Game Theory, Indirect Modeling and the Origin of Morality[1]

The final version of this paper appears in *The Journal of Philosophy, 2011, CVIII (4): 171-187)*

Arnon Levy

The Van Leer Institute, Jerusalem

1. INTRODUCTION

The idea that the origin of moral norms can be explained in Darwinian terms has been with us since Darwin himself. But recently it has taken a more promising form with the advent of evolutionary game theory. Combining concepts and formal tools from evolutionary biology and economics, current evolutionary game theoretic models embody a distinct picture of the origin of morality. On this picture morality is the product of cultural evolution. That is, the evolution of norms happens within culture as the aggregate product of individual-level behavioral learning. The circumstances in which moral norms come into play are represented as strategic interactions such as bargaining and cooperation/defection scenarios. A basic assumption is that social agents generally seek to improve their lot in such interactions and that they do so by learning from peers. Strategies that have secured higher payoffs will tend to spread, because social learners will tend to adopt them more readily. The result is a Darwinian process operating on top of the basic

[1] Many thanks to Peter Godfrey-Smith for reading several earlier drafts of this article. For their comments and assistance I am also indebted to Arthur Applbaum, David Haig, Ned Hall, Frances Kamm, Daniel Rosenbloom, Susanna Siegel and the participants at a colloquium in the Department for History and Philosophy of Science at the Hebrew University of Jerusalem.

set of human cognitive skills and motivations. Recent modeling has shown that such a process can give rise to conduct akin to cooperation, sharing, trust and commitment.

Successes in evolutionary game theoretic modeling have blown wind in the sails of an optimistic approach according to which, as Brian Skyrms puts it, we have "the beginning of an explanation" of the origins of morality.[2] Among modelers as well as commentators, there appears to be a sense that substantial progress has been made in understanding where moral norms comes from, how they change over time, and where, in principle, they might be going. But I will argue that while progress in modeling has been made, the sense that we are closer to an explanation of the actual origin of moral norms is mistaken, and the reasons underlying this mistake are worth attending to. Models of the evolution of morality contain a number of idealizations, especially with regard to the way in which they portray individual-level moral learning. Assumptions of this sort were made early in the project and follow-up work retains them, sometimes without noticing their effect on the explanatory power of the models. Taking a close look at the idealizing assumptions involved can teach us something important, I will argue, about the applicability of Darwinian ideas to culture.

Moreover, the case of evolutionary game theory reveals an error characteristic of a certain mode of modeling. Specifically, I think it will be helpful to think of the models in question as employing a strategy of indirect representation and analysis. In indirect analysis one explores a phenomenon in a once removed fashion: to learn about a target system in the world one analyzes a constructed scenario ("the model"), relying on similarities between the model and the target to gain knowledge of the target from an analysis of the self-standing model. I will argue that work in evolutionary modeling of morality has largely provided us with an understanding of models,

[2] Brian Skyrms, "Sex and Justice", *Journal of Philosophy*, XCI (1994): 305-320, p. 320.

but that some of the crucial similarity relations are missing. Thus, what has been achieved is more in the nature of understanding a stipulative picture of how moral change might occur in a world rather different from ours. It is not, importantly, knowledge about the actual dynamics of norms. I will try to offer a diagnosis of the source of the conflation of such internal progress, as I shall call it, with genuine explanatory success.

For concreteness, the discussion will focus on a family of models exploring the game of "Divide the Cake". These models have been developed over the last 15 years or so by Brian Skyrms and several collaborators in order to capture the origin of distributive justice.[3] The case of distributive justice is a central one which deserves attention in its own right. Since much of what I have to say about it generalizes in fairly immediate ways, I will not look at other examples. To forestall potential concerns about the kind of discussion I am engaged with, let me emphasize that what follows does not address the normative status of morality. The models I discuss target the process through which moral attitudes and moral conduct are shaped. This is a purely descriptive matter. The issue of whether an evolutionary account affects the norms we ought to hold or their epistemic standing is separate, and I do not address it here.[4] I begin, in the next section, by describing the work of Skyrms and his collaborators, paying special attention to the diachronic aspect of its development. There follows a discussion of the explanatory merits of "Divide the Cake" models, emphasizing the significance of some key mismatches between the

[3] I'll focus on the work of Brian Skyrms and Jason Alexander, discussed below. Related work includes Ken Binmore, "*Game Theory and the Social Contract Vol. I: Playing Fair",* (Cambridge: MIT, 1994); Zachary Ernst, "Explaining the Social Contract", *British Journal for the Philosophy of Science, LII* (2001): 1-24 and "Robustness and Conceptual Analysis in Evolutionary Game Theory", *Philosophy of Science,* LXXII (2005): 1187-1196; H. Payton Young, *Individual Strategy and Social Structure* (New Jersey: Princeton, 1998); Kevin Zollman, "Explaining Fairness in Complex Environments", *Politics, Philosophy and Economics VII (2008): 81-97.*

[4] Within the descriptive project, a more specific set of questions concerns the explanandum of an evolutionary explanation. What, exactly, is the explanatory target here? Is the aim to explain the origin of moral intuitions? Of evaluative tendencies? Or perhaps of conduct that merely conforms with moral standards? I do not have the space to provide an adequate discussion of this issue here. I will assume, in a charitable spirit, that whatever evolutionary game theory explains is closely connected with morality, and worth explaining.

models and reality. The final part of the paper is devoted to the issue of progress in indirect modeling.

