The Stag Hunt and the Evolution of Social Structure, by Brian Skyrms. Cambridge University Press, 2004, 149 pages.

In this short but substantial book, Brian Skyrms draws together a number of results from several previously published papers on game theory, signalling, and the origin of social structure to provide an insightful analysis of the problem of cooperation and collective action. Although this book may be viewed as a sequel to Skyrms' 1996 *Evolution of the Social Contract* –the text on the rear jacket suggests as much – *The Stag Hunt* fails to be a sequel on two grounds. First, a typical sequel extends the story of an earlier work by reworking themes within the same narrow framework of previously established constraints.[1] Second, thanks to Hollywood, we also tend to expect each sequel, upon completion, to leave increasing amounts of uninteresting ground for future exploration. Neither of these statements is true. *The Stag Hunt* is much more than a mere sequel.

The Stag Hunt agrees with the general spirit of its predecessor in that traditional problems of the social contract (such as the emergence of fairness and cooperative behaviour, the persistence of both in the face of deviants, to name a few) are better addressed from an evolutionary point of view than from that of traditional decision and game theory. However, whereas *Evolution of the Social Contract* employed the replicator dynamics as its primary tool, *The Stag Hunt* shifts attention away from the replicator dynamics to other models of greater realism and conceptual interest. Two alternative approaches to modelling cultural evolution — local interaction models (the subject of part I) and dynamic social networks (the subject of part III) — dominate the discussion, with the replicator dynamics relegated only to the middle chapters (part II) involving signalling.

The move away from the replicator dynamics serves to address one important, general problem with the methodology of *Evolution of the Social Contract,* first noted for the particular case of the evolution of fair division by D'Arms (1996) and elaborated on by D'Arms et al. (1998). If it is the case that "the key to the evolution of cooperation, collective action, and social structure is correlation" (Skyrms, 2004, pg. xii), wherefrom does the correlation between behaviours or strategies originate? In *Evolution of the Social Contract*, correlation between strategies was loaded into the replicator dynamics as a free parameter (see Skyrms, 1996, pg. 19). The absence of a plausible, independent story underwriting the introduction of such correlation rendered the explanatory story suspect. What prevented one from invoking correlation of strategies whenever it proved expedient to do so?

In *The Stag Hunt*, the discussion of local interaction models in Part I illustrates how correlation between strategies can naturally arise if interactions between individuals are constrained according to a previously given network of social relations. To block objections that this simply pushes the explanatory story back a step, Part III provides a model of how such a social structure may come about from purely random interactions between individuals who modify their behaviour according to plausible (and experimentally corroborated) learning rules, with one caveat that I shall return to later. The capstone of part III combines the evolution of social structure with the evolution of strategies, arguing that "the most empirically realistic version of

[1] *The Godfather: Part II* being the notable exception to the rule.

reinforcement learning, fluid-interaction structure and slow imitation decisively and unambiguously tip the scales in favour of cooperation" (Skyrms, 2004, pg 122). Very nice.

However, those acquainted with yet other criticisms of *Evolution of the Social Contract* may find their concerns with the overarching project unaddressed. Consider, for example, Kitcher's objection regarding evolutionary game theory's ability to explain morality[2]:

> "[I]t's important to demonstrate that the forms of behaviour that accord with our sense of justice and morality can originate and be maintained under natural selection. Yet we should also be aware that the demonstration doesn't necessarily account for the superstructure of concepts and principles in terms of which we appraise those forms of behaviour." (Kitcher 1999)

The Stag Hunt does relatively little to address these charges of overreaching behaviourism. A primary analytic concern remains identifying the basins of attraction for strategies which are behaviourally equivalent to cooperative, or just, actions. When the underlying models have several free parameters, Skyrms identifies — in many cases — how the basins of attraction change as the parameters are varied. Yet the focus on forms of behaviour, rather than concepts or principles or sentiments, still dominates.

One should keep in mind, though, that the rhetoric has changed between *Evolution of the Social Contract* and *The Stag Hunt* to mitigate the criticism of excessive behaviourism. Whereas the former, for example, claimed to have provided a possible beginning "of an explanation of the origin of our concept of justice" (Skyrms, 1996, pg. 21), *The Stag Hunt's* express aim is more muted. The "fundamental question of the social contract", writes Skyrms, is "[h]ow can you get from the noncooperative hare hunting equilibrium to the cooperative stag hunting equilibrium?" If *this* is the fundamental question of the social contract, and there is room here for debate, we *do* receive an outline of a general answer to this question. I suspect, though, that moral and political philosophers will still be generally unsatisfied with the answer, and may want more. I understand entirely; I still want a pony.

