The Dark Side of the Force: When computer simulations lead us astray and "model think" narrows our imagination

— Pre conference draft, *Models and Simulations*, Paris, June 12-14 —

Eckhart Arnold

May 31th 2006

Abstract

This paper is intended as a critical examination of the question of when the use of computer simulations is beneficial to scientific explanations. This objective is pursued in two steps: First, I try to establish clear criteria that simulations must meet in order to be explanatory. Basically, a simulation has explanatory power only if it includes all (known) causally relevant factors of a given empirical configuration and if the simulation delivers stable results within the measurement inaccuracies of the input parameters. If a simulation is not explanatory, it can still be meaningful for exploratory purposes, but only under very restricted conditions.

In the second step, I examine a few examples of Axelrod-style simulations as they have been used to understand the evolution of cooperation (Axelrod, Schüßler) and the evolution of the social contract (Skyrms). These simulations do not meet the criteria for explanatory validity and it can be shown, as I believe, that they lead us astray from the scientific problems they have been addressed to solve and at the same time bar our imagination against more conventional but still better approaches.

Contents

1 Introduction

Computer simulations have become a popular tool in various branches of science, including even the social sciences. The reasons are easy to understand: Computer simulations provide a simple and yet powerful tool to explore the implications of theoretical assumptions. They are cheaper than experiments and often easier to construct and to handle than mathematical models. At the same time they confine the realm of what can be modeled only to what can be described algorithmically, which gives them a very broad scope. With this tool at hand it should be possible to bring into the reach of exact treatment even such questions that have traditionally seemed to defy the use of formal methods.

However, upon closer inspection it becomes apparent that computer simulations do not always deliver what they promise. Often they remain in the state of purely theoretical "toy simulations" and never get to the ground of empirical testability. In the following, I will first try to put forward a few

straight forward criteria for proper explanatory computer simulations. After that I will analyze some examples of computer simulations that fail to meet these criteria and I will try to point out the bad consequences this failure has.

2 Different aims of computer simulations in science

Computer simulations can be employed in science not only for generating explanations but for various different purposes. They can, for example, be used to merely express certain theoretical assumptions or concepts. In this sense they provide a sometimes weaker and sometimes stronger but usually simpler and more flexible alternative to mathematical modelling. Or they can be used to prove the "logical possibility" of certain general assumptions such as the assumption that cooperation is possible among egoists. Or they can be used to explore the possible consequences or implications of certain assumptions. All of these previously mentioned uses of computer simulations can be subsumed under the general title of *exploratory simulations* or, as it is sometimes also called, *speculative simulations*. It is the distinctive mark of this type of simulations that these simulations do not need to resemble empirical reality. If there exists any resemblance at all then it is typically vague and consists in the plausibility of the assumptions.

Another – potentially more important – class of computer simulations are *predictive simulations*. The purpose of predictive simulations is to generate true predictions of some empirical process. An example might be simulations in meteorology that predict how the wheather is going to be in the future. The assumptions that enter into predictive simulations do not need to be in any way realistic. As long as the predictions prove to be reliable, it is permissible to use strongly simplified assumptions about the modeled process or even assumptions which are known to be false. This shows that just because a simulation produces successful predictions it does not necessarily also provide an explanation for the predicted phenomena, even though successful predictions may be one among several indicators for a simulation to

be explanatory valid.

The most desired case, however, would be that of an *explanatory simulation* that is a type of computer simulation that actually allows us to explain the empirical phenomena that are modeled in the simulation. It is this class of simulations or, rather, the class of simulations that pretend to be explanatory but really are not that I will be concerned with in this paper.

3 Criteria for "explanatory" simulations

But in what sense can a computer simulation be explanatory? And what are the criteria a computer simulation must meet in order to be explanatory?

A computer simulation can be called *explanatory* if it adequately models some empirical situation and if the results of the computer simulation (the *simulation results*) coincide with the outcome of the modeled empirical process (the *empirical results*). If this is the case, we can conclude that the empirical results have been caused by the very factors (or, more precisely, by the empirical correspondents of those factors) that have brought about the simulation results in the computer simulation.

To take an example, let us say we have a game theoretic computer simulation of the repeated prisoner's dilemma where under certain specified conditions the strategy "tit for tat" emerges as the clear winner. Now, assume further that we know of an empirical situation that closely resembles the repeated prisoner's dilemma with exactly the same conditions as in our simulations. (Probably, the only way to bring this about would be by conducting a game theoretic experiment, where the conditions can be closely monitored.) And let us finally assume that also in the empirical situation the "tit for tat" strategy emerges as the most successful strategy. Then we are entitled to conclude that "tit for tat" was successful in the empirical case, because the situation was a prisoner's dilemma with such and such boundary conditions and because – as the computer simulation shows – "tit for tat" is a winning strategy in repeated prisoner's dilemma situations under the respective conditions.

Now that we have seen how explanations by computer simulations work

in principle, let us ask what are the criteria a computer simulation must fullfill in order to deserve the title of an *explanatory simulation*. The criteria should be such as to allow us to check whether the explanation is valid that is whether the coincidence of the results is due to the congruence of the operating factors (in the empirical situation and in the computer simulation) or whether it is merely accidental.