2. THE INITIAL MODEL

The initial model, as I will call it, appeared in Brian Skyrms' elegant 1994 paper "Sex and Justice".[5] There he proposed an account of the origin of a share-and-share-alike rule of distributive justice. Skyrms' point of departure was an apparent puzzle that arises in the context of the game Divide the Cake. In this game two individuals are faced with the problem of dividing a unit good (the cake). The allocation procedure is the following. At one and the same time both players present a demand corresponding to their desired portion of the cake. If the demands sum to one or less, each gets what she demanded; otherwise both get nothing. A salient solution to this problem is an equal 50-50 split. Game theorists as well as laymen feel an intuitive pull towards equal splitting. In experimental tests in which subjects have been asked to play this game it is by far the most common outcome.[6] When asked, subjects typically explain that equal splitting seems to be the just (or fair) way to divide the cake. However, from the point of view of rational decision making there is nothing special about a 50-50 split. It is a Nash equilibrium of the game, i.e. neither player can benefit from unilaterally changing their strategy. But this holds for splitting 70-30 or 80-20 as well. The equal split is maximally efficient – none of the cake goes to waste – but again, so are many less equitable division schemes.

[5] Skyrms, op. cit.

[6] John H. Kagel & Alvin E. Roth, *Handbook of Experimental Economics,* (New Jersey: Princeton, 1998).

Skyrms proposed to look at the matter not as a decision-making problem, but as the outcome of a Darwinian process. He offered the following model.[7] Suppose a large population of agents plays Divide the Cake repeatedly. With a cake that can be sliced into *n* equal portions, the composition of the population is given by the vector $\vec{x} = x_1, x_2, x_3, \dots x_n$ where x_i is the proportion of players demanding *i* slices of cake. Over time, the frequency of types in the population changes in accordance with the so-called replicator dynamic:[8]

$$x_i' = x_i \frac{f_i(\vec{x})}{\overline{f}(\vec{x})}$$

Here f_i is the payoff of those demanding *i* slices – we can call them the "i-strategists" – and $\overline{f}$ is the average payoff in the population.[9] The replicator dynamic is one of the most commonly used equations of change in evolutionary modeling. Its key feature is that the proportion of a strategy in the population is a direct function of how it does relative to the average payoff in the population. As f_i increases relative to $\overline{f}$, so does the growth of x_i: the more successful a strategy is the faster it spreads. The replicator equation originates in evolutionary biology, but under certain assumptions it can be interpreted as representing a cultural dynamic driven by payoff-based imitation. If players imitate successful strategies more often than unsuccessful ones, the result is a Darwinian dynamic in which the successful strategies spread though imitation learning.

[7] H. Payton Young had previously offered a bargaining model for the evolution of the equal split, focusing on the same game. See: "An Evolutionary Model of Bargaining", *Journal of Economic Theory*, LIX (1993): 149-168. Young's model is similar in spirit to Skyrms', but some of its technical aspects are more complex and it is not as overtly designed to account for the evolution of justice.

[8] There are two differences between the initial model and a later model by Skyrms and Jason Alexander, which I discuss below. First, the later one allows slicing the cake in ten whereas the initial model employed a more restricted strategy space. Second, the initial model used a continuous-time version of the replicator equation. Here I present the discrete-time version used in the later work. Both points do not affect my arguments. For a comprehensive discussion of the replicator equation see Jörgen W. Weibull, *Evolutionary Game Theory* (Cambridge: MIT, 1995).

[9] The vector $\vec{\mathbf{x}}$ figures into the equation because the payoff gained by a strategist will depend on who it faces.

Analyzing this model, Skyrms' main finding was that what made the difference was how players were paired. In its simplest form the model posits a well-mixed population where players pair-up completely at random. In this case an equal split can evolve but unequal splits are quite likely too. The population may get "trapped" in an unjust state in which some players demand (and receive) more, others less. One can, however, modify the model so that pairing is non-random – specifically, so that an *i*-strategist has a higher-than-chance likelihood of playing against similarly-minded *i*-strategist. In this case we say that interactions are (positively) correlated. So modified, the model results in "demand 1/2"[10] invariably taking over the population and the equal split predominates.

