. If the logical possibility is a key element in a best explanation of some phenomenon. This is the case, if the explanation of some phenomenon merely hinges on the question whether the phenomenon can be produced by a particular mechanism and if this can be demonstrated by a simulation. This can be called the *best explanation* condition.
3. If the logical possibility is also a real possibility (the difference between logically possible and really possible will be explained later) that can at least in principle be identified in some particular empirical setting then, again, it is useful to know this possibility. I call this the *real possibility* condition.

These three conditions are to be considered disjunctive, that is, it suffices if one of them is met to render a theoretical computer simulation epistemically valuable. For the last case, the *real possibility* condition, Schelling's neighborhood segregation model (again, a well known example from the history social simulations) will be discussed as an example. It will be argued that the logical possibility of neighborhood segregation produced by the mechanism postulated by Schelling can be considered a real possibility, if the causal connection it describes can be traced by appropriate empirical methods. However, even a real possibility may not be the actual cause of some phenomenon, but it needs to be checked against other possible causes first.

The article concludes with a brief summary of the most important points and conclusions from the previous sections.

I would like to emphasize that I am aware that Axelrod's approach is history and that today "Evolution of Cooperation" is the heading for a much wider array of research programs, many of which certainly avoid the deficiencies of their ancestral role model. But my point is not to criticize the ongoing research on the evolution of cooperation. Instead my aim is to better understand the epistemological challenges of justifying simulation studies that remain purely theoretical and prove no more than logical possibilities. Arguably, for this purpose the examination of a well known historical example is better suited than meddling with current research. Despite the enormous popularity of Axelrod's simulations, there seem to exist no instances in which Axelrod's simulations have successfully been validated empirically. But without empirical validation these simulations can demonstrate nothing more then merely logical possibilities.[1]

The question would of course not be interesting if the deficiencies of these historical models weren't an issue any more at all. However, in many areas of the computational social sciences and life sciences they still are an issue. For the field of agent based modeling this has been highlighted in an article by Heath, Hill and Ciarello (2009), who criticize agent-based models for their lack of empirical validation:

> It could be argued that validation is one of the most important aspects of model building because it is the only means that provides some evidence that a model can be used for a particular purpose. Without validation a model cannot be said to be representative of anything real. However, 65% of the surveyed articles were not completely validated. This is a practice that is not acceptable in other sciences and should no longer be acceptable in ABM [Agent-Based-Modeling] practice and in publications associated with ABM. (Section 4.11)

While I strongly sympathize with their criticism of empirically not validated agent-based models, I do not intend to take the stance that all simulation models must be empirically validated or, what amounts to the same, that simulations that only prove logical possibilities are always useless. Rather, I treat this as one of the philosophical questions to be pursued in this context. My conclusion is that under the conditions listed above purely theoretical simulations may be insightful. However, often none of these conditions is met. I therefore agree with Heath, Hill and Ciarello in so far as I believe that there is an overabundance of purely theoretical simulation models which are indeed not useful for any "particular purpose".

---

1 There are kinds of computers simulations of which one can with some measure of justification claim that they prove real possibilities even without direct empirical validation. But this requires that the modeling assumptions are highly realistic and that all (natural) laws governing the simulated phenomena are well known as, for example, in quantum chemistry. Arguably, in the social sciences these requirements are never met.

2. Simulations of the evolution of cooperation: The mixed blessings of the reiterated Prisoner's Dilemma model

2.1. The success story of the reiterated Prisoner's Dilemma model

It is always dangerous to date the definite historical beginning of some trend or idea or fashion, but one can safely say that the use of computer simulations for the study of the evolution of cooperation really took on with the publication of Robert Axelrod's book “The Evolution of Cooperation” (1984). To be sure, the very concept of reciprocal altruism – which, without going into too much detail here, amounts more or less to the same as Axelrod's concept of cooperation – had been devised much earlier in an evolutionary biological context by Robert Trivers (1971). But Axelrod's use of the computer to explore the evolution of cooperation was quite innovative. An annotated bibliography assembled by Axelrod himself and Lisa D'Ambrosio (1994) ten years after the publication of the book lists 209 titles. The research report by Hoffmann (2000) “Twenty Years on: The Evolution of Cooperation Revisited” confirms the impression that Axelrod's famous computer simulations of the reiterated Prisoner's Dilemma continued to be a success story. However, the just mentioned research report relates conspicuously little empirical research on the topic. In fact, only one title is explicitly mentioned (Milinski 1987), although it had, by the time Hoffmann's report was written, already turned out not to be successful in confirming Axelrod's simulation model (Milinski and Parker 1997; Dugatkin 1997). We will come back to this in more detail later.

Axelrod's model virtually spawned myriads of similar simulation studies of the reiterated Prisoner's Dilemma (an incomplete list can be found in Dugatkin 1997), often only changing a few details or adding parameters to the original setup. Whole books with simulation series' of the reiterated Prisoner's Dilemma (Arnold 2008) or similar models (Schüßler 1997) had been written in that way. Why did Axelrod's simulations find so many followers in the scientific community? I believe that the reasons were a mixture between the simplicity of the simulation method and the assumed power of its research design. As to the simplicity: Nothing was easier than conducting research on the evolution of cooperation. You write a computer program, analyze the results, draw some more or less careful conclusions from them and publish the results. A single person with the necessary programming skills could write an Axelrod-style simulation within a few weeks. No comparison to the painstaking process of empirical research (as in Milinski 1987) or to the costs for conducting game-theoretical experiments. And what could be wrong with it? Hadn't Axelrod demonstrated how powerful the method is? At least it seemed so. But in fact, there exist considerable flaws in Axel-

rod's research design.

2.2. The deficiencies of Axelrod's research design

The research design of Axelrod's "Evolution of Cooperation" consists of three basic steps.[2] First, take a game theoretical model and run computer simulations on this model. Then, analyze the results carefully and try to explain why happened what happened. Doing so, general conclusions can be drawn from the simulation results. These general conclusions can potentially be used to explain phenomena that fall within the scope of the model. Proceeding in this fashion Axelrod came to the following conclusions about what factors encourage cooperation:

- The evolution of cooperation depends crucially on repeated interaction. Only the "shadow of the future", i.e., each player's expectations about the future behavior of other players, encourages players to cooperate.
- Successful strategies have the following characteristics (Axelrod 1984, chapter 6):
    1. they are friendly, i.e. they start with cooperative moves
    2. they are envy-free
    3. they punish defectors
    4. they are forgiving (after punishment is finished)
- TIT FOR TAT has all of these characteristics and is an extremely good strategy in the reiterated two persons prisoner's dilemma.

Unfortunately, for each of these recommendations counterexamples can be constructed by only slightly changing the parameters of the reiterated Prisoner's Dilemma or the initial strategy set or the setup of the game. For example, in contrast to the second characteristic mentioned above, Bernd Lahno has constructed a strategy ENVY that makes sure to defect at least as often as the opponent. ENVY is usually about as good as TIT FOR TAT, but more robust against random disturbances (Lahno 2000).[3] Or, to take another example, being forgiving is not always advisable: In order to avoid the problem that the initial strategy set of the simulation is more or less arbitrary, one can set a complexity limit and run the simulation with all strategies that fall within the given complexity limit. One way of doing so is to use finite automata as strategies. Finite automata have a fixed finite number of states they can be in. The state determines their action (i.e. *cooperate* or *defect* in the Reiterated Prisoner's Dilemma) in the current round. Whether and to which state an automaton changes in the next round depends on the present state of the automaton in connection with the opponent's action. If one limits the number of possible states to two, there are only 26 different automata, which makes it easy to simulate. In a setting where all possible 26 two-state automata form

2 For a description of the reiterated Prisoner's Dilemma model that Axelrod used, see the appendix.

3 For a software implementation of Axelrod's model that also contains the strategy ENVY by Bernd Lahno see Arnold (2004).

the initial strategy set, the strategy GRIM emerges as the evolutionary winner (Binmore 1994, 315; Arnold 2008, 216-218). But GRIM is the most unforgiving strategy possible, because it never stops punishing if the opponent has dared to defect but once. The example of the two-state automata also demonstrates that TIT FOR TAT is not always a great strategy. (And it is not clear what exactly it would mean to say that it is so most of the time.) It is not even generally true that indefinite repetition of interaction is necessary to encourage cooperation. Rudolf Schüßler (1997, 61ff.; see also Arnold 2008, 285-289 and the appendix of this paper) has constructed a simulation where players can break up the interaction at will. The reason why this does not generally encourage a hit and run-tactic is that those that employ hit and run must find their partner in a pool of strategies that is mostly made up of other hit and run-players, because the honest cooperators tend to keep up their relationships. Thus, enforced continuation of interaction is not a requirement, though, in a slightly different sense, the shadow of the future still is. But this is nearly the only general conclusion of Axelrod that remains intact.