As criteria that a computer simulation must meet in order to be an explanatory model of an empirical process, I propose the following:

1. *Adequacy Requirement*: All causally relevant factors of the modeled empirical process must be represented in the computer simulation.

2. *Stability Requirement*: The input parameters of the simulation must be measurable with such accuracy that the simulation results are stable within the range of inaccuracy of measurement.

If both criteria are met, we can say that there exists a *close fit* between model and modeled reality. The claim I wish to hold is that only if there is a close fit between model and reality we are entitled to say that the model explains anything. Even though these criteria are very straight forward, a little discussion will be helpful for a better understanding.

Regarding the first criterion, it should be obvious that if not all causally relevant factors are included then any congruence of simulation results and empirical results can at best be accidental. Two objections might be raised at this point: 1) If there really is a congruence of simulation results and empirical results should that not allow us to draw the conclusion that the very factors implemented in the computer simulation are indeed all factors that are causally relevant? 2) If we use computer simulations as a research tool to find out what causes a certain empirical phenomenon, how are we to know beforehand what the causally relevant factors are, and how are we ever to find out, if drawing reverse conclusions from the compliance of the results to the relevant causes is not allowed?

As to the first objection: If the simulation is used to generate empirical predictions and if the predictions come true then this can – with some

hesitations – indeed be taken as a hint to its capturing all relevant causes of the empirical process in question. The hesitations concern the problem that even if a simulation has predictive success it can still have been based on unrealistic assumptions. Sometimes the predictive success of a simulation can even be increased by sacrificing realism. Therefore, in order to find out whether the factors incorporated in the computer simulation are the causally relevant factors we cannot rely on predictive success alone, but we have to consult other sources as well, such as our scientific background knowledge about the process in question.

As to the second objection: If we have a simulation that predicts correctly then we are – with the hesitations mentioned above – entitled to draw reverse conclusions from the compliance of the results to the exclusive causal relevance of the incorporated factors or mechanisms. However, this is impermissible if the simulation does not generate predictions but is just meant to give an ex-post explanation. For, if we only try long enough, we are almost sure to find some computer simulation and some set of input parameters that matches a previously fixed set of output data. The task of finding such a simulation amounts to nothing more than finding any arbitrary algorithm that produces a given pattern. But then we will only accidentally have hit on the true causes that were responsible for the results in the empirical process.[1]

Therefore, only if we make sure that all causally relevant factors are included in the simulation, we can take it as an explanation. And usually we cannot assure this by relying on the conformance of the simulation results and the empirical results alone without any further considerations. Summarizing we can say: *If the first criterion is not fullfilled, then the computer simulation does not explain anything.*

[1]The problem here is in some respects similar to the problem of curve fitting, where one has to deal with the danger of overfitting. One could try to apply similar tricks here as are often used with curve fitting. For example, one could try to turn an ex-post explanation into a quasi-prediction by dividing the data set (that describes the empirical results) and then designing and calibrating the simulation on only one part of the divided data set. The thus calibrated simulation is then used to "predict", or rather "quasi-predict" the other part of the data set. If the "quasi-predictions" prove to be true, we have some reason to assume that we have hit upon the real causes. But, even if we use such methods to create quasi-predictions, the above mentioned caveats apply.

The second criterion is even more straight forward. If the model is unstable, then we will not be able to check whether the simulation model is adequate. For, if it is not stable within the inevitable inaccuracies of measurement, it does not deliver one result, but a range of different results. But then we cannot say for sure, whether the empirical results are due to the factors the model captures. Imagine for example, we had a game theoretic model that tells us whether some actors will cooperate or not cooperate. Now assume, we had some empirical process at hand where we know that the actors cooperate and we would like to know whether they do so for the very reasons the model suggests. In other words: We would like to know whether our model can explain why they cooperate. If the model is unstable then – due to measurement inaccuracy – we do not know whether the empirical process falls within in the range of input parameters for which the model predicts cooperation or not. Then there is no way to tell whether the actors in the empirical process cooperated, because of the reasons the model suggests or, quite the contrary, inspite of what the model would predict.

A special case of this problem of model stability and measurement inaccuracies occurs when we can only determine the ordinal relations of greater and smaller of some empirical quantity, but not its cardinal value (perhaps, because it does not have a cardinal value by its very nature such as the quantity of utility in economics for example), even though the simulation crucially depends on the ordinal value of the respective input parameter.[2] Briefly put, the morale of the second criterion is: *If condition two is not met, we cannot know whether the computer simulation explains.*

In connection with the first criteria the requirement of model stability (in relation to measurement inaccuracy) gives rise to a kind of dilemma. An obvious way to make a model more adequate is by including further parameters. Unfortunately, the more parameters are included in the model the harder it becomes to handle. Often, though not necessarily, a model looses stability by including additional parameters. Therefore, in order to assure that the model is adequate (first criterion), we may have to lower

[2]This is a well known restriction for modelling in economics, but it seems to have fallen into oblivion when computer simulations hit the scene.

the degree of abstraction by including more and more parameters. But then the danger increases that our model will not be sufficiently stable any more (second criterion).