3. ELABORATING THE PICTURE

The initial model was elegant and the analysis yielded an interesting result, but it was also a highly abstract representation of the social process in question. Agents and their strategies were represented statistically, and so were the overall effects of social learning. Correlated interaction, the key ingredient assuring reliable emergence of the equal split, was posited but no mechanisms for its generation was incorporated into the model. As a consequence some critics challenged the rationale for Skyrms' stipulations, whereas others alleged that once more details were included, the results would not hold up.[11] Partly as a response to these criticisms, the years since the initial model have seen Skyrms and a number of collaborators seek to provide more concrete follow-ups that flesh out the initial model.

[10] Or, correspondingly, "demand-5" when the cake is sliced into 10 parts, as in the models discussed below.

[11] Examples include: Justin D'arms, Robert Batterman, & Krzyzstof Górny, "Game Theoretic Explanations and the Evolution of Justice", *Philosophy of Science,* LXV (1998): 76-102; Neil Tennant, "Sex and the Evolution of Fair Dealing", *Philosophy of Science*, LXVI (1999): 391-414.

To do so they turned to agent-based modeling. An agent-based model represents players explicitly – as individuals with a strategy, a learning rule, a record of success in the game, often also as situated within a social network. In 1999 Skyrms and Jason Alexander proposed a model of this kind for Divide the Cake, as an extension of the initial model.[12] They considered a population of 10,000 agents playing Divide the Cake on a large lattice. In every iteration of the game each player is paired to her 8 immediate neighbors in turn. Each game involves, as before, the players making demands according to their strategies and garnering payoff. At the end of each round each player compares herself to her neighbors and adopts the strategy of the neighbor with the highest overall payoff for that round.[13] This way of updating one's strategy is known as "imitate your best neighbor".

The results were consonant with the main message of the initial model. Skyrms and Alexander found that for upwards of 99% of population configurations they simulated, the outcome was an equal split.[14] All that was required was the presence of a small cluster of interacting demand 5-ers – these then went on to "infect" the rest of the lattice. Recall that in the initial replicator model the key requirement for the evolution of the equal split was correlated pairing (positively correlated pairing, that is). In the agent-based model this is, in an important sense, still true. Positive correlation arises from the network structure, as "new" equal-splitters are inserted nearby the neighbors they take after. Recall also that the initial model involved a payoff-driven dynamic generated by imitation learning. This too is true in the later model, where imitation is given a concrete "imitate your best neighbor" form.

[12] Jason Alexander & Brian Skyrms, "Bargaining with Neighbors: Is Justice Contagious?", *Journal of Philosophy*, XCVI (1999): 588-598.

[13] So long as the neighbor is more successful than her. Ties are broken by a coin toss.

[14] Due to the large number of parameters in such models, they are typically simulated on a computer, rather than solved analytically.

In recent work, Jason Alexander (now McKenzie Alexander) has delved deeper into agent-based models of Divide the Cake.[15] In a book-length treatment of the origin of morality, he explores a range of network models, including "small-world" networks, where phenomena like six-degrees-of-separation are common and bounded-degree networks, which have a constrained random structure. McKenzie Alexander also enlarges the class of imitation rules to include, e.g., "imitate with proportion relative to success"; or a rule that has players play the game with their neighbors but imitate the agent with the best score in the entire population ("imitate the winner"). In these models too there is a remarkable tendency for populations to stabilize at an equal split. As in earlier models, the combination of structure-generated correlation and payoff-based learning ensures that the strategy demand-5 is superior to more (or less) demanding alternatives.

A few words about the developmental trajectory of this work has will serve to summarize the discussion so far and set the stage for the argument to follow. The move from the initial replicator-based model to current agent-based models involved primarily a process of concretization: filling in specifics where the initial model contained abstract posits. The early model presupposed a payoff-driven populational dynamic, supervening on individual level imitation learning. But it provided little by way of detail on how social learning works. Later models made explicit various forms it could take. The initial model posited a like-meets-like bias in the pairing of players, but did not fill in this assumption – it took the form of a correlation parameter. In later models correlation was generated endogenously due to the fact that interactions occurred within a network and updating was neighborhood-based. Thus, the newer models offered an elaboration of the earlier model, proposing mechanisms for how its features

[15] Jason McKenzie Alexander, *The Structural Evolution of Morality* (Cambridge, UK: Cambridge, 2007).

arise. Significant details were added, but the basic picture was retained: a population of success-seekers in which individuals have a better-than-random chance of meeting their ilk will tend to stabilize at a norm of equal sharing.