Chapter 1, and Part I: Location

One important contribution *The Stag Hunt* makes to the social contract literature[3] is its refocusing of the discussion upon the Stag Hunt, rather than the Prisoner's Dilemma. The Stag Hunt takes its name from Rousseau's story in *A Discourse on Inequality* (I shan't describe the game here) and provides a better reformulation of the problem of the social contract for one crucial reason: although the Prisoner's Dilemma illustrates the conflict between individual optimality and collective optimality, All Cooperate is not an equilibrium. Since social contracts are — at least apparently — stable, it seems odd to begin with a game theoretic model that makes a stable social contract rationally impossible. The Stag Hunt better models the problem of the social contract because the "all cooperate" outcome (All Hunt Stag) is an

[2] D'Arms (2000) raises a similar point.
[3] Although it is not the first, as Skyrms notes.

equilibrium as well as the "all defect" outcome (All Hunt Hare). It does not assume the social contract to be rationally impossible at the outset.

This is why Skyrms claims that the problem of social contract formation becomes a problem about equilibrium selection or, alternatively, equilibrium transition. If a society is in a nonequilibrium state (or, heaven forbid, the All Hunt Hare equilibrium, the Stag Hunt representation of the state of nature), how can that society settle upon a social contract — i.e., how can it move into the All Hunt Stag equilibrium? Here we find another argument against using the replicator dynamics to model social evolution: according to those dynamics, the social contract cannot form through gradual means. If society happens to be in the state of nature (All Hunt Hare), under the replicator dynamics any minor deviation (say, any mildly revisionary proposal to move society more closely towards a social contract) will ultimately be discarded, and society will return to the state of nature. It doesn't make sense, argues Skyrms, to model social evolution using dynamics which rule out, *a priori*, the gradual formation of the social contract. As such, he suggests, we ought to consider alternative dynamics.

The first alternative dynamic considered, and the focus of Part I, are local interaction models. Suppose people are positioned on the squares of a checkerboard and play a game by interacting with their eight nearest neighbours. People receive a total payoff equalling the sum of the payoffs from each pairwise interaction. After everyone has interacted, players engage in a round of strategic learning and modify their strategies. According to the learning rule known as *Imitate-the-best*, players compare how well they did to their nearest neighbours, and adopt the strategy which received the highest payoff, provided that the highest payoff exceeds the player's current payoff.[4] In contrast, the learning rule known as *Best Response* assumes that one's neighbours will continue to follow the same strategies in the next generation that they follow in the current generation, and selects the strategy which provides the highest expected payoff. Should a tie between several strategies occur, it is broken with a coin flip.

Given this brief sketch, and without discussing the core results, it should be clear how *The Stag Hunt* leaves a great deal of interesting ground unexplored. For example, the local interaction models considered in chapters 2 and 3 involve only square lattices and rings (at the very end of chapter 3, Skyrms briefly considers one alternate structure). One may wonder, do the convergence results depend upon the topology of the local interaction model, or the size of the population?[5] What happens if more sophisticated learning rules are employed? What happens if more than one type of learning rule is present in the population? And so the questions multiply. Yet, to my mind, this is a good thing: in its discussion of local interaction models, *The Stag Hunt* has identified a fertile area of research, and much work remains to be done.[6]

Part II: Signals

[4] *Imitate-the-best* is a form of dissatisfaction-driving learning.
[5] Yes, they do. Unfortunately, the margin is too narrow to contain a demonstration of this fact.
[6] Do note that I might harbour some bias on this point.

Although a natural unity links Part I and Part III, the middle section of *The Stag Hunt* concentrates on the evolution of inference (chapter 4) and cheap talk (chapter 5), and hence functions more as a conceptual detour than a bridge. The discussion of the evolution of inference revisits Lewis' well-known signalling game from *Convention*, and will be familiar to readers of *Evolution of the Social Contract* from the treatment of the evolution of meaning. The chapter on cheap talk illustrates how, somewhat counter-intuitively, meaningless babble can radically transform the basins of attraction of equilibria in games like the Stag Hunt, in addition to creating entirely new equilibria.

While it is undoubtedly important to show how our inferential capacity could have been produced by evolution as a solution to adaptive problems, the chapter on the evolution of inference is the least developed part of the book and does not actually provide a model of how this might happen. Here we find a conceptual sketch of how one might go about constructing such a model. It would be interesting to see this done, primarily to determine whether the envisioned path from proto-truth functions to full-fledged inference is as easily travelled as intimated.

By contrast, the treatment of Stag Hunt and bargaining games with cheap talk is much more thoroughly fleshed out. Cheap talk — the exchange of costless, meaningless signals over the course of play — has been thought to be relatively ineffective in influencing equilibrium outcomes, even by game theorists such as Robert Aumann. To the extent that cheap talk has been considered effective, it has been thought to primarily destabilise equilibria. Skyrms gives the example of a "secret handshake" used to identify cooperators in the Prisoner's Dilemma. If talk is cheap, mutants which extend the secret handshake yet defect can invade the population.