Even worse for Axelrod, Ken Binmore (1998, 293ff.) has pointed out that from the viewpoint of mathematical game theory Axelrod's result yielding TIT FOR TAT as the winner strategy is both a trivial and at the same time highly contingent consequence of the well known folk theorem, which states that any payoff within the positive payoff region (in the reiterated Prisoner's Dilemma the range between the mutual defection and mutual cooperation payoff) is an equilibrium. (Just imagine a pool of strategies that punishes “grimly”, i.e. forever, all deviations from the path leading to the equilibrium payoff.)

Now, if the simulation results of the reiterated Prisoner's Dilemma model and the conclusions that can be drawn from them are contingent on very specific conditions of the simulation setup like parameter values, initial strategy set, noise etc., then this means that before any empirical phenomena can be explained with the help of conclusions drawn from the observation of simulations of the reiterated Prisoner's Dilemma, one better ought to make sure that the setup of the simulation really matches the empirical situation. This, unfortunately, seems to be close to impossible for the reiterated Prisoner's Dilemma model.

2.3. The continuing lack of empirical confirmation of the model

I have mentioned earlier that the research report by Hoffmann on the “Evolution of Cooperation” merely quotes one example of empirical research, which is Milinski's (1987) study of predator inspection by sticklebacks. This study has become a kind of running gag of the empirical confirmation of the reiterated Prisoner's Dilemma model of the evolution of cooperation. Whenever an example for empirical research in the vicinity of Axelrod's model is needed, this study is quoted. For

example, it reappears in Osborne's "Introduction to Game Theory" (2003, 445). Unfortunately the study is not a good example for the explanatory power of Axelrod's model.

It all started so well, though. Predator inspection is a kind of behavior exposed by various species of shoal fishes. When a predator comes close to the shoal it can often be observed that one or several of the shoal fishes moves towards the predator, though carefully avoiding to come to close. Usually, the distance up to which a pair of shoal fishes approaches the predator is short than that for a single fish. Milinski examined the hypotheses that if a pair of fishes approaches the predator they play a reiterated Prisoner's Dilemma. He did so by simulating partner fishes with a mirror and he found that if the simulated fish stays behind, the real fish will stop to advance as well. Because of this as well as some other reasons Milinski (1987) drew the conclusion that there is indeed evidence that the inspecting fishes play a Reiterated Prisoner's Dilemma and that they employ TITt-FOR-TAT or a similar strategy. However, the ensuing scientific debate (see Dugatkin 1997) called this result into question, because a fish might continue to advance towards the predator if the partner keeps up simply because the risk of being eaten is lower than when advancing alone and not as a reward for cooperation. After ten years of debate among experts, Milinski and Parker (1997) come to the conclusion that the empirical data does not suffice to decide whether the exposed behavior is indeed cooperative (as Milinski's earlier study assumed and tried to explain with Axelrod's model) or not. But this also means that we cannot be sure that Axelrod's model provides an adequate description at all. In fact, Milinski and Parker (1997) do not make any use of Axelrod's model any more.

Generally, the problem with Axelrod's model is that it is not very robust[4] and would require exact measurements of the payoff parameters before it can be applied empirically. Now, the big problem is: How can we measure the payoff of some kind of altruistic behavior in the animal kingdom in terms of reproductive success? It is probably beneficial for ape's to groom each others backs to get rid of the lice. But how can we measure quantitatively the increase in reproductive success an ape with a well-groomed back enjoys? The problem has so far not been solved. Unsurprisingly, Dugatkin finds hardly any examples in his survey on "Cooperation among Animals" (1997) where such quantitative measurements of payoff parameters have been attempted, let alone been successful. The situation for Axelrod-style modeling in biology twenty years after Axelrod's book had been published is nicely summarized by the biologist Peter Hammerstein:

> "Why is there such a discrepancy between theory and facts? A look at the best known examples of reciprocity shows that simple models of repeated games do not properly reflect the natural circumstances under which evolution takes place. Most repeated animal interactions do not even correspond to repeated games.
> ...

4 See the appendix, where this problem of Axelrod's model is described in more detail.

> Most certainly, if we invested the same amount of energy in the resolution of all problems raised in this discourse, as we do in publishing of toy models with limited applicability, we would be further along in our understanding of cooperation.“ (Hammerstein, 2003, 83, 92)

So much for the animal kingdom. What about the social sciences, though? Again, it is hard to find empirical studies that make more than merely inspirational use of Axelrod's model. One very striking example, however, is Axelrod's reinterpretation of Tony Ashworth's historical study on the Live and Let-Live that emerged and was sustained between enemy soldiers during the First World War on some stretches of the Western front line (Axelrod, 1984, chapter 4; Ashworth 1980). Before any model such as Axelrod's can be applied to a specific historical situation a lot of factual knowledge about this situation is required. This is what Ashworth's historical study has to provide. But Ashworth does of course not confine himself to relating bare factual knowledge. He also gives an explanation for the Live and Let-Live System where soldiers avoided shooting to kill (as they had been ordered to do) in the hope that the enemy on the other side would behave likewise. His explanation identifies an intricate set of causes encompassing the desire to survive the war, empathy with the soldiers on the other side of the front line, esprit de corps among the soldiers, similar routines, e.g. similar breakfast times when no one would shoot anyway, initial causes such as Christmas truces and bad whether periods, and, most importantly, whether the troops where elite troops or not.

How can Axelrod's model account for this intricate set of causes? The answer is: It cannot account for most of these causes other than by hiding them in its payoff parameters. But while the existence of the causes that Ashworth mentions can be concluded from the historical sources that Ashworth quotes, the payoff parameters can only vaguely be guessed by plausible assumptions. Axelrod argues with some good reason that the situation the soldiers in the trenches were in is indeed a repeated Prisoner's Dilemma situation (but see Schüßler 1997, 33ff. or Batterman et al. 1998, 89, footnote 19). However, there is no way of determining the parameters of the payoff matrix precisely. Moreover, the outcome in reality does not match the conclusions drawn by Axelrod from his simulations very well. For in his simulations cooperation prevails. But in the trenches the Live and Let-Live system only prevailed in one third of all cases (Ashworth 1980, 171ff.).

Even worse, Ashworth found out that the most crucial factor for Live and Let-Live was the elite status of the troops. On front sections where elite troops were deployed, Live and Let-Live would not take place. This fact can be brought in accordance with the model by assuming that elite troops have different payoff parameters that express valuing soldier's valor higher than life. But then we are merely making assumptions in order to make our model work. For we might as well guess without a model that elite soldiers won't shirk their orders as we may guess that the elite soldier's payoff parameters are such that our model does not predict cooperation any more.

So why use the model at all, if it does not add any epistemic surplus to the standing historical explanation of the phenomenon? Some might argue it is about the generalization that the model allows. However, if we loose the grip on the actual causes – as we do when they are hidden in the payoff parameters of some game theoretical model – then generalization is bought at the price of such a loss of information and explanatory power that it is hardly worthwhile any more (Arnold, 2008, 174ff.).

A possible objection to this criticism of Axelrod with respect to two instances of empirical studies on particular instances concerning reciprocal altruism, might be that this criticism confuses generalist with particularist explanations (Batterman et al. 1998, 84ff.). In evolutionary theory generalist explanations describe the evolutionary forces through which a certain phenotype like, for example, certain cooperative or non-cooperative behavioral traits are brought about. These evolutionary forces are also often called the *ultimate causes* of the trait in question (Mayr 1961). In contrast to that particularist explanations concern the particular mechanisms, or *proximate causes,* through which the trait is realized. For example, the proximate cause of the stotting behavior (i.e. a very ostentatious way of jumping into the air) of the Thomson Gazelle is that the Thomson Gazelle wants to signal to predators that it is healthy and not worth pursuing (Dugatkin 1997, 94f.). The ultimate cause however is that the behavioral trait is evolutionarily advantageous because it reduces the gazelle's risk of being chased and eaten by the thus informed predator. Regarding particularist and generalist explanations (or proximate and ultimate causes for that matter), it is important to understand that both types of explanation work hand in hand but operate on different levels of causation. This means both types of explanation must be compatible with each other if they concern the same phenomenon, but at the same time a generalist explanation cannot be blamed for not yielding the proximate causes of a phenomenon and *vice versa*. However, to be explanations at all, both types of explanations must be empirically testable. Thus, the distinction between generalist explanations and proximate explanations is not to be confused the with the distinction between purely theoretical models and empirically validated models, even though generalist explanations just like purely theoretical explanations tend to be more abstract than their respective counterpart.