How much have we learned from these models about the origins of justice? Below, my goal is to answer this question. Let me emphasize that I am construing it as a substantive, largely scientific issue: I want to ask whether the models explain (or begin to explain) the origins of justice in the familiar, primarily causal sense in which scientific theories explain phenomena. My discussion does not rely on an expectation that the models in question yield testable or fine-grained predictions. To expect that would be unreasonable, as we are dealing with a proposal for an explanatory framework, the subject matter of which is a set of highly complex phenomena. But I will take it that the goal is to understand the origins of actual justice in the actual world. My aim to assess whether the picture offered by game theoretic models is on the right track.

4. SYMMETRY

Let us first focus on some key symmetry assumptions. The game of Divide the Cake depicts two agents that have identical preferences, no special needs and no claims on the good they are dividing. Moreover, the good itself is nondescript: it is a windfall with no history nor an intended use, none of the agents involved had any role in producing or procuring it, it does not belong to anyone nor are there effects to distributing it beyond the immediate payoffs involved. The game describes in a simple form a pervasive kind of interpersonal interaction – bargaining for a resource in scarce supply – but the interaction occurs against an asocial background.

These symmetry assumptions greatly affect the kinds of norms that can be represented. For instance, many actual norms of distribution, including a variety of egalitarian norms, take into account the relative contributions of agents – direct or indirect – to the production of the goods being distributed. Many norms also take into account the effects of distributing a good, how needy the distributees are, as well as the ends to which they will use their share.[16] Such attributes of players are not represented in the game of Divide the Cake and, at least with respect to some of them (e.g. neediness), it is unclear that a model of this sort can represent the relevant facts. Moreover, it is highly doubtful that an evolutionary model that incorporated these aspects would yield an egalitarian result similar to equal splitting.

Thus, while the equal split might be representative of egalitarian rules of justice, its degree of generality is fairly limited. Even if we were to take the Divide the Cake models to provide a full explanation of the tendency of laboratory subjects to demand ½, and the concomitant intuitive pull of the equal split solution, it is far from clear that we can extrapolate from this to other cases where distributive problems arise, even cases in which an egalitarian solution is salient.

5. TYPES OF LEARNING

Now let us set aside worries about generality and look at a more fundamental explanatory question. The evolutionary game theoretic models that are our focus portray moral change as a cultural evolutionary process. This type of populational change is not biological, but it bears

[16] Empirical results in psychology shore up these casual reflections. See Menachem Yaari & Maya Bar-Hillel, "On Dividing Justly", *Social Choice and Welfare,* I (1981): 1-24. Maya Bar-Hillel & Menachem Yaari, "Judgments of Distributive Justice" and Richard J. Harris, "Two Insights Occasioned by Attempts to Pin Down the Equity Formula", both in Barbara A. Mellers, & Jonathan Baron (eds.), *Psychological Perspectives on Justice*, (New York: Cambridge, 1993).

important similarities to evolution by natural selection.[17] The units of cultural moral evolution (on the game theoretic picture) are moral "strategies" which have consequences for population-level ways of dividing goods. The process is driven by the differential adoption of strategies at the individual level, where learning serves in the role of heredity. Thus we have an analogue of "heritable variation in fitness" – to use Richard Lewontin's phrase – the basic pattern of a Darwinian process of change. The key to assessing the explanatory value of the game theoretic framework lies in determining whether it is plausible that a Darwinian dynamic underlies cultural moral change. To do so we need to take a closer look at social learning, and some different forms it may take.

Success learning. On the picture underlying game theoretic models such as Divide the Cake, social learning is the engine of moral change. The central idea is that moral learning is driven by success in strategic interactions. The fact that learning results in the spread of successful strategies is what gives the model its Darwinian character. 'Learning' is used here in a generalized fashion to refer to ways in which an individual adopts new behavioral routines. What distinguishes different learning rules are the biases they introduce into the adoption of new behaviors. In success-based learning, as I will call it (or simply 'success learning')The adoption of new behaviors is biased towards those that increase payoff. That is, individuals adopt new behaviors in a payoff-driven fashion, seeking strategies that will increase their gains in future rounds of the game. The idea that learning is success-based is pivotal to the game theoretic framework and to the explanation of justice here at issue, so it is worth going into more detail.

[17] Many (arguably most) writers in this area would nowadays reject the idea that the contours of moral norms in present day societies can be explained purely, of even primarily, as the result biological evolution.