There exists no general strategy to avoid this dilemma. In many cases it may not be possible at all to get around the dilemma. But this should not come as a surprise. It merely reflects the fact that the use computer simulations is, of course, limited. With the tool of computer simulations many scientific problems get into the reach of formal modelling that would be hard to handle with pure mathematics alone. Still, many scientific problems remain outside the realm of what can be handled with formal methods, either because of their complexity or because of the nature of the problem. This remains especially true for many areas of the social sciences.

Apart from the two criteria listed above it is important that the output of the computer simulation should reflect the empirical results with all the details that are regarded as scientifically important, and not just – as it sometimes happens – merely a much sparser substructure of them. [3] For example, we may want to use game theoretic model like the prisoner's dilemma to study the strategic interaction of states in politics. The game theoretic model will tell us whether the states will cooperate or not, but most probably it will say nothing about the concrete form of cooperation (diplomatic contacts, trade agreements, international contracts etc.) or non cooperation (embargos, military action, war etc.). Therefore, even if the model or simulation really was predictively accurate, it does at best provide us with a partial explanation, because it does not explain all aspects of the empirical outcome that interest us. In the worst case it's explanatory – or, as the case may be, it's predictive – power is almost as poor as that of a horoscope. The prediction of a horoscope that, for example, tomorrow "something important" will happen easily becomes true, because of its vagueness. Similarly, if a game theoretic simulation predicts that the parties of a political conflict will stop cooperating at some stage, but does not tell us whether this implies, say, the outbreak of war or just the breakup of diplomatic relations then it only

[3] This requirement could also be regarded as a second adequacy criteria, but to keep things simple it has been left out from the list of criteria above.

offers us comparatively unimportant information. We could also say that if the simulation results fail to capture all important features of the empirical outcome then the computer simulation "misses the point".

Summing it up: Only if a computer simulation closely fits the simulated reality – that is if it adequately models the causal factors involved, if it is stable and if it is descriptively rich enough to "hit the point" – it can claim to be explanatory.

4 Simulations that fail to explain

In the following I will to discuss two examples of computer simulations that were designed by its authors to explain certain empirical phenomena but ultimately fail to do so. But it is not only the failure to explain that I am interested in. What concerns me more is the consequences these failures had. In the one case it lead scientists away from the relevant questions and made them indulge into the study of computer simulations that remained almost completely useless for scientific explanation. In the other case it had, as it seems, the effect of limiting the imagination so that some of the most important features of the respective subject matter got overlooked.

Admittedly, both examples are examples of bad simulations. Why bother looking at bad simulations? Because, in these cases the failures are just the more obvious and they help us understand what to avoid. Also, at least the type of simulations described in the first example has been immensely popular for a long time.

4.1 Axelrod style simulations of the "evolution of cooperation"

4.1.1 Typical features of Axelrod style simulations

My first example is concerned with the sort of computer simulations of "the evolution of cooperation" that have become very popular after the publication of Robert Axelrod's book (Axelrod, 1984) with the same title. Robert Axelrod's book is a surprising phenomenon for two reasons: First of all, be-

cause of the extraordinary success it had as far as its impact on the scientific community is concerned. It spawned virtually myriads of subsequent studies on the repeated prisoner's dilemma (the model Axelrod used) and the "evolution of cooperation" that went more or less along the same lines and employed similar methods as Axelrod. An annotated biography from 1994 (ten years after the first publication of "The Evolution of Cooperation") lists more than 200 articles that directly relate to Axelrod's study. But Axelrod's approach is also surprising for a second reason: The almost complete uselessness his and his follower's computer simulations of the reiterated prisoner's dilemma proved to have for the empirical research in the field.

How did Axelrod arrive at his results about cooperation and why did it prove so difficult to support them empirically? In order to find out, if and how cooperation can emerge among egoistic agents, Axelrod started off with a game theoretical model of a certain type of cooperation dilemma, the well known prisoner's dilemma. Since the one shot prisoner's dilemma does not offer many strategic opportunities (no rational player will ever cooperate in the one shot prisoner's dilemma, and any player who does fares worse than if he or she did not), Axelrod built his simulation on top of the repeated prisoner's dilemma. He conducted his famous computer tournaments of the repeated two player prisoner's dilemma with strategies that he had got from many different participants. On top of the computer tournament he built an "evolutionary simulation" simulating a population dynamical process among these strategies by using the payoffs they gained in the tournament to calculate their fitness values.[4] Already at this point we may notice that the setup of Axelrod's simulation does not resemble any empirical situation whatsoever. The prisoner's dilemma itself provides a concise abstract description of the essential features of many dilemma situations that occur in reality, but nowhere in this world we find an arrangement that really corresponds to Axelrod's computer tournament that is built on top of it. How are we then to draw conclusions from the computer tournament with respect to empirical

[4]The details are not important here. There exist many descriptions of Axelrod's procedure the best of which is probably still Axelrod's own book (Axelrod, 1984). Simulations of the repeated prisoner's dilemma similar to Axelrod's computer tournament can easily be found on the web. (Google for "CoopSim" to find one of them.)

cooperation dilemmas?