In the case of the Live and Let Live in the First World War it would therefore not be fair criticism that Axelrod's explanation does not capture the causes that Ashworth's historical narrative describes, if Axelrod's explanation was about the ultimate causes and Ashworth explanation about the proximate causes. However, what is at stake in this example as well as the example of the predator inspection behavior is not ultimate or proximate causes but whether the model of the suggested ultimate cause (evolutionary altruism as described by the reiterated Prisoner's Dilemma model) can be identified empirically. In both examples this remains highly doubtful.

Thus, Axelrod's research design does not only fail when drawing general conclusions from specific simulation results, but also when trying to apply these conclusions to empirical subject matters. Others who have more or less followed Axelrod's research design (Schüßler 1997; Skyrms 1996; Skyrms 2004; Arnold 2008) have most of the time been more careful about drawing general conclusions from their simulation results. However, hesitating to draw conclusions they end up with the opposite embarrassment, namely, explaining what their simulations are good for, if no tenable empirical conclusions can be drawn from them. This question will concern us now.

3. What do models prove, if they merely prove logical possibilities?

We have seen that the research design of Axelrod's "Evolution of Cooperation" is flawed, because it is not possible to draw general empirical conclusions from theoretical simulations. If the simulation model is highly stylized – as Axelrod's model is – it is often not possible to relate the model to concrete empirical problems other than in a purely metaphorical form. So, what if anything can highly stylized computer simulation models demonstrate? The answer is that they can demonstrate logical possibilities, or – as one could also say – theoretical possibilities.Thus, Axelrod's simulations demonstrate that cooperative strategies can evolve under the competitive conditions of the reiterated Prisoner's Dilemma. What they do not demonstrate (without further empirical research) is that any of the instances of cooperative behavior that we find in nature is indeed the result of the mechanism that produces cooperation in the model. Moreover, a computer model as such cannot even prove that the mechanism at work is possible in nature. For what is possible in a model may still be impossible in nature. There may be natural laws that forbid to happen in nature what is possible in the computer; the conditions required to make the mechanism of the model work may nowhere be given in nature; or the same effect may, as a matter of fact, be brought about by totally different causes in nature. Thus, when speaking of simulation models as demonstrators of logical possibilities it is important to understand that these logical possibilities are not necessarily real possibilities as well.

One could also say that models demonstrate that certain concepts work or can be rendered conclusively within the sphere of theoretical imagination. Borrowing a term from software engineering one could speak of a simulation model as a "proof of concept". Thus, Axelrod's model proves the feasibility of the concept of reciprocal altruism. Again, this does not say anything about its feasibility in reality. It merely demonstrates that we can describe such a mechanism without logical contradiction.

Having understood what "proof of concept" or "proof of logical possibility" means and, more

importantly, what it does not mean, the question still remains, under what circumstances a proof of logical possibilities is valuable in the sense of providing us with important insights or adding anything of relevance to our scientific knowledge. I believe that there are (at least) three different cases where the proof of logical possibilities can indeed provide an important piece in the puzzle of scientific research:

1. *Novel Discovery.* When it reveals a phenomenon that was formerly unknown and unexpected or believed to be impossible.

2. *Best Explanation.* When the explanation of some phenomenon merely hinges on the proof that a particular mechanism can produce a given result. This can become important in the context of an inference to the best explanation.

3. *Real Possibility.* If the proven logical possibility is also a real possibility *and* if the modeled mechanism can be identified empirically.

Any one of these conditions suffices to render a theoretical model epistemically useful. I am now going to describe the three cases in more detail and one by one.

3.1. Revealing surprising phenomena

One of the best known examples for surprising simulation results are the simulations that gave rise to chaos theory. In the early 1960ies the mathematician and meteorologist Edward Lorenz, while running long series of calculations of non-linear but deterministic systems on the computer, found out that even small deviations of the initial conditions can lead to a totally different outcome (Lorenz 1963). The effect was later termed the Butterfly-effect, because, since the weather is a non-linear system, it could be possible that a butterfly's wing flap in Australia sets off a hurrican in America. If we assume that this phenomenon was unknown or, to put it more carefully, went largely unnoticed before, then its logical possibility as demonstrated by computer simulations provides a scientifically valuable insight.

Another example, that will concern us further below is Thomas Schelling's (1971) model of neighborhood segregation. Schelling's model demonstrates that the macro phenomenon that in many American cities the neighborhoods are either inhabited by blacks or whites, but seldom by a mixed population of blacks and whites, can be the result of micro motives quite different from the intention to live in a strictly segregated neighborhood. As Schelling's model shows, the effect is brought about already by a fairly mild preference not to live in an area that is strongly dominated by another ethnic group. If more than 50% of neighbors from another group are still acceptable then

this means that people would in fact be willing to live in an integrated neighborhood. Nonetheless, such a preference already suffices to produce highly segregated neighborhoods in the model. Thus, the model shows among other things that one cannot conclude from an observed neighborhood segregation pattern that the population must be racist. That Schelling's model can also be misleading if not interpreted carefully will be discussed below. Prima facie, the possibility proof of Schelling's model provides a valuable insight because it refutes an otherwise rather natural conclusion.

On the contrary, the insights that Axelrod's simulation model provides are far less surprising. That in the reiterated Prisoner's Dilemma cooperative strategies can prevail is, as has been mentioned earlier, a trivial consequence of the folk theorem, which was already known in the 1950ies (Binmore 1998, 293ff.). Axelrod's model can serve as a nice illustration for the concepts of reciprocal altruism and evolutionary stability, but both of these concepts had been detected and described before: Reciprocal altruism by Trivers (1971) and evolutionary stability by Maynard-Smith (1982). However, even if providing a good illustration of the concept reciprocal altruism is undoubtedly a merit of Axelrod's model, the same excuse cannot justify the large number of variants and follow-up simulations.

3.2. Inference to the best explanation

Another possible justification for an otherwise purely theoretical proof of concept simulation can be that it plays an important role in the context of an inference to the best explanation. A classical example may help to demonstrate under what conditions a theoretical derivation can be justified as best explanation. Gregor Mendel's (1866) experiments in plant hybridization are a very successful example of an inference to the best explanation. Mendel observed that if plants of the same kind but with different characteristics are matched, the characteristics will reappear in the offspring in certain typical numerical proportions. For example, Mendel crossed (pure) green and yellow peas and found that all descendants would be green, while in the following generation the yellow peas would reappear but be dominated in number by the green peas with a ratio of 3 to 1. Mendel found that the same holds for round and wrinkled peas. And he found that when both characteristics are combined in one experiment this yields a characteristic ratio of 9 (green and round) : 3 (green and wrinkled) : 3 (yellow and round) : 1 (yellow and wrinkled). Thus, the task was set to explain the reappearance of characteristics in the specific empirically determined numerical proportions in subsequent generations.

Mendel offered an explanation by describing a mechanism of genetic inheritance, where each characteristic (in this case form and color) is determined by a genotype that contains two alleles[5]

5 Mendel himself did not use these terms. I am using the modern terms “genotype” and “allele” here for the sake of simplicity.

that determine the characteristic. If it is assumed that one of these alleles is dominant and the other recessive and if it is furthermore assumed that in the case of the peas the allele for green dominates the allele for yellow and the allele for round dominates the allele for wrinkled then the observed ratios of characteristics in the descendant plants can be deduced from this mechanism. The mechanism and the example case is simple enough to make those deductions with pen and paper. But had Mendel lived in our time he might have used a computer simulation. This does not make a difference here, because in either case, the proof of concept consists in demonstrating that a certain mechanism is capable of producing a given result.

Now in Mendel's time molecular genetics did not yet exist, and it was not possible to determine whether the alleles were really existing entities. The mechanism he devised was largely a theoretical construction (just as so many game theoretical or agent-based computer simulations today). So, why do we accept this an explanation? The main reason is, I believe, the following: There are no real alternatives. It is hard to imagine an alternative mechanism that is equally powerful with respect to producing the observed result and at the same time equally simple and elegant as the genetic mechanism postulated by Mendel. Thus, we can say that Mendel's demonstration that a certain ratio of hereditary characteristics can be produced by the postulated mechanism constitutes a best explanation.

The crucial requirement for accepting a theory or a model or a computer simulation as a best explanation for some phenomenon is that either there are no sensible alternatives or all possible alternative explanations can be excluded for good reasons. In the case of Axelrod-style computer simulations this is usually not the case. Given a certain pattern of human or animal behavior, it is usually possible to construct many different game theoretical (or non game theoretical) models that yield this pattern, each of which could be defended on plausible grounds. Often computer simulation models contain rich sets of adjustable parameters which makes it easy to accommodate them to different desired outcomes – though at the price of the reduction of their explanatory power.