The initial model employed the replicator equation. As noted above, this equation of change has strategies spread as a function of how well they are doing relative to the population average. Strategies that do better than average proliferate. The replicator equation can receive a cultural interpretation. The idea is that if individual agents tend to mimic those who perform better, strategies that do better will get mimicked at a higher rate across the population. The replicator equation thus represents the aggregate effects of individual-level success learning. In the later agent-based models success learning is given a more explicit form. Individuals are represented severally and so for every individual the model must specify an update rule. This allows the injection of some real-life complexity into update rules. Skyrms and Alexander used a characteristic success-based rule – "imitate your best neighbor". Other success-based imitation rules include "imitate proportional to success" and "imitate the winner". But success learning is not especially tied to imitation. Consider "best response": this rule has the player monitor the behavior of all her neighbors and then calculate which strategy will give the highest overall payoff in the next move, assuming neighbors will continue to behave as they did. "Best response" is a high-rationality mechanism compared with imitating the more successful. Less sophisticated rules also exist. In reinforcement learning, for instance, a player chooses her strategy based on the cumulative payoff it has yielded in the past.[18]

What I have called success learning is therefore a bias in the direction of adopting behaviors associated with specific expectations about success – those that are likely to yield higher payoffs in the future, or have done so for others (the latter, of course, is often a good proxy for the former). The appeal to success-based rules represents a picture in which agents are driven to

[18] For a discussion of these and other learning rules see H. Payton Young, *Strategic Learning and its Limits, (*New York: Oxford, 2004); Drew Fudenberg & David K. Levine, *The Theory of Learning in Games,* Cambridge: MIT, 1998).

change their behavior by directly attending to the relation between their success and that of their peers.

To better understand what is distinct about success learning, it will help to consider other types of social learning. Here are two important ones.

Source learning involves a bias in favor of particular role players in the learner's environment – parents, teachers, superiors, celebrities. In source-based learning an individual's adoption of behaviors is guided by the fact that other individuals, those who occupy a particular role, exhibit this behavior. In recent work, Kevin Laland has referred to this as "who" learning.[19] There may be a (biological) evolutionary explanation for why children and young adults tend to emulate the morality of parents or other salient individuals in their surroundings.[20] The resultant sharing of knowledge or the strengthening of in-group cohesion may have enhanced ancestral fitness. But models of cultural evolution start, as it were, where biological evolution ends.[21] Their concern is with figuring out how the process of cultural change looks *given* a preexisting cognitive toolkit, including patterns of learning and social adaptation. From the perspective of modeling a process of cultural change, source-based learning behaves very differently from success learning.

In source learning one follows in the footsteps of salient individuals such as parents despite of, or irrespective of their success. It is easy to translate this idea into specific modeling suggestions. Consider a network model of the sort discussed above. In any given neighborhood one individual is designated the leader. A simple (and rather boring) "follow the leader" rule has

[19] Kevin Laland, Social Learning Strategies, *Learning & Behavior*, XXXII (2004): 4-14.
[20] They emulate, of course, their behavior and beliefs more generally.
[21] It is likely that both processes occur at once. But it is also likely that their time scales are substantially different, and for this reason one can treat the biological as fixed when studying cultural evolution.

all individuals adopt the behavior of their local leader. Probabilistic leader-following is a reasonable next step. A plurality of leaders, with different locations and strengths could be introduced, and so on.

Frequency-based learning involves another type of bias in the adoption of behaviors. Here an individual adopts behaviors as a function of how common they are in her environment. The clearest examples involve herd phenomena, or conformism more generally. Of course, as with source-based learning, there may be evolutionary explanations for frequency biases.[22] But from the point of view of modeling a social dynamic, frequency-based learning differs greatly from success learning.

As with source learning, the idea of frequency-bias can be readily modeled. It is also quite apparent that processes in which herd effects play an important role will tend to behave differently than success-based processes. The initial configuration of a population will matter more, and one would expect to find snowball effects owing to the self-reinforcing nature of conformism and herd phenomena. Frequency bias also leads to populations that are less susceptible to small perturbations, erratic flips in an individual’s behavior, trial and error and the like. When being common is what determines a strategy’s fate the appearance of a small number of outliers is less likely to make a significant impact.

6. JUSTIFYING SUCCESS LEARNING?

[22] See e.g. Jeremy Kendal, Luc-Alain Giraldeau & Kevin Laland, The Evolution of Social Learning Rules: Payoff-Biased and Frequency-Dependent Biased Transmission, (forthcoming), *Journal of Theoretical Biology*.

I have presented success learning, source learning and frequency bias in their pure forms. It is natural to suppose that all three play a role in moral learning, as well as other factors besides (e.g. random experimentation, inertia).[23] Highlighting the contrast between the different types of learning is meant to bring out the fact that in evolutionary game theoretic modeling an assumption is being made, namely that moral learning is purely success-based (call this the assumption of success learning). The explanatory value of the Divide the Cake model and others like it depends on this assumption. How plausible is it?

There are two kinds of ways, I think, of justifying the assumption of success learning, consistent with the models having substantial explanatory value. One could provide empirical evidence that moral learning is largely success-based. Or one could show that it does not matter: the results of the model do not depend very strongly on the assumption. Neither argument has been made, as far as I know, in the context of investigating the origin of morality. If made neither, I think, would succeed.