The way Axelrod proceeded was to examine the simulation results and to draw generalizing conclusions from them. This is how Axelrod arrived at such conclusions like: the strategy *Tit For Tat* is generally a very good strategy in the repeated prisoner's dilemma, a strategy should be friendly in the sense that it should not start to defect, a strategy should punish defection but not be too unforgiving, the evolution of cooperation depends crucially on the continuation of interaction and the like. Unfortunately, subsequent research showed that none of these conclusions was generally true. It suffices to change the simulation setup but a little bit and it pays to be a cheater, or to be unforgiving (as is the case when the simulation is run with all two state automata as a base strategy set). And, of course, *Tit For Tat* does not always win the race. The general finding that cooperative strategies can be successful in the repeated prisoner's dilemma as such is just a trivial consequence of the game theoretical folk theorem (Binmore, 1998, p. 313ff.). And all other generalizing conclusions Axelrod drew were simply not warrented.

Nonetheless, Axelrod's pioneering work triggered off a multitude of similar computer simulations of the prisoner's dilemma or other games. Most of their author's were too cautious to draw such sweeping conclusions as Axelrod did. Still, regarding their design and the kind of reasoning they rely on, many of these simulations follow the pattern that was set by Axelrod's role model. In order to classify this type of simulation, we may speak of *Axelrod style simulations*.

Generally speaking, *Axelrod style simulations* are computer simulations that share the following typical features:

1. They are constructed from a set of plausible assumptions or on top of a common mathematical model. In many cases they are derived from existing Axelrod style simulations by adding new parameters or changing other boundary conditions. The concrete shape of the model remains largely arbitrary and at the discretion of the scientist who builds it.

2. They are not related to any particular empirical situation. (And most

certainly there exists no *close fit* to empirical reality in the sense explained before.) Thus they remain a primarily theoretical endeavor.

3. If any conclusions are drawn from the simulation, they are usually drawn by means of inductive generalizations from the simulation results. The simulation is thus used to establish very general points or rules of thumbs about its subject matter.

4.1.2 How Axelrod style simulations work

Let us look in more detail on a typical exponent of this tradition of simulation based research to see how Axelrod-style simulations work in practice. An in many respects good example for this tradition is provided by Rudolf Schüßler's "Kooperation unter Egoisten" (Schüßler, 1990). Schüßler called into question Axelrod's assumption that continued interaction is a necessary precondition for the evolution of cooperation. Quite the contrary to Axelrod's thesis, Schüßler wanted to show that cooperation can even emerge on "anonymous markets". In order to do so he set up is own Axelrod-style simulation where agents are free to break up the cycles of interaction whenever they want. This encourages a kind of hit and run tactic where agents do not cooperate in the last round before stopping the interaction on their behalf and take away the benefit of single sided non cooperation without being punished. With the help of his computer simulation Schüßler could demonstrate that even in this case cooperative strategies could – under certain specific simulation conditions – outcompete the cheaters. The reason for this astonishing phenomenon is quite easy to understand: When the interaction is broken up, the previous partners of interaction are forced to pick their new partner from the pool of free players. As the cooperative players tend to be bound in partnerships by other cooperative players, the pool is made up mainly of cheaters. Therefore a cheater has only a small chance to find a new partner that can be exploited.

As can be seen, Schüßler started off with some arbitrary and at best plausible assumptions about an "anonymous market" that are in no way related to any specific empirical situation (points one and two in the above list of

features of Axelrod-style simulations). But Schüßler also had a deeper motivation for his simulation experiments, which brings us to the third point: the general conclusions that are derived from the simulation results. With his simulation that showed that cooperation could even emerge on "anonymous markets" Schüßler wanted to provide arguments against sociological normativism. Sociological normativism is by Schüßler understood as the thesis that social order cannot be upheld without social cohesion and the appeal to common norms. The classical proponents of sociological normativism are – among others – Ferdinand Tönnies with his distinction of "Gesellschaft" and "Gemeinschaft" and Emile Durkheim, who greatly emphasized the importance of social bonds. By some modern sociologists (including Schüßler himself) this question is linked with what they call Hobbe's problem, the problem whether and to what degree social order and coordinated action is possible without central authority. Schüßler's simulation is linked with the problem of sociological normativism in so far as it proves the "logical possibility" (Schüßler) of norm conformant behavior (if cooperation is taken as normatively desired in this case) even under absence of authority or other previously fixed coordination mechanisms such as cohesion. But does the proof of this "logical possibility" really establish a strong point against sociological normativism? This is not at all the case. The fact that something is logically possible does not even remotely imply that it is possible in reality. When sociological normativists speak for the importance of social bonds they usually do not mean to assert that it is by logical necessity that the social order requires some level cohesion to function properly. Rather, they draw on the social character of human nature. Therefore, in order to refute them, one has to show why their conception of human nature is wrong or that the empirical support for their claims is inconclusive and could be interpreted otherwise. Claims about mere logical possibilities as they appear in the highly stylized and artificial setting of agent based simulations are notoriously weak arguments in sociological discussions. Not the least so because it would probably be easy to draw up Axelrod-style computer simulations where under different but equally plausible boundary conditions cooperation is bound to break down when social ties are weakend.