I have mentioned earlier that Milinski's (1987) attempted explanation of predator inspection with Axelrod's reiterated Prisoner's Dilemma model suffered – as the subsequent discussion revealed – from the fact that alternative explanations are still possible (see Milinski and Parker 1997). And in the case of the Live and Let-Live system in the trenches of the Western front of World War I, we have, from the historical account, knowledge about causes of this phenomenon which do not even appear in the model. Thus, notwithstanding the question whether the application of Axelrod's simulation model and its results to this empirical case study can be justified by some other story, it most certainly cannot be justified as a best explanation of the Live and Let-Live system.

Summing it up and generalizing the above considerations, simulation models that merely prove

logical possibilities can be justified in the context of an inference to the best explanation only if no alternative possible explanations exist or if those that exist can be excluded for good reasons.

As a side note, it is imaginable to explain some observed cooperative animal interaction[6] (by inference to the best explanation) as the expression of reciprocal altruism if the limited number of alternative forms of genetic altruism (kin selection based altruism and group selection based altruism) can safely be excluded. Still, it would be dangerous to conclude that any particular model of reciprocal altruism, like the reiterated Prisoner's Dilemma, provides the best explanation.

3.3. Logical possibilities, real possibilities and identifiable causes

If a proof of concept simulation can be justified as a best explanation then it is not necessary that the existence of the simulated mechanism can be proven empirically. However, this type of justification is only possible if different alternative mechanisms for the explanation of the same phenomenon can be excluded. Where this is not possible, a proof of a logical possibility might still be scientifically useful, if it can be shown that it also constitutes a real possibility. Or, in other words, it is useful when the simulated mechanism can be retraced empirically. Because simulation models that prove logical possibilities are often highly stylized, this can be quite a challenge. Milinski's (1987) attempt to identify the cooperative behavior described with Axelrod's reiterated Prisoner's Dilemma model in the predator inspection behavior of shoal fishes shows just how difficult this can be. In the vicinity of Axelrod's reiterated Prisoner's Dilemma model it is hard to find any example where this has been achieved. Arguably, this is due to its lack of robustness and its employment of input parameters (payoff utility values) that cannot be measured empirically. To illustrate the case where a proof of concept model can be justified because it models an empirically identifiable cause, I will therefore discuss the well-known neighborhood segregation model by Thomas Schelling (1971).

Schelling's (1971) model captures one of many possible causes how neighborhoods in residential areas become segregated by some group characteristic, e.g. color of skin, of their inhabitants. In its simplest form the model consists of a checkerboard landscape, each field of which is inhabited by either a green or a red agent or – as a comparatively rare case – left empty. Each agent has a lower limit with how many neighbors of the same color the agent feels happy in its neighborhood. It should be observed that if this limit is below 50% then this means that the agent would be perfectly happy with an integrated neighborhood. Agents that are unhappy move to the nearest empty field where they can be happy. The most interesting results are revealed, if the limit is set to a comparatively low level like 30%. For even though this means that the agents would be quite happy to live in

6 I do not enter into the non-trivial question how we can conclude whether some observed pattern of behavior is cooperative or egoistic.

an integrated environment or even an environment that is (up to a certain limit) dominated by the other group, they end up in segregated environments.[7]

What does this result prove? And, more importantly, can we learn something from the model about reality outside the model? As to the first of these questions, Schelling's segregation model proves the logical possibility of segregated neighborhoods to be the result of individual choices which are based on preferences that are not at all unfavorable to integrated neighborhoods. But is this logical possibility also a real possibility? In other words: Is it possible not only under the artificial conditions of the model but in the real world as well that choices of individual humans or families that do not oppose integrated neighborhoods lead to segregated neighborhoods because of their having a limit concerning how few member of the same group in the neighborhood are acceptable to them?

I believe[8] that it is legitimate to consider a logical possibility a real possibility if a) the prerequisites of the model, most notably the values of its input parameters can be retraced empirically, *and* if b) there is reason to believe that any unrealistic assumptions the model makes do not significantly affect its outcome.

Here I understand under prerequisites the initial conditions with which the model starts. These are, evidently, the values of the input parameters, but also structural conditions such as, for example, that the situation is a reiterated Prisoner's Dilemma, or that there are more or less segregated neighborhoods that consist of individual households. In a more general sense, the prerequisites of a model can be understood as those features of a model on which it depends to which (empirical) scenarios a simulation model can be applied or, what amounts to the same, whether it can be applied to any particular given scenario.

The assumptions of a simulations model concern furthermore those features that determine the course of the simulation and which can (though do not need to) describe hidden mechanisms. Schelling's model, for example, contains assumptions concerning the mechanisms by which households decide to move and by which they select the new destination. Evolutionary simulations like Axelrod's typically contain assumptions concerning the population dynamics. Mathematical simplifications and correction terms can also be counted among the assumptions found in simulation models. Roughly speaking, while it depends on the prerequisites whether a simulation model can be applied to a particular scenario, it depends on the assumptions whether its results can be trusted. The distinction between prerequisites and assumptions is not razor sharp, but it is helpful as a guidance

7 This can easily be verified with the segregation model from the NetLogo library (Wilensky 1997).

8 It would lead too far to justify the following two criteria epistemologically, here, but I am convinced that intuitively they are plausible enough.

to distinguish between those aspects of a model that must be empirically tractable before a model can be considered a realistic model and those aspects where this requirement can be lessened without interfering with its being realistic or at least its potential to deliver empirically reliable results.

It seems that in the case of Schelling's model both points can be granted. Regarding a) it does not seem impossible to inquire the preferences regarding integrated neighborhoods with a survey. Of course, the results of a survey are always imprecise to some degree. But luckily Schelling's model is quite robust in this respect.[9] Also, it is worth mentioning that the preferences to be inquired are not opaque utility values (as in Axelrod's "Evolution of Cooperation" model), but concern the least percentage of families of the same ethnic group that an individual requires in his or her immediate neighborhood for not wanting to move out.

Concerning b) it is important that Schelling's model is also robust, i.e. its results remain essentially the same, if the basic setup of the model is changed, for example, by using a different geometry than the checkerboard (Adinoyat 2007, 441). But if this is the case then there is hope that the simplifying assumptions of the model do not affect its power to capture reality. Behind this reasoning lies the idea of derivational robustness analysis (Kourikoski and Lethinen 2009) which can roughly be described as follows: If unrealistic assumptions of the model are changed in various ways, but the results remain essentially the same, then it (hopefully) will not make a difference if the unrealistic assumptions were substituted by the real conditions. Clearly, this is a case of non-demonstrative inference. But if this is granted then we can assume the model to be realistic enough for its purpose if it withstands derivational robustness analysis.

If this line of reasoning is accepted then we are entitled to assume that the logical possibility that Schelling's model proves is also a real possibility. This means, for example, that policy makers that want to set up a program to develop and sustain integrated neighborhoods would need to take this possibility into account. It is, however, very important to understand that even a real possibility is not automatically an actual cause that explains existing or emerging patterns of neighborhood segregation. Schelling's model does not require group members to be racist to bring about neighborhood segregation, but in an empirical case of neighborhood segregation racism could still be the cause.

Generally speaking, to make sure that an identified possibility is an actual cause, it is necessary to check it empirically against other possible causes, some of which might be much stronger and, therefore, preempt the effect of this particular possible cause. Also, it needs to be checked whether

9 This can be verified by varying the similar-wanted parameter in NetLogo's segregation model (Wilensky 1997).

there are no other factors that block the possible cause described by Schelling's model. Possible alternative causes include housing prices in connection with differences in average income levels between different groups. It has also been suggested that it is not so much the actual number of neighbors with a different color of skin that triggers the decision to move away, but the increase of the number of neighbors with a different color (Ellen 2000, 124-125). Finally, it might be the case that it is not a single cause but a network of several causes that bring about neighborhood segregation. In this case the epistemological situation would be similar as in the case of Axelrod's attempt to relate the reiterated Prisoner's Dilemma model to the Live and Let Live system of World War I. However, in contrast to Axelrod's model, Schelling's model captures a factor that is empirically identifiable. Thus, Schelling's model can become useful for explaining neighborhood segregation even if it captures only one of several causes that are at work. Axelrod's model – because the cause it models is not really identifiable empirically – cannot add anything relevant to the explanation of Live and Let Live.

Thus, proofs of logical possibilities also provide scientifically valuable insights if the logical possibilities are at the same time real possibilities and if it can be determined empirically if they represent actual causes or not.