The empirical literature on moral education and development is not couched in terms of the categories I have discussed. But there is ample evidence that social learning in general, and moral learning in particular, involves source and frequency biases. Many psychologists studying moral development take parental guidance to be the key to the acquisition of moral norms.[24] There are cognitive psychological experiments that demonstrate that people differentially pick up

[23] How these factors interact, and how to model their interaction, are interesting questions in their own right. It is possible their effects are additive, in which case each factor can be assigned a weight in an overall sum. More complex possibilities include default/threshold rules, e.g. “follow the leader, unless you can gain more than x by doing otherwise” or “maximize payoff, unless you find yourself in too small a minority”.

[24] See essays collected in Jerome Kagan & Sharon Lamb, *The Emergence of Morality in Young Children*, (Illinois: Chicago, 1983) and Judith G. Smetana & Melanie Killen (eds.), *Handbook of Moral Development*, (New Jersey: Laurence Erlbaum Associates, 2006).

and retain information from designated sources in their environment.[25] Psychologist Joseph Henrich and anthropologist Francisco Gil-White have emphasized the importance of such mechanisms for cultural change in general.[26] They discuss evidence showing that prominent individuals in a community tend be listened to and emulated, both with respect to factual beliefs and with regards to standards of conduct. Interactions with socially prominent individuals have a particular behavioral-cognitive signature too. Anthropological observations in traditional communities show that elders are observed more closely and that information from and about them is retained better.[27] More generally, it is highly plausible that power asymmetries and hierarchical organization are produced and sustained by their influences on the conduct of individuals. Such hierarchical power structures occur within small-scale communities, within the family and in the context of specific institutions in politically organized societies. In all these contexts top-down influences clearly play a role in shaping individual moral conduct – either via mechanisms of emulation and internalization, or through fear and compliance. The extent of these kinds of effects is hard to assess, but the range, strength and frequency of asymmetries in social standing suggest that such influences are important indeed.[28]

The second way one could justify the assumption of success learning is via robustness analysis. Roughly put, robustness analysis shows how well a result holds up in a variety of set-ups, thereby demonstrating which factors make a difference to it and which do not. [29] Modeling

[25] Thomas Holtgraves, John Srull & Janet Socall, "Conversational Memory: the Effects of Speaker Status for the Assertiveness of Conversation Remarks", *Journal of Personality and Social Psychology*, LVI (1989): 149-160.

[26] Joseph Henrich & Francisco Gil-White, The Evolution of Prestige Freely Conferred as a Mechanism for Enhancing the Benefits of Cultural Transmission, *Evolution and Human Behavior*, XXII (2001): 165-196.

[27] *Ibid*.

[28] There are also theoretical reasons for thinking that source-specific influences would have biologically adaptive advantages. See Robert Boyd & Peter J. Richerson, *Culture and the Evolutionary Process*, (Chicago, 1985) as well as Sabine Coussi-Korbel, & Dorothy M. Fragaszy, "On the Relation between Social Dynamics and Social Learning", *Animal Behavior, L (1995): 1441-1453.*

[29] For a philosophical account see Michael Weisberg, Robustness Analysis, *Philosophy of Science,* LXXIII (2006): 730-742.

the pressure and volume of a gas (of fixed volume and temperature) while varying the shape of the chamber is a way of examining whether the geometry of the chamber is a difference making factor (for the dependence of volume on pressure). If varying the geometry has no effect on the relationship between pressure and volume then this is evidence that it is not a relevant difference maker.[30] Analogously, if we substituted for the success-based learning rules used in Divide the Cake models a variety of rules like the ones discussed in the previous section and if, doing so, we still observed a strong tendency to arrive at an equal split, this would tell us that equal splitting is not heavily dependent on the assumption of success learning. Such analysis has not, to the best of my knowledge, been done. Were it to be done it is rather unlikely, I think, that equal splitting would be shown to be robust across different learning rules. If initial conditions include a substantial number of non equal-splitters it seems that a conformist bias would land the population in one of the common strategies. If source learning plays a significant role, it would matter greatly what a small number of leader nodes in the social network do, less so the strategies of the mass of followers. Thus I think it is doubtful that one can justify the focus on success learning by arguing that little causal information is lost by the exclusion of other types of learning.