To do Schüßler justice it must be mentioned that he is fully aware of the just mentioned explanatory limits of his computer simulations and that he discusses them frankly and with great intellectual honesty. It is only that doing so he makes the reader wonder why he did care to fill a whole book with computer simulations that demonstrate so little. The same questions could be asked for many of the simulations that have been carried through on the topic of the "evolution of cooperation". Most later authors were, like Schüßler, more careful than Axelrod in drawing sweeping conclusions from their computer simulations. But if no conclusions can be drawn from them, the question inevitably arises what these computer simulations are good for after all. It is this question that has become crucial in the case of Axelrod style-simulations. In order to answer it, let us see how Axelrod-style simulations fare when it is attempted to employ them in the context of an explanation of some real world phenomenon.

4.1.3 The explanatory irrelevance of Axelrod-style simulations in social sciences

The probably most dramatic example for Axelrod's theory of the "evolution of cooperation" is given in his chapter on the trench war on the western front in the First World War. During the long phases when no great battle took place, a rather surprising phenomenon occurred on many parts of the front in this war: Hostilities lost in intensity and the number of casualties was reduced to a figure that is surprisingly small given the fact that the soldiers virtually eyeballed their opponents on the other side. The phenomenon has been extensively studied by the historians of the epoch, among others by the sociologist Tony Ashworth (Ashworth, 1980), who found out that it was due to a kind of "live and let live" system that emerged on many (roughly one third) of the quieter parts of the front line: The soldiers hoped that if they weren't taking too hard on their enemies then the enemies would do the like to them. Thus, contrary to standing military orders, a kind of cooperation between the opposing front soldiers emerged on the basis of an unspoken "live and let live" agreement. Axelrod draws heavily on the description of Tony Ashworth as a source and he fully acknowledges Ashworth's achievements.

Axelrod treats the "live and let live" system in trench war as an excellent confirmation case for his theory. But would his theory really be able to explain the "live and let live" system? In order to find this out, let us see, whether Axelrod's computer simulations can add anything to the explanation of the "live and let live" system that goes beyond the explanation that is already given in Ashworth's historical narrative. To do so we first have to briefly reconstruct the explanation that is given by Ashworth and then check whether there exist aspects of the phenomenon that Axelrod can explain better.

Ashworth, in his historical treatment, identifies the following causes for the "live and let live" system:

1. The strategical deadlock. It was virtually impossible to move the front-line for either side.

2. The natural desire of most soldiers to survive the war.

3. The impersonal, "bureaucratic structure of aggression" (Ashworth, 1980, p. 76ff.).

4. Empathy with the soldiers on the other side of the front.

5. Whether elite troops or non elite troops were fighting on either side. "Live and let live" was much less frequent where elite troops were involved. (According to Ashworth this was the most decisive factor of all.)

6. The "esprit de corps" that can, however, become either conductive or, in the case of elite troops, impedimental to the emergence of the "live and let live" system.

7. The branch of service. Infantry soldiers had to face a much greater danger and consequently had a greater interest in "live and let live" than artillery soldiers.

8. The limited means of the military leadership to suppress "live and let live". (Only later they found an effective way to do so by organizing raids on the enemy trenches.)

9. Initial causes such as Christmas truces, bad wheather periods when fighting was impossible, coincidental temporary ceasefire due to similar daily routines on both sides (mealtimes).

At first sight it would seem quite obvious that Axelrod's computer model hardly captures any of these causes. If at all then only the first cause, the strategical deadlock situation the soldiers were caught in, could roughly be interpreted as a repeated prisoner's dilemma. But then, this is only one in a long list of causes, which means that Axelrod's model is far from fullfilling the adequacy requirement. It would therefore mean to strongly distort the historical situation if we were to maintain that the soldiers cooperated in the "live and let live" type fashion, because they were caught in a repeated prisoner's dilemma situation and because – as computer simulations demonstrate – "tit for tat" often is a good strategy in such situations.

However, if the model helps to give us a deeper or more precise understanding of one of the different factors that contributed to the "live and let live"-system, Axelrod's model would still have some explanatory value, even if only as a partial explanation. Also, we could still try to link some of the other causes to Axelrod's model by assuming that they determine the preferences of the soldiers and thereby the payoff parameters of the repeated prisoner's dilemma that – according to Axelrod's interpretation – they play with their enemies. For example, it is plausible to assume that the status of the troop (elite troop or non elite troop) had a bearance on how the soldiers valuated the situation they were in. While a non elite soldier would prefer to be a coward and live an elite soldier might prefer to fight and risk death. Consequently, elite soldiers might not even face a prisoner's dilemma. Quite in harmony with Axelrod's model, which suggests that cooperation is the rule and non cooperation the exception, this could help to explain why "live and let live" appeared only in one third of all cases.

The way Axelrod proceeded when determining the payoff parameters was

to asses by plausible reasoning the ordinal relations between the different alternatives for soldiers according to their assumed preferences. Unfortunately this is not enough, because the outcome of Axelrod's simulation is strongly sensitive to the cardinal values of the payoff parameters. This violates the stability requirement. Therefore we cannot really know whether the soldiers followed the "live and let live" strategy because of what Axelrod's model suggests.

More generally, the difficulty of applying Axelrod style simulations to political or historical science results from the problem that the values of the required input parameters cannot be found ready made in the historical records. They must be reconstructed through a complicated and error-prone interpretation process. It is therefore hard to see, how the stability requirement can be fullfilled at all for simulations that are not extremely robust right from the beginning. As we shall see later, a similar problem applies for the application of Axelrod-style simulations in biology. Only that there we have more reason to hope that it can be overcome by simulations that are more closely knit to the measurable quantities of the empirical processes.