4. Summary and conclusions

In this paper I have discussed what simulation models of the “Evolution of Cooperation” can possibly prove about their subject matter. I have discussed the question in the historical context of the tradition of modeling initiated by Robert Axelrod. With respect to this tradition the result is quite sobering. The research design that Axelrod's role model suggested is heavily flawed. And most of the models created in the tradition of Axelrod are not empirically applicable and therefore pretty much useless. This is not to say that a model is useless just because it is purely theoretical. But we must keep in mind that it is only under fairly restrictive conditions that a purely theoretical simulation can become useful. I have described three different sets of conditions under which this is the case. While I cannot claim that the list is complete, I am quite convinced that any complete list will still be very restrictive.

The most important conclusion that can be drawn from this discussion is that modeling is not per se a useful scientific activity. Modelers need to be aware of the contexts of explanation of empirical phenomena to which their models are meant to contribute. Otherwise there is a considerable danger that their models will not be useful at all. Most importantly they need to take into account the inevitable restrictions under which empirical research labors. For example, it is useless to construct simulation models that rely on input parameters that cannot be measured precisely enough for the

model to yield robust results.

I anticipate two objections against this conclusion: 1. Some simulation models only aim at deepening the theoretical understanding of some phenomenon and are not constructed with the intention of contributing to any empirical explanation in the first place. 2. The problem I describe is mostly history and – if at all – concerns only specific schools of modeling, most notably Axelrod-style models of the reiterated Prisoner's Dilemma models.

As to the first objection: The technical and intellectual level of social simulations is certainly not high enough to justify social simulations as a purely theoretical endeavor or *l'art pour l'art*. Also, one cannot seriously claim to have achieved a deepening of theoretical understanding if that deepening of theoretical understanding does not pay off in terms of empirical explanatory success sooner or later. Therefore, if the simulations cannot be validated empirically themselves, the least that is to be asked for is that they sooner or later become useful in the context of empirical science, *somehow*. Detached from their empirical object, social simulations are merely more or less trivial programming exercises.

As to the second objection. In this paper I have indeed only covered historical cases, but there is evidence that similar problems still exit. In the beginning of the paper I have quoted Heath, Hill and Ciarello (2009) who complain about the lack of empirical validation of many agent-based simulations. Regarding the research on the evolution of cooperation in particular, a more recent survey by Guala (2012) on the topic of weak or strong reciprocity allows to draw the conclusion that the problem has shifted. While the contemporary modeling on weak or strong reciprocity apparently links fairly well to the experimental empirical research, there is still a considerable gap between models, simulations and lab experiments on the one hand side and field research on the other hand side. Often it seems that field research is not even seriously taken into consideration as the following quotation from Guala (2012, 8) illustrates:

> According to Bowles and Gintis (2002, p. 128), for example, “studies of contemporary hunter-gatherers and other evidence suggest that altruistic punishment may have been common in mobile foraging bands during the first 100,000 years or so of the existence of modern humans.” In support of this claim, however, they cite a study (Boehm 1999) that does not endorse a costly punishment account of human sociality. Richerson and Boyd (2005, p. 219) write that “in small-scale societies, considerable ethnographic evidence suggests that moral norms are enforced by punishment.” Among their references, however, one finds only two ethnographic surveys, a laboratory experiment, and a study of dominance that do *not* support the costly punishment story (cf. Richerson & Boyd 2005, p. 280, n. 60).
>
> Most of Richerson and Boyd’s (2005) case is, in fact, based on Fehr and Gächter’s (2000a; 2002) experiments.

One can get the impression that the attitude of disregard for empirical research that has been symptomatic for many simulation studies of the reiterated Prisoner's Dilemma is carried onward in a different form in the contemporary model-oriented research on the evolution of cooperation. Thus it is still important to be aware of the limitations of pure simulation studies. And it continues to be a challenge to find simulation research designs that allow to link simulation studies with empirical field research.

5. Appendix: Examples for the weaknesses of Axelrod-inspired research designs

5.1. A Brief introduction to the repeated Prisoner's Dilemma model

The model that Axelrod investigated in his “Evolution of Cooperation” is the reiterated Prisoner's Dilemma Model in a tournament and an evolutionary (or, more precisely, population dynamical) setting. The reiterated Prisoner's Dilemma – as its name suggests – consists of repeated rounds of the one shot Prisoner's Dilemma. In the *one shot Prisoner's Dilemma* two players each have the choice to either *cooperate* with the other player or to *defect*. There are four different payoffs depending on the four possible combinations of choices the players make. If both players cooperate both get a *reward* payoff which is represented by the parameter R (default value 3); if both players defect they both get a *punishment* payoff (parameter P with default value 1); if one player defects while the other cooperates, the player who defects gets a *temptation* payoff, which is the highest possible payoff in the game (parameter T with default value 5), while the player who cooperates receives the lowest possible payoff called the *sucker's payoff* (parameter S with default value 0). For the game to be a Prisoner's Dilemma the inequality $T > R > P > S$ must hold. In the reiterated Prisoner's Dilemma also $2R > T+S$ must hold. The dilemma consists in the fact that it would be best for both players to cooperate, but they both have an incentive to defect, because, no matter what choice the other player makes, defecting yields a higher payoff than cooperating. As a consequence, both players – if they behave in a strictly rational fashion – end up with the punishment payoff instead of the reward payoff. (In technical terms: The Nash Equlibrium of the game is not Pareto-optimal.)

In the *reiterated Prisoner's Dilemma* several rounds of the game are played. (Axelrod used 200 repetitions. It is important that the exact number is unknown to the players to prevent endgame effects.) Since players can adjust their choice in the subsequent rounds to the previous choices of the other player, they get the opportunity to either punish or reward the other player. This opens up a host of strategic opportunities. The simplest strategies are either always cooperate (called DOVE) or always defect (HAWK). Other reasonable strategies are TIT FOR TAT that chooses cooperation in the first round and cooperates in the subsequent rounds whenever the other player has cooperated in

the previous round, but defects when the other player has defected, or the strategy PAVLOV (win stay, loose shift).

This game can be played by human agents in economic experiments. While it can be difficult to find out what the exact strategies are that humans use in an experiment, there is a large body of experimental evidence that shows that humans frequently cooperate in reiterated Prisoner's Dilemma games as well as other, related games like the Dictator Game or various Public Goods games (see, for example, Andreoni and Miller (1993), Gintis (2000) or Clark and Sefton (2001)). In fact the level of cooperation was so high that it was a challenge for economists to find explanations that would reconcile these results with economic theory (for example, Dufwenberg and Kirchsteiger (2004)). In a computer simulation like that of Axelrod, however, the players are computer agents the choices of which are determined by algorithms that represent their strategies.

In the *tournament setting* every strategy from a set of strategies plays the pairwise reiterated Prisoner's Dilemma with every other strategy. The goal is not to beat as many opponents as possible (in this case HAWK would be the best strategy, because it cannot be beaten in a pairwise match) but to gain the highest average score over all matches.

In the *evolutionary setting* each strategy gets a share of a population of players which breeds offspring according to the success of the strategy in a tournament of all players. Thus, if a strategy is successful its share of the population increases. A sequence of tournaments is played where each tournament resembles a new generation. The result of an evolutionary simulation of the reiterated Prisoner's Dilemma is shown on figure 1.

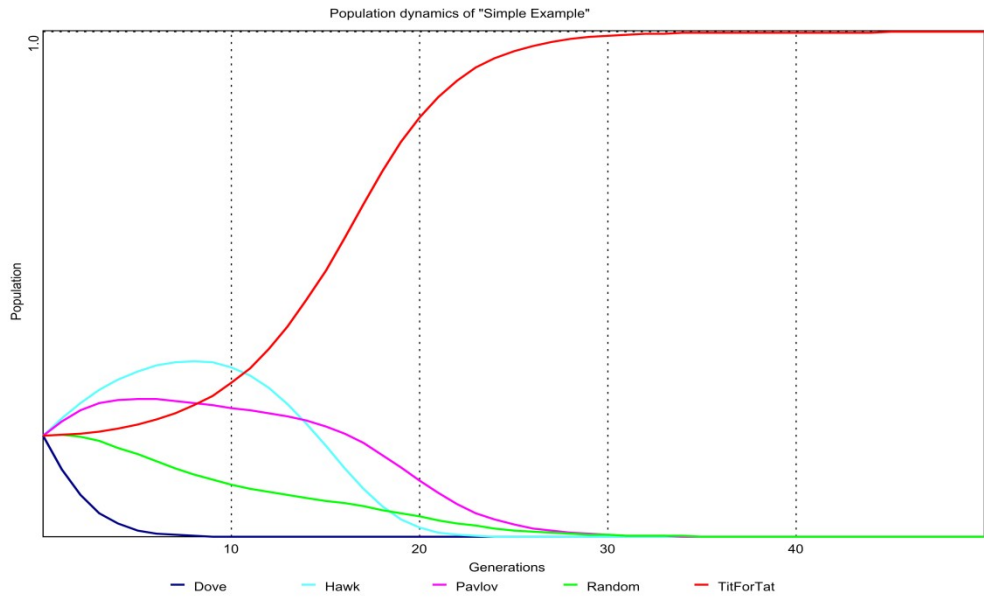

Figure 1: Evolutionary Simulation of the reiterated Prisoner's Dilemma

It should be observed that the strategy TIT FOR TAT that ends up dominating the population is superseded by HAWK and PAVLOV during the first generational cycles. The reason is that HAWK and PAVLOV both exploit DOVE, which TIT FOR TAT doesn't. Once DOVE drops out of the population they do not enjoy this advantage over TIT FOR TAT any more.