A different kind of reply to the criticism I have made is that it is premature. Models of the evolution of morality, and the models we have looked at in particular, are at an early stage. Like most modeling, its beginnings are simple and limited. The models target a restricted set of phenomena, make simplifying assumptions and aim for a partial fit with the world. Specifically,

[30] Of course, one needs to ascertain that geometry was varied in the right ways, and that other significant factors have been held fixed.

one might concede that the assumption of success learning represents an over-simplification, but argue that it is merely a natural starting point, not an end to the investigation. Criticisms of it are founded on an unreasonable expectation that the model take the full complexity of real-life social phenomena into account. Such “start simple” responses are encountered fairly often in discussions of models. Impatient outsiders seek completeness while modelers exhibit a more pragmatic, one-factor-at-a-time approach, seeking a good understanding of simple scenarios before moving to more complex ones.

The “start simple” response is important and as far as it goes, correct. But it cuts both ways: If models of the evolution of morality are at an early stage in which it is hard as yet to justify the idealizing assumptions being made, then optimism about the explanatory value of this work should be kept in check. Indeed, caution should be taken with respect to the basic message of work on the evolution of morality – namely, that moral conduct evolves, in a way closely analogous to how biological traits evolve. The assumption of success learning is what gives the models their Darwinian character; it is what ensures that proliferation tracks payoff. Doubts concerning this assumption undermine the basic picture of morality as evolving through a form of Darwinian cultural selection.

Success learning appears as a natural starting point if and to the extent that one thinks that it is likely that morality evolves by something like natural selection. Now, it is sometimes thought that if there is to be a scientific account of cultural change and the stabilization of norms, then this account must in some important sense be akin to Darwinian natural selection. But it is important to see that is not so. What is characteristic of Darwinian dynamics is that the differential spread of types in a population tracks success. In biology success equals increased offspring; in cultural evolution it translates into the rate at which individuals adopt a behavior or an idea. Some populational trends are not Darwinian in character. Bank runs, dress fads and other herd phenomena are not primarily Darwinian. Neither are top-down changes, such as the near cessation of religious activities in the former U.S.S.R. Herd behavior is not driven by success but by a sheer snowball effect: people emulate what appears to be most common. Top-down changes are sometimes driven by success, but only that of the individuals at the top. Thus, neither process is Darwinian.[31] Now, importantly, moral change might be more like herd behavior or top-down changes than it is like natural selection. It would be if frequency effects and source learning are rife. If so, the idea that morality evolves is fundamentally incorrect. In this sense the most basic message of work on the evolution of morality will not have been borne out.

7. MODES OF PROGRESS

The discussion of simple beginnings directs our attention to the diachronic aspect of model-building. In modeling one starts simple but, ordinarily, seeks to add complexity as time goes by. It is to the diachronic aspect of modeling the evolution of morality, and of modeling more generally, that I want to turn in this final section.

I have throughout talked about explanations of the origins of morality in terms of models and modeling, but I have not yet said much about model-building as such. As I understand it, modeling is a part of scientific (and perhaps philosophical) theorizing characterized by a pragmatic approach. Instead of proposing a broad, unified theory of a certain domain of phenomena modelers proceed piecemeal by offering a family of partial theoretical proposals, in

[31] For a related discussion see Peter Godfrey-Smith, *Darwinian Populations and Natural Selection* (New York: Oxford, 2009), especially Chapter 8.

the hope of attacking a large and complex problem by breaking it down into tractable parts. Theorizing about the evolution of morality is in this respect a typical case of modeling. Models, in general, are usually local in scope and tend to contain significant idealization – assumptions that are known to be false are introduced in order to simplify the analysis and isolate causal factors.

One important kind of model-based inquiry involves the indirect investigation of a real-world phenomenon ("the target") via the analysis of a hypothetical construct ("the model"). [32] Substantial parts of theory in evolutionary biology and in economics have this character. If model and target sufficiently resemble one another in their causal structure and/or in their input-output behavior, the analysis of the model can provide information about the target. Indirectness is often indispensible in that it affords a separation of the analysis of theoretical possibilities from an evaluation of their empirical plausibility. In the sciences of complex phenomena theorizing that is overly focused on the real world may hamper understanding rather than promote it. But indirectness can lead theorists astray precisely because it legitimizes a looser fit between theory and reality. One goal of the present paper has been to evaluate the explanatory merits of game theoretic models of the evolution of morality. Another is to connect this evaluation to considerations that apply more generally to the process of modeling and its pitfalls.