What then are we left with? Since Axelrod's simulation as applied to the "live and let live" system of the first world war violates the both the adequacy requirement and the stability requirement (the latter is the case even, if we treat it as a merely partial explanation), it cannot claim to be explanatory. At best it delivers us an alternative metaphorical description for the strategical situation the soldiers found themselves in in terms of game theoretical concepts. Offering no more than that it has hardly anything to add to the detailed explanations Ashworth offers within his historical narrative.

The example shows how difficult it is to make any good use of Axelrod style simulations in the social sciences. Partly this has to do with typical difficulties that all formal approaches face in the social sciences outside economics. There are two main reasons for the limited success of formal methods in social sciences. First of all, social processes do often result from an intricate set of interwoven causes (see the example above), for only some of which we have a formal description ready at hand. But if we cannot single out the causes that can be described formally then any accuracy that is gained by

the formal description inevitably gets lost when we reintegrate the formally described causes with the other causes in a comprehensive explanation. The second reason is that measurement is difficult in social sciences and that only few quantities can be measured with accuracy. (In the above example, how would you measure the empathy the soldiers felt for the likes of them on the other side of the fontline?) It is not only true for computer simulations that our formal modelling is just as good as our measurement capabilities. Partly, however, the reason why Axelrod style simulations fare so badly is due to the fact that it is just a very incautious type of modelling.

4.1.4 Do Axelrod-style simulations do any better in biology?

The skeptical conclusion about Axelrod-style simulations the last section closes with becomes even more inevitable when we look at examples from biology, a field were the obstacles against formal modelling are much smaller than in social sciences. Being not a biologist myself, it would of course be difficult for me to estimate the usefulness of Axelrod style simulations for the explanation of cooperative behavior in biology. Luckily, there exists a comprehensive survey by the biologist Lee Allen Dugatkin on "Cooperation among Animals" (Dugatkin, 1997) that pays some particular attention to the manyfold of game theoretical computer simulations that have come up in the aftermath of Axelrod's "Evolution of Cooperation". In the beginning of his book Dugatkin list a whole number game theoretic computer simulations and their results, which – being the results of computer simulations alone – are purely theoretical of course. The major part of his book consists of a survey of the empirical research on the various instances of cooperative behavior that can be found in the animal kingdom. Interestingly, there exists not a single instance of cooperative behavior in the animal kingdom to which any (!) of these computer simulations could be applied in a strict sense.

This is not to say that biologists did not try to do so. The attempt has been made, for example, to apply Axelrod's and Hamilton's theory of the evolution of cooperation to the behavior of predator inspection that is found among various types of shoal fishes. In an early paper by Manfred Milinski on the topic (Milinski, 1987), Milinski tries to find out – with the help of an

inventive experimental setup – whether pairs of inspecting fishes play "Tit for Tat" like Axelrod and Hamilton postulated it for the repeated prisoner's dilemma. In order to do so Milinski also assesses (or rather estimates) the payoff parameters of Axelrod's model as applied to this particular case. Like Axelrod in the case of the "live and let live" system in the trench war of the First World War, he confines himself to an assessment of the ordinal relations between the payoff parameters. But unfortunately Axelrod's model is sensitive to the cardinal values of the payoff parameters. In later studies on the topic of predator inspection the attempt to explain this type of behavior with Axelrod's theory of the "evolution of cooperation" seems to have been completely dropped. In a paper that appeared ten years later (Milinski and Parker, 1997) than the first study, Milinski and Parker, even leave the question open, whether pairwise predator inspection is an instance of cooperative behavior at all, although this still appears likely. A major methodological problem is that – despite some very ingenious experiments – it is very difficult to measure or to estimate reliably both the risk a fish runs when inspecting a predator and the fitness relevant payoff a fish receives from inspecting. (The former has to some degree been achieved by Milinski and Parker, but the latter remains an open riddle).

As Dugatkin summarizes the situation in the concluding chapter of his book, there exists, with one exception, no case of cooperative animal behavior where the payoff parameters required as input for the game theoretical computer models could be measured. Therefore it is no surprise that none of the many Axelrod style simulations of the evolution of cooperation could be applied strictly to any of the empirical instances of cooperation in biology. It is therefore very doubtful whether this type of simulations (which remains remote from concrete empirical research and rests purely on "plausible" assumptions) is of any use for biologists at all. Another leading exponent of the game theoretic approach in biology puts it the following way: "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." (Hammerstein, 2003, p. 83) And after a long discussion of problems that the study of cooperative behavior of animals faces the same expert concludes: "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, S. 92)

One might object that maybe some of the models can be further developed so that they actually fit some of the empirical examples of reciprocity. This is of course true: It does not matter whether one starts constructing a model with a certain empirical application case in mind and builds it around measurable quantities (*bottom up approach*) or whether one starts with arbitrary plausible assumptions and only later on tries to adjust the model to specific empirical situations (*top down approach*). But the one way or the other, our the models and the empirical processes they are related to should be brought together. For, just because we have a model that shows us that for this or that reason cooperation evolves or breaks down, we cannot conclude for any empirical case of the evolution of breakdown of cooperation that it did so by virtue of the very same causes for which it did in the model. It could also have been the effect of quite different causes. Unless there is a *close fit* between model and reality we will never know.