5.2. The lack of robustness of Axelrod's model

One of the main reasons why Axelrod's simulation model of the pairwise reiterated Prisoner's Dilemma fails to explain empirical instances of cooperative behavior is its lack of robustness. Without entering into the details of robustness analysis (see Kourikoski, Lethinen, Marchionni 2010 for a thorough treatment) I will consider only two aspects of robustness here: 1. In order to be robust with respect to empirical validation, the input parameters must be measurable and the model must yield stable results within the measurement inaccuracies of the input parameters. 2. Moderate changes of the setup of the model should yield similar results. Examining Axelrod's repeated Prisoner's Dilemma model under these aspects with the simulation software published by Arnold (2004), we find that:

1. Axelrod's model is not robust with respect to changes of the input parameters.

If a population dynamical simulation of the Strategies DOVE, GRIM, HAWK, JOSS, RANDOM, TAT FOR TIT, TIT FOR TAT is run with the payoff parameters T,R,P,S = 5,3,1,0 then TIT FOR TAT leads the evolutionary race after 50 generations with a population share of 38 %, but if the reward parameter is only slightly changed from R = 3 to R = 3.5, then DOVE emerges as winner with a population share of 38 %. The consequence of this lack of robustness is that before one can conclude from the model that it is better to retaliate for defections (TIT FOR TAT) than not to retaliate (DOVE) one would need to be able to measure the reward parameter R with sufficient accuracy to distinguish R = 3 from R = 3.5 in the actual application case of the model.

2. The results of Axelrod's model strongly depend on the strategy set.

If the strategies DOVE, GRIM, HAWK, JOSS, PAVLOV, RANDOM, TAT FOR TIT, TESTER, TIT FOR TAT, TRANQUILIZER[10] play a tournament, then the population dynamics yields TIT FOR TAT as the winner with TESTER[11] placing second. If then the single strategy DOWNING 0.9[12]

10 For a description of theses strategies see Axelrod (1984) or, alternatively, the source code of CoopSim (http://www.eckhartarnold.de/apppages/coopsim.html).

11 TESTER defects in the first two rounds. If the opponent reacts with punishment, TESTER plays a cooperative move (as consolation, if you like) and then switches to playing TIT-FOR-TAT during the rest of the game. Otherwise, TESTER defects every second round for the rest of the game.

12 DOWNING 0.9 tries to estimate the probability of being punished. It cooperates if this probability is greater than a certain value (0.9 by default), otherwise it deceives. (Note: This strategy is somewhat different from the strategy DOWNING described in Axelrod's "Evolution of Cooperation" (1984)).

is added, the image changes completely with HAWK coming out first and GRIM on the second place. Again, for any particular application case we would need to be able to determine the set of strategies that are in the race. It should be noted in this context that the inference from observed behavioral patterns to strategies can be ambiguous.

Because the repeated Prisoner's Dilemma model lacks the robustness that would make it applicable in an empirical context, I hold the opinion that it is not more than a good metaphor for reciprocal altruism in nature or society. One can, of course, also put it positively and say that the model provides a metaphorical explanation of reciprocal altruism – no more, no less.

5.3. Schüßler's model of cooperation on anonymous markets

One of the more interesting follow-ups to Axelrod's model has been developed by Rudolf Schüßler (1997). Schüßler's approach resembles Axelrod's model and research design in so far as he uses a highly stylized game theoretical model and draws conclusions with respect to a much more general debate.

The most important deviation from Axelrod's model is that Schüßler's model allows the players to exit the reiterated game, thus encouraging a hit- and run technique. A stripped down variant of Schüßler's simulation model[13] works as follows: There are just two strategies in the game: HAWK and QUIT FOR TAT. HAWK always defects and quits the interaction when the other player starts retaliating. QUIT FOR TAT cooperates but quits the interaction after the first defection of the other player. Interactions can also be ended by a random chance event. If an interaction sequence has ended, the players need to seek a partner from the pool of free players. After a certain number of rounds a new generation starts and each strategy's share of the population is updated according to its average payoff.

The result (figure 2) is that even though HAWK can employ a hit and run tactic (that has diligently been made impossible in Axelrod's original model) the cooperative QUIT FOR TAT strategy is more successful and soon dominates the population at a stable equilibrium with a high number of QUIT FOR TAT players and a low number of HAWK players.

13 The source code can be downloaded from:
http://www.eckhartarnold.de/apppages/downloads/RPD_with_exit_option.zip

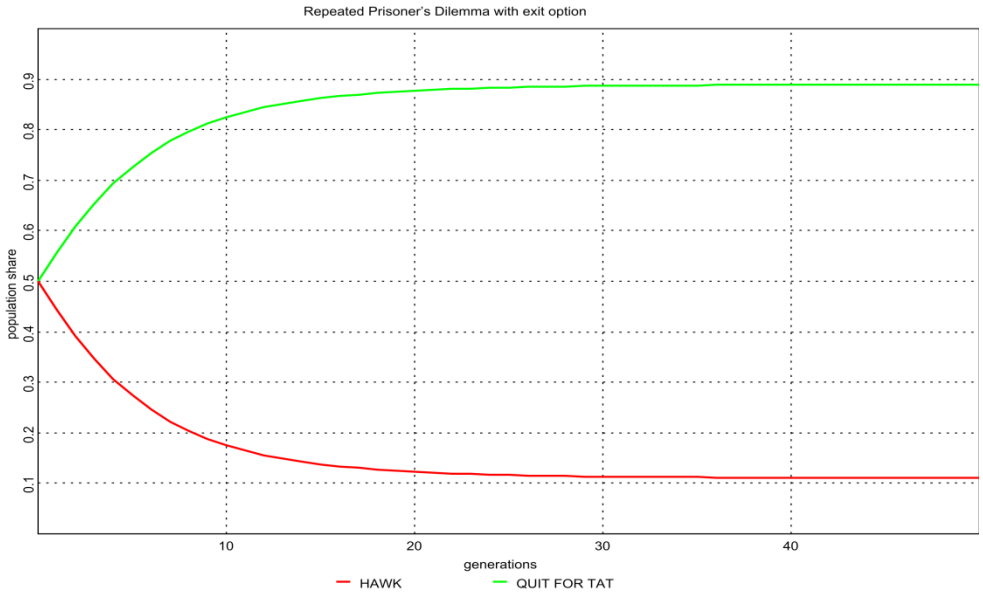

Figure 2: A simplified version of Schüßler's (1997) simulation that allows players to end the interaction at will

The explanation for this phenomenon is simple: Since the cooperative QUIT FOR TAT players tend to keep up the interaction with other cooperative players, the pool of free players will mostly consist of non-cooperative HAWK players. This effectively ruins the bargain for HAWK players, because most of the time they are on the run and have to pick a partner from the mostly non-cooperative lot they find in the pool.