Let me distinguish two modes in which progress in indirect modeling can be made. Both have epistemic significance, but of a different sort. Both characteristically involve moving from simple beginnings to more complex analyses. One mode of progress may be labeled the "target-oriented" mode. Here the move towards complexity is sensitive to factors that affect the behavior

[32] I draw here, with some modifications, on a framework developed by Ronald Giere in *Explaining Science: A Cognitive Approach* (Chicago: Chicago, 1988); and more recently by Peter Godfrey-Smith, "The Strategy of Model-Based Science", *Biology & Philosophy*, XXI (2006): 725-740; and Michael Weisberg, "Who is a Modeler?", *British Journal for the Philosophy of Science*, LVIII (2007): 207-233.

of the real world target, elements that make a causal difference to it. In the simplest (and perhaps ideal) case a model starts by examining one or a small number of important factors, typically in a simplified way, and initial results are obtained. As time goes by an effort is made to incorporate further factors and to gauge the importance of the factors included in the initial proposal. Additional elements are either brought in or discovered to be irrelevant. If this kind of effort goes well then the result is greater understanding of a real-world target: progressively better knowledge about the causal process underlying the explanandum. We can contrast target-oriented progress with an “internal” mode of progress, which is more conceptual in spirit. Here a particular picture or set of assumptions is laid out, and then explored in its own terms. Detail is added primarily in order to refine the understanding of elements that are present to begin with. Internal progress is made to the extent that the initial construct and its behavior are well understood, and that refinements are seen to either bear out and illuminate the initial results or complicate them in interesting ways. Thus, in a target-oriented mode a modeling endeavor makes progress by incrementally adding causal information. This can be done via experimental testing and empirical observation or via robustness analysis, but the crucial point is that the work is regulated primarily by its empirical adequacy. In contrast, in an internal mode one explores the subtleties of a constructed set-up, but this is done largely independently of the actual causal relevance of the set-up being explored.

Skyrms’ initial model depicted a cultural process akin to evolution by natural selection and showed that it could lead to an egalitarian rule of justice. The underlying causal picture was of individuals haggling over a good and adapting their behavior to achieve greater bargaining success. Skyrms found that correlated pairing of players was vital. Later work was primarily devoted to filling in the initial picture. Network structure, which generated the correlation among players endogenously, was introduced. So were specific success-based learning rules which gave concrete form to the statistical trend embodied in the replicator equation. This amounted to real progress, I think, but progress of the internal kind; it was an exploration of the same sort of set-up that was posited at the start. It was not progress of the target-oriented, causally informative variety. Little or no attempt was made to provide empirical grounding for the idea that "straight" bargaining is the right way to capture justice or for the key simplifying assumption that moral learning is success-based. Nor was there a theoretical attempt to examine how much of an effect these assumptions have on the basic message that morality evolves through a cultural process akin to natural selection.

Now, some statements made by modelers in this area suggest otherwise. They assert that substantial explanatory progress has been made. Skyrms and William Harms, for instance, state that in the past "the nature and source of [moral] standards have remained something of a mystery." But that "[r]ecent work on the evolution of norms has changed this picture dramatically." [33] In light of the preceding discussion, I think that what is going on here, at least some of the time, is that internal progress is being mistaken for target-oriented progress. This is a mistake which, I want to suggest, is quite closely connected with theorizing in the indirect style, and the two modes of progress I've outlined. For while they differ methodologically and epistemically, internal progress and target-oriented progress often look and feel quite similar from the point of view of model development. Both characteristically involve incremental concretization –adding detail, positing specific mechanisms, substituting aggregates by explicit representation of individuals and so on. This phenomenological similarity can be mistaken for

[33] William Harms & Brian Skyrms, "Evolution of Moral Norms", in Michael Ruse, ed., *The Oxford Handbook of Philosophy of Biology* (New York: Oxford, 2008), p. 434.

epistemic parity, with the result that research that has taken a largely conceptual route appears to have yielded progress in answering explanatory questions about the empirical world.

While my argument only licenses conclusions about the case of models of the evolution of morality, I think it is important to recognize that this *kind* of situation can, in principle, occur in other instances of indirect modeling. That is to say, there is a type of error that can arise when an indirect modelling project, with its associated phenomenology, is developed in one way, but interpreted another way. Notice that this diagnosis differs from a suggestion that is commonly made when models are thought to float free from reality, namely that mathematical elegance or computational tractability are over-valued relative to empirical adequacy. Such complaints have recently been made, in extra-academic settings, against macroeconomists and string theorists. [34] It is possible that tractability and elegance occupy too prominent a role in these instances. But it often seems that accusations of this sort present modelers in an unreasonably crude light. On the present proposal the error is more subtle than that. It does not involve a simple conflation of mathematical appeal and truth. Rather, it arises when one fails to distinguish two kinds of motion towards greater complexity which are genuinely similar in some respects, but differ in their epistemic import. Whether and to what extent this has occurred in any particular instance depends on the details. A healthy skepticism would counsel in favor of bearing the possibility in mind.

[34] With respect to the former see, e.g., Paul Krugman, “How Did Economists Get it so Wrong?”, *The New York Times Magazine*, September 2, 2009. (I am indebted to Peter Godfrey-Smith for drawing my attention to this aspect of Krugman’s article). For the latter see: Lee Smolin, *The Trouble with Physics* (New York: Haughton Mifflin, 2006).