But instead of seeking to achieve a fit between model and reality, the tradition of Axelrod-style modeling of the "evolution of cooperation" largely proceeded a different course. Computer simulation followed after computer simulation, each of them changing the basic configuration in some way or other or trying the addition of new and different parameters. But most of these simulations never got to the ground of empirical testability. This way, however, computer simulations only lead away from the real scientific problems.

4.2 Can we simulate the "Social Contract"?

But it is not just because it leads us scientifically astray that too much indulgence into pure model research is bad. The other problem is that it

may prevent us from seeing the most obvious, because our imagination is limited by the narrow lens of our own models. This is what seems to have happened to some of the modern game theoretical interpretations of social contract philosophy.

Such a game theoretical interpretation has been put forward, among others, by Brian Skyrms in two books, "Evolution of the Social Contract" and "The Stag Hunt and the Evolution of Social Structure" (Skyrms, 1996) (Skyrms, 2004), in both of which he presents computer simulations to deal with classical questions of social contract philosophy. The theory of the social contract is rooted in the philosophy of the 17th and 18th century with Thomas Hobbes' "Leviathan" being the most famous work dedicated to social contract philosophy and Skyrms believes that we can raise the discussion to a higher level by applying the modern tools of game theoretical computer simulations to it. It is as a stag hunt game that Skyrms presents the central question of the social contract in the latter of the two works (though he acknowledges that the prisoner's dilemma usually is the more common candidate): "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, p. 10)

Skyrms is by no means alone with his belief in the superiority of the "modern tools" when it comes to social contract philosophy. His belief is shared by many analytic philosophers. Thus we read in a recent introduction to philosophy: "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, p. 232) The author goes then on to introduce the prisoner's as a "modern tool" which – as the reader is to believe – allows to "make more precise the sort of argument Hobbes offered".

Is it really possible to "make more precise the sort of argument Hobbes

offered" by translating the metaphoric language of 17th and 18th century philosophers (e.g. Hobbe's state of nature, Rousseau's stag hunt metaphor) into precise game theoretic models? In order to answer this question, let's first see what kind of problems the social contract theories of the 17th and 18th century philosophers deal with and then examine Skyrm's treatment of these problems. Social contract philosophy is traditionally concerned with two different questions:

1. Normative social contract philosophy is concerned with justification of political order and, furthermore, with the requirements a political order must fullfill in order to be just.

2. Descriptive social contract philosophy tries to answer the question how political order evolves out of anarchy and how just (or democratic) order evolves from authoritarian order.

It is primarily the descriptive question of social contract philosophy that Skyrm's game theoretic discussion is addressed to. For that matter Skyrms presents a number of simple game theoretical models, including his simulation of the stag hunt game. His simulation of the stag hunt game is a territorial simulation, where players on a two dimensional plane play a (one shot) stag hunt game pairwise with their neighbors. They change their strategy (to cooperate or not to cooperate) depending on to the most successful strategy in their neighborhood. Skyrms then examines the effects of the respective sizes of the interaction and reproduction neighborhood. He summarizes his results as follows:

> How much progress have we made in addressing the fundamental question of the social contract: "How can you get from the noncooperative hare hunting equilibrium to the cooperative stag hunt equilibrium?" The outlines of a general answer have begun to 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. This can come to pass through pure chance and the passage of

> time in a situation of interaction with neighbors. ... The small group of stag hunters prospers and can spread by reproduction and imitation. The process is facilitated if reproduction or imitation neighborhoods are larger than interaction neighborhoods. (Skyrms, 2004, p. 123)

As far as the stag hunt game goes Brian Skyrms is surely right, but if this model is to tell us anything about how political order evolves from anarchy, Brian Skyrms completely misses the point. If Skyrms computer simulations of the stag hunt game really was an adequate model for the evolution of the social contract then we would have to conclude that political order could – if only the neighborhood structure were favorable enough – evolve from anarchy even without the institution of a Leviathan, merely by the gradual propagation of cooperation through neighborhoods. This is of course a most delightful prospect, though, sadly, one for which there exists not a single precedence in history. Other than Skyrm's simulations suggest and as the authors of the 17th and 18th century well knew, there is no way to bring about political order without a Leviathan of some kind. Apart from this lapse, which obviously the "modern tools" could not prevent, there is another omission that must appear striking to anyone who has ever wasted a thought on what the requirements and conditions of political order are: Nowhere in the various game theoretic models Skyrms presents in his two books is the phenomenon of rulership (Herrschaft) and submission (beherrscht werden) reflected, even though this is probably the most basic phenomenon of politics and a condition that surely any serious theory about the evolution of political order must take account of.

Surprisingly, this blunder went largely unnoticed in the lively discussion following the publication of Skyrm's first book. Only Philipp Kitcher points out that dominance hierarchies play an important part in evolution and that the sort of symmetric games Skyrms looks at do not properly reflect these (Kitcher, 1999). Kitcher's remark leads in the right direction, but it does not hit the point, because dominance hierarchies as they exist among animals as well as among humans are not the same as rulership, which is an exclusively

human phenomenon. The decisive difference is that a ruler can order a subject to do something, while dominance merely means that the others will give way to the dominant person (or animal), which is much less than carrying out orders.