Generalizing this finding, Schüßler relates his model to one of the "central, classical assumptions of normativistic sociology" which implies "that in an exchange society of rational egoists no stable cooperation-relations can emerge" (Schüßler 1997, 91). Schüssler ascribes this assumption to the sociologists Émile Durkheim and Talcott Parsons. Without quoting any sources Schüßler furthermore maintains that "alleged proofs for this thesis attempt to show that already simple analytical reflections suffice for this conclusion." And he believes his simulation "to be fit to shake this evidence".[14] But even if Parsons or Durkheim did indeed believe that "simple analytical reflections" suffice to back this assumption (rather than relying on their own empirical work to back it), it is most likely that they would embed their argument in a theoretical framework where the "inability to emerge stable cooperation-relations" is not proven of rational egoists in the sense of rational choice theory or of agents in the highly stylized environment of a computer simulation but of humans with their distinct psychological signature. A normativistic sociologist does not at all need to deny that the emergence of stable cooperation-relations is possible among the artificial agents in a highly styl-

14 This is my translation. The German original reads: "Eine der zentralen, klassischen Annahmen der normativistischen Soziologie besagt, daß in einer Austauschgesellschaft rationaler Egoisten keine stabilen Kooperationsverhältnisse entstehen können (vgl. Durkheim 1997, Parsons 1949). Angebliche Nachweise für diese These versuchen zu zeigen, daß bereits einfache, analystische Überlegungen zu diesem Schluß ausreichen. Die vorliegende Simulation sollte geeignet sein, diese Sicherheit zu erschüttern." (Schüßler 1997, 91).

ized simulation model in order to still maintain that it is *impossible* in the real world given the human nature as we know it and are acquainted with it. Thus, far from being "fit to shake this evidence" Schüßler's simulation does in fact not allow him to make any substantial contribution to the discussion about normativistic sociology. Incidentally, the example highlights the difference between logical and real possibilities and it shows how notoriously weak proofs of logical possibilities usually are in the social sciences.

5.4. Skyrm's take on Game Theory and Social Contract Philosophy

Brian Skyrms (1996; 2004) has written two books that present game theoretical models and vaguely relate them to social contract philosophy. What I am concerned with here is not his models that may have their merits on their own right,[15] but the background story that he sells them with. In "The Stag Hunt and the Evolution of Social Structure" (2004) he writes:

> How do we get from the hunt hare equilibrium to the stag hunt equilibrium? We could approach the problem in two different ways. We could follow Hobbes in asking the question in terms of rational self-interest. Or we could follow Hume by asking the question in a dynamic setting. We can ask these questions using modern tools – which are more than Hobbes and Hume had available, but still less than we need for fully adequate answers. (Skyrms 2004, 10)

Skyrms uses game theoretical simulation models to explore these questions. And he suggests that these "modern tools" allow us to gain insights about the topics that Hobbes and Hume explored beyond that what Hobbes and Hume can offer us ("more than Hobbes and Hume had available").[16] Almost all of the book is simply dedicated to the discussion of various highly stylized simulation models. One of the models Skyrm's presents is a highly stylized spatial model of the stag hunt game. I am not going to discuss this model in detail. Only so much: Depending on the setup and the particular boundary conditions the cooperative stag hunting equilibrium spreads and eventually replaces the non cooperative hare hunting equilibrium completely. In a concluding postscript Skyrms returns to the topic of classical social contract philosophy:

> How much progress have we made in addressing the fundamental question of the social contract: "How can you get from the non cooperative hare hunting equilibrium to the cooperative stag hunting equilibrium?" The outlines of a general answer have begun to

15 For the discussion of these see Ernst (2001) and Batterman et al. (1998). Both of these papers are, of course, only concerned with Skyrms first book (1996) only. While both papers take critical notice of the lack of robustness of Skyrms' simulations, they both give it a more gentle reading than I do here.

16 This echos a belief in the superiority for formal methods that is widespread in contemporary analytical philosophy. In a similar vein Kwame Appiah writes "It would be interesting and important if we could make more precise the sort of argument Hobbes offered, so that we could say just why it is that the advantages of civil society over the state of nature ought to appeal to anyone." (Appiah 2003, 232) And he carries on with describing the Prisoner's Dilemma as game-theoretical model for the state of nature.

> emerge. Over time there is some low level of experimentation with stag hunting. Eventually a small group of stag hunters comes to interact largely or exclusively with each other. ... The small group of stag hunters prospers and can spread by reproduction and imitation. (Skyrms 2004, 123)

The first oddity of this story is that the reader has to assume that the fundamental question of the social contract is a question about game theoretical equilibriums. Historically, the classical questions of the social contract that Hobes, Locke, Hume and others discussed in the 17th and 18th century, were questions like: How can we overcome or prevent anarchy? How is political order possible? Is a peaceful order without the institution of government possible at all? How can we justify the institution of government? And similar questions. Now, it is *prima facie* far from clear that these questions can be rephrased as the game theoretical question "How can you get from the non cooperative hare hunting equilibrium to the cooperative stag hunting equilibrium?" Rather it seems that, just like in Axelrod's and Schüßler's case, we have a highly stylized computer model on the one hand side and we have concrete real life questions on the other hand side and the resemblance between both is at best metaphorical.

However, if it is assumed that the "fundamental question of the social contract" when rephrased in game theoretical terms really is the same question, then the second oddity is the fact that the answer that Skyrms arrives at contradicts completely the answer that Hobbes, Locke, Hume and about any other classical social contract philosopher arrived at. In Skyrms model it is possible that order (resembled by the cooperative hare hunting equilibrium) emerges from disorder (resembled by the non cooperative stag hunting equilibrium) even without the institution of government. But the whole point of classical social contract philosophy was to justify the existence of government (and to determine its proper form), because none of the social contract philosophers from the 17th and 18th centuries seriously believed that a peaceful order was possible without government. It is surprising that Skyrms does not comment on this contradiction. If we exclude the possibility that the thinkers of the 17th and 18th century were simply wrong (which for the extreme scarcity of historical examples of peaceful states of anarchy we can safely do), then the most benevolent reading we can give to Skyrms is that under the guise of social contract theory he discusses different questions with the help of game theoretical models that may be appropriate to address these, but which appear inadequate when related to the classical questions of social contract theory that circle around the institution of government in large scale societies.

5.5. Arnold and the story of "slip stream" altruism

I have myself conducted some Axelrod-style simulations and published the results in Arnold (2008). Do they suffer from the same limitations as the other examples? I am afraid they do and

here is why.

In Arnold (2008) I conduct large scale simulation series of Axelrod-style simulation of the repeated Prisoner's Dilemma. Then I analyze the results, both the aggregated results and some of the individual simulation results. In these I found "interesting" phenomena that I then examined thoroughly and tried to explain in the sense in which one "explains" the results of computer simulations, i.e. by relating them to the mechanisms of the simulation and trying to understand why these mechanisms bring the results about. What I refrained from doing is to draw general conclusions from the simulations, but, as has been said earlier, such modesty does not rescue the project, because it only raises the question what a simulation is good for, if no conclusions can be drawn from it.

One of these interesting phenomena is the fact that in many simulations of the series naïve cooperators gained a moderate success. In mixed strategy sets that contain reciprocators (e.g. TIT FOR TAT), naïve cooperators (e.g. DOVE) and exploiting strategies (e.g. HAWK) it can happen that the reciprocators wipe out the exploiters before the naïve cooperators have been wiped out, which allows the naïve cooperators to survive at a low level. I called this phenomenon "slip stream"-altruism (Arnold 2008, 103ff.). Its philosophical significance, if any, lies in the fact that it shows that under evolutionary conditions even genuine altruists, i.e. altruists that do good without expecting to be treated well in return, can survive.[17] Some philosophers tend to interpret reciprocal altruism as merely a kind of time-deferred egoism. (It should be observed that punishment isn't costly in this model, so the naïve cooperators do not live at the expense of the reciprocal cooperators.)

An interesting special case is that where naïve cooperators are even more successful than reciprocal cooperators. This case is in its simplest form depicted on figure 3. Here the strategies TAT FOR TAT (a variant of TIT FOR TAT that starts with a defection) and TIT FOR TAT keep out the exploiter HAWK. But since TIT FOR TAT and TAT FOR TIT do not play very well against each other, DOVE earns the highest score. I used the term *conflicting reciprocators* to describe the situation when there are different types of reciprocal altruists in the game that do not play well with each other.