But what caused this rather grave oversight? How come that Skyrms offers an answer to the fundamental question of the social contract that is obviously wrong? What Skyrms and other analytical philosophers seem to forget when they seek to make arguments from 17th and 18th century philosophers more precise by reformulating them in game theoretical terms, is that metaphors (like the state of nature metaphor or the stag hunt metaphor) do not get any better if one makes them more precise only on the side of the metaphorical image without paying due attention to the relation between the metaphor and its object. If the relation between the formal model that is to replace the metaphor and the object of the metaphor is not made more precise (in terms of adequacy, stability and descriptive appropriateness) the epistemological strength of the model is not any greater than that of the metaphor.[5] Or, to put it briefly: A metaphor remains a metaphor even if it employs a formal model as its object of comparison. For this reason it is also a fairly irrelevant question whether the state of nature is better described as a prisoner's dilemma or as a stag hunt game or as some other game. As the failure of Skyrms to provide a sound argument for either of the questions of social contract philosophy shows, the miscarried attempt to translate metaphoric descriptions into formal models may even disrupt the whole argument.

In the case of the Axelrod-style simulations of the "evolution of the social contract" discussed in the previous section it seemed that too much attention was spend on the construction of models and too little attention on whether the models are adequate. But when looking at Skyrms treatment of the social contract one may easily get the impression that he has never thought about the subject matter in question at all. At this point, however, it might be

[5]This is not to say that it is never useful to replace metaphors by models, for the epistemological strength of a model can – if the subject matter in question permitts – be increased, while that of a literal or poetical metaphor cannot.

asked if Skyrms really wanted to tell us anything about the social contract, or if he just wanted to show how some of the metaphors from the political philosophy of enlightment could be represented by game theoretical models without any specific claim about their applicability in any (including the original) context. But then, he explicitly relates his models to the social contract. If this is to be taken serious then the severe misunderstandings that result can only be due to the fact that he perceives his subject matter exclusively through the narrow lens of his own models. To give a name to the narrowing of perception or, rather, imagination as a result of the exclusive occupation with the technical aspects of formal modelling, I propose to call it "*model think*". "Model think" occurs when we conceive reality only through one specific brand of models and when we let other possibilities of conceiving reality escape our attention just because they cannot properly be represented with this brand of models.

5 Conclusions

Quite a few lessons that can be learned from the previous examples of failures of computer models. Some of them are truisms, but as they are often neglected they are important nonetheless.

First of all, if our models are to be explanatory then the establishment of a *close fit* between model and reality is at least as important as the construction of the model itself. The biological examples such as Milinski's and Parker's studies on predator inspection suggest that establishing this fit may even be much harder and more time consuming than constructing the model itself.

Secondly, when there is no *close fit* between model and reality, then the model has approximately the epistemological status of a metaphor. The results of such non explanatory simulations are hardly more than *computer generated metaphors*. Therefore, one must be very careful when drawing conclusions from them. At best one can regard these conclusions as mere hypotheses that still require an independent empirical confirmation. It should be clear that explanations based on non explanatory computer simulations amount to nothing more than *model based story telling*. I am introducing

these terms, because I believe that we need some negative catch phrases to characterize the misuses of formal models and, specifically, computer simulations.

Finally, we should be aware of the fact that although the ease and power of formal modelling has been greatly increased with the advent of the computer, there still remain scientific areas where the advantages of formal modelling are doubtful or where it is not possible at all. Computer simulations are just one scientific tool among others. It is helpful in some situations but useless in others. In my opinion the employment of the tool of computer simulations should be seen as something that requires justification. Apart from the aim to prove logical possibilities or to produce predictions it can be justified when there exists a close fit to the sort of empirical situation the simulation models or there is at least a realistic prospect of developing the simulation further so that a *close fit* can be established. Where computer simulations cannot not go beyond a merely metaphorical resemblance of empirical reality they are probably not worthwhile.

References

Appiah, K. A. (2003). *Thinking it Through. An Introduction to Contemporary Philosophy*. Oxford University Press.

Ashworth, T. (1980). *Trench Warfare 1914-1918. The Live and Let Live System*. MacMillan Press Ltd.

Axelrod, R. (1984). *Die Evolution der Kooperation* (deutsche Übersetzung, 5. Auflage (2000) ed.). R. Oldenbourg Verlag.

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

Dugatkin, L. A. (1997). *Cooperation among Animals*. Oxford University Press.

Hammerstein, P. (2003). Why is reciprocity so rare in social animals? a protestant appeal. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution*, Chapter 5, pp. 83–94. Cambridge, Massachusetts / London, England: MIT Press in cooperation with Dahlem University Press.

Kitcher, P. (1999). Games social animals play: Commentary on brian skyrms's *evolution of the social contract. Philosophy and Phenomenological Research LIX, No. 1, March*, 221–228.

Milinski, M. (1987). Tit for tat in sticklebacks and the evolution of cooperation. *nature 325, January*, 433–435.

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

Schüßler, R. (1990). *Kooperation unter Egoisten: Vier Dilemmata* (2. Auflage (1997) ed.). München: R. Oldenbourg Verlag.

Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge University Press.

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