But can we learn anything about reality from this finding? It is imaginable that a biologist eventually stumbles upon a constellation of behavioral types in the population of some animal species that contains exploitative, naïve and reciprocal strategies and where the naïve altruistic behavioral types enjoy the highest population share. Then, this simulation might serve as one of (undoubtedly) several possible explanatory metaphors for this behavior. It is not more than a metaphor because,

17 Another evolutionary mechanism that can produce genuine altruists is that of group selection (Sober and Wilson 1998).

again, we cannot measure the payoff parameters. The simulation depicted on figure 3 yields TIT FOR TAT instead of DOVE as the winner if the payoff parameter R is set back to its default value R = 3. Therefore, I believe that at the end of the day my simulation series is just another one of the "toy models with limited applicability" that Hammerstein (2003, 92) complains about.[18]

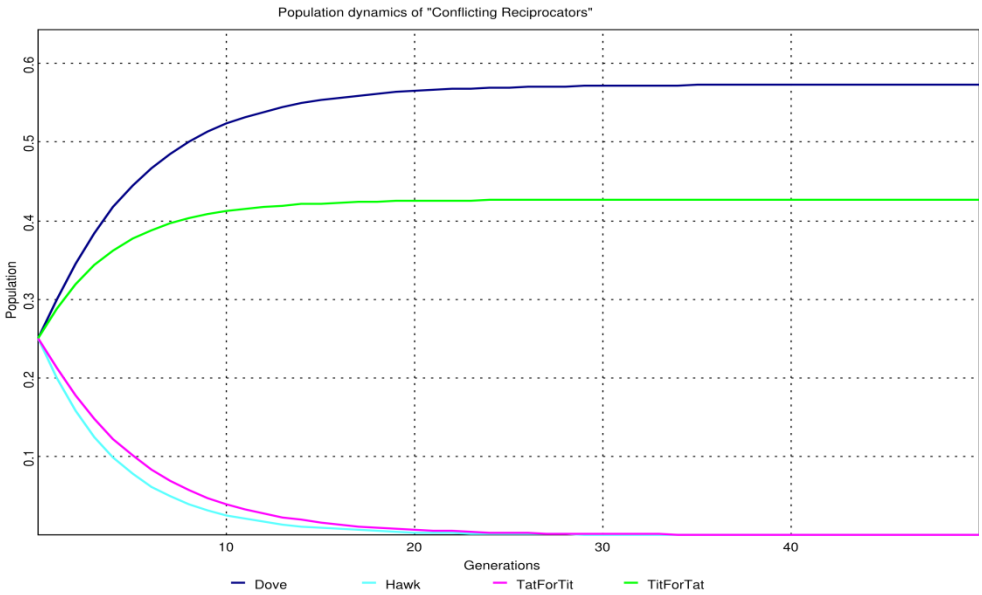

Figure 3: DOVE wins because there are conflicting reciprocal cooperators in the game. (Please note that the parameter R = 4 in this simulation!)

References

Andreoni, James and Miller, John H. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *The Economic Journal* 103 (1993): 570-585.

Arnold, Eckhart. *CoopSim.* A Python implementation of Axelrod's "Evolution of Cooperation" Simulation. 2004. http://www.eckhartarnold.de/apppages/coopsim.html.

Arnold, Eckhart. *Explaining Altruism. A Simulation-Based Approach and Its Limits*, Heusenstamm, Ontos Verlag, 2008.

Appiah, Kwame: *Thinking it Through. An Introduction to Contemporary Philosophy.* Oxford University Press, 2003.

Ashworth, Tony: *Trench Warfare 1914-1918. The Live and Let Live System*. MacMillan Press Ltd., 1980.

Axelrod, Robert. *The Evolution of Cooperation*. Basic Books, 1984.

Axelrod, Robert and D'Ambrosio, Lisa. "Annotated Bibliography on the Evolution of Cooperation". Center for the Study of Complex Systems, University of Michigan, 1994. http://www-personal.umich.edu/~axe/research/Evol_of_Coop_Bibliography.htm.

18 Some of my simulations are quoted favorably in Schurz (2012). Now, I am very grateful for that and happy that the simulations were useful to someone else. However, in view of the epistemological limitations of purely theoretical simulations such as mine, I still feel a bit uneasy about it as if with some kind of undeserved praise.

Aydinonat, N. Emrah. “Models, conjectures and exploration: an analysis of Schelling’s checkerboard model of residential segregation.” *Journal of Economic Methodology* 14,4 (2007):429–454.

Bateson, William. *Mendel's Principles of Heredity. A Defense.* Cambridge University Press 1902. http://www.esp.org/books/bateson/mendel/facsimile/title3.html (facsimile).

Battermann, Robert, D'Arms, Justin and Górny, Krzystof. “Game Theoretic Explanations and the Evolution of Justice.” Philosophy of Science 65 (1998): 76-102.

Binmore, Ken. *Game Theory and the Social Contract I. Playing Fair*. Fourth printing (2000). Cambridge, Massachusetts / London, England, MIT Press, 1994.

Binmore, Ken. *Game Theory and the Social Contract II. Just Playing*. Cambridge, Massachusetts / London, England, MIT Press, 1998.

Brian Heath, Raymond Hill, and Frank Ciarello. “A survey of agent-based modeling practices (January 1998 to July 2008).” *Journal of Artificial Societies and Social Simulation (JASSS)*, 12(4):9 (2009).
http://jasss.soc.surrey.ac.uk/12/4/9.html.

Clark, Kenneth and Sefton, Martin. “The Sequential Prisoner's Dilemma: Evidence on Reciprocation” *The Economic Journal* 111 (2001): 51-68.

Dufwenberg, Martin and Krichsteiger, Georg. “A theory of sequential reciprocity” *Games and Economic Behavior 47 (2004): 268-298.*

Dugatkin, Lee A. *Cooperation among Animals*. Oxford University Press, 1997.

Ellen, Ingrid Gould. *Sharing America’s Neighborhoods: The Prospects for Stable Racial Integration*. Harvard University Press, 2000.

Ernst, Zachary. “Explaining the Social Contract” British Journal for the Philosophy of Science 62 (2001): 1-24.

Gintis, Herbert. “Beyond *Homo economicus*: evidence froSchelling's (1971) model captures one of many possible causes how neighborhoods in residential areas become segregated by some group characteristic, e.g. color of skin, of their inhabitants.m experimental economics.” *Ecological Economics 35 (2000): 311-322.*

Guala, Francesco. “Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate.” *The Behavioral and Brain Sciences* 35, no. 1 (February 2012): 1–15. doi:10.1017/S0140525X11000069.

Hammerstein, Peter: “Why Is Reciprocity So Rare in Social Animals? A Protestant Appeal.” In *Genetic and Cultural Evolution*, edited by Peter Hammerstein. Cambridge, Massachusetts / London, England, MIT Press in cooperation with Dahlem University Press, 2003, Chap. 5: 83–94.

Hoffmann, Robert. “Twenty Years on: The Evolution of Cooperation Revisited.” *Journal of Artificial Societies and Social Simulation* 3:2 (2000). http://jasss.soc.surrey.ac.uk/3/2/forum/1.html.

Kourikoski, Jaakko and Aki Lethinen. “Incredible Worlds, Credible Results.” *Erkenntnis* 70 (2009):119–131.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. “Economic Modelling as Robustness Analysis.” *The British Journal for the Philosophy of Science* 61, no. 3 (September 1, 2010): 541–567. doi:10.1093/bjps/axp049.

Lahno, Bernd “In Defense of Moderate Envy.” *Analyse und Kritik* 22:1 (2000), 98-113.

Lorenz, Edward N. “Deterministic Nonperiodic Flow.” *Journal of the Atmospheric Sciences* 20,2 (1963): 130–141. doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Mendel, Gregor. "Versuche über Pflanzen-Hybriden." *Verhandlungen des naturforschenden Vereines, Abhandlungen,* Brünn 4 (1866): 3-47.
http://www.mendelweb.org/MWGerText.html (German).
http://www.mendelweb.org/Mendel.html (English).

Milinski, Manfred. "TIT FOR TAT in sticklebacks and the evolution of cooperation." *nature* 325 (1987): 433–435.

Milinski, Manfred and Parker, Geoffrey A. "Cooperation under predation risk: a data-based ESS analysis." *Proceedings of the Royal Society* 264 (1997): 1239–1247.

Osborne, Martin J. *An Introduction to Game Theory*. Oxford University Press, 2003.

Schelling, Thomas C. "Dynamic Models of Segregation" *The Journal of Mathematical Sociology* 1, no. 2 (1971): 143–186.
doi:10.1080/0022250X.1971.9989794.

Schurz, Gerhard. *Evolution in Natur und Kultur: Eine Einführung in die Verallgemeinerte Evolutionstheorie.* Heidelberg, Springer, 2011.

Schüßler, Rudolf. *Kooperation unter Egoisten: Vier Dilemmata.* 2nd edition, München, R. Oldenbourg Verlag, 1990.

Sober, Elliott and Wilson, David S. *Unto Others. The Evolution and Psychology of Unselfish Behaviour.* Harvard University Press, 1998.

Skyrms, Brian. *Evolution of the Social Contract*. Cambridge University Press, 1996.

Skyrms, Brian. *The Stag Hunt Game and the Evolution of Social Structure*. Cambridge University Press, 2004.

Smith, John Maynard. *Evolution and the Theory of Games*. 2nd ed. Cambridge University Press, 1982.

Trivers, Robert L. "The Evolution of Reciprocal Altruism." *The Quarterly Review of Biology* 46 (1971): 35–57.

Wilensky, U. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL., 1999.
http://ccl.northwestern.edu/netlogo/.

Wilensky, U. *NetLogo Segregation model*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL., 1997.
http://ccl.northwestern.edu/netlogo/models/Segregation