However, Skyrms shows that both of these views underestimate the power of cheap talk. Cheap talk can create new equilibria, and also can change the size of the basins of attraction in the Stag Hunt, such that All Hunt Stag, rather than All Hunt Hare, becomes the state with the largest basin of attraction. Although this latter claim may seem unintuitive, it makes sense upon reflection. It's wrong to think that, with respect to sending signals, the only two possibilities are that the signals be meaningful (in the sense of a Lewis signalling system) or meaningless. If signals are correlated with hunting stag or hare, and strategies allow one to condition one's response upon the signal received, then the *only* time when receipt of a signal fails to convey information is in the special case where all combinations of signals and responses occur with equal frequency. In all other cases, signals — even if they are "meaningless" in the sense of a Lewis-style signalling system — become correlated with response-types, and convey information. The full effects of cheap talk, though, remain an open question.

Part III: Assocation

Part III addresses the second conjunct of the title — the evolution of social structure. This work, performed in collaboration with Robin Pemantle, considers how social structures, like those appealed to in the local interaction models of Part I, come about. The model is as follows: given a population, each person assigns a numeric weight to every other player in the population. These weights, if all nonnegative (and with a sum strictly greater than zero), can be converted into interaction probabilities by

straightforward normalisation; if the weights are negative, interaction probabilities can be obtained with a minor amount of fiddling. People choose to play a game with others according to these interaction probabilities. After each interaction, some of the players (either the initiator, or both, depending on the model) receives a payoff. These payoffs are used by the players to adjust the weights they assign to each other. As the weights increase, and weight tends to concentrate on some players rather than others, interaction structures emerge. The final chapter considers what happens when strategic dynamics (i.e., learning by *Best Response*, *Imitate-the-best*, or other learning rules) are combined with structural dynamics.

Much can be said about the technical content of the final two chapters, which cover a great deal of original and interesting material. Since this would be of little benefit, though, to readers unfamiliar with the book, I shall concentrate on a few overarching philosophical questions which, to my mind, remain unanswered.

First, regarding the caveat mentioned earlier: although we do get, in Part I, a story about how social structure influences the evolution of justice and cooperation, and we do get, in Part III, a story of how social structure might emerge, how well do the two stories line up? For starters, the interaction structures considered in part I are highly regular: two-dimensional square lattices (or toruses) or one-dimensional lines (or rings). The social structures which are shown to evolve in Part III lack any such topological regularity. The emergence of justice and cooperation according to the local interaction models of Part I is unlikely, then, if the model of social structure *formation* is that of Part III. Moreover, the interaction model of Part I is strictly deterministic, with the interactions being hard-wired and always taking place in each generation. How might a probabilistic model of the evolution of social structure like that of Part III be modified to give rise to deterministic interaction structures within a reasonable (i.e., human) time-frame, rather than in the infinite limit?[7]

Second, a different concern regarding the type of explanation offered. The penultimate section of chapter 3 ends with the statement, "[t]he *structure* of local interaction makes a difference" (Skyrms, 2004, pg. 42). This sentiment ramifies over the course of the book. We find that *signals* make a difference. The *co-evolution of strategic and structural evolution* makes a difference. The *relative rate* of the co-evolution of strategy and structure makes a difference. How are we to assimilate all of these things-which-make-a-difference into a coherent explanatory account of cooperation, collective action, and the problem of the social contract? This remains unclear.

Consider, by way of comparison, the following crude way of putting the meta-narrative of *Evolution of the Social Contract.* The replicator dynamics are a model of cultural evolution. Strategies (i.e., behaviours) which have very large basins of attraction are highly likely to evolve, if the initial conditions were selected at random. Many behaviours of interest, such as fair division in resource allocation problems, retribution in the ultimate game, and coordination upon meaningful signalling

[7] Notice that the model of social network formation given in chapter 6 might not even converge to a deterministic interaction structure even in the limit. In cases where stars form (Skyrms, 2004, pg. 90), the interaction probabilities can converge in the limit to a person splitting his time between two individuals.

systems, turn out to evolve — Lo and behold! often with surprising frequency — under the replicator dynamics. And, thus, we have an explanation for how certain core features of society could have come about through the operation of a blind, dumb, gradual process of social evolution.

It's not at all clear that *The Stag Hunt* permits anything like a similarly crude meta-narrative to be formulated. And perhaps this is a good thing. (Although I'm no postmodernist, I'm suspicious of grand meta-narratives, myself.) However, this also means that it's difficult to see exactly how the explanation works. If everything makes a difference, then all of the various parameters which factor into the evolution need to be set *just so* in order for us to find the behaviour we find in society. Yet if only a very small set of possible paths leads from the primordial state to our current social state, what have we discovered? We *already know* that at least one such path exists, namely, whatever path we followed to get us here. If too many things make a difference, the project, then, risks becoming a game-theoretic archaeology for the human sciences (no — not in the sense of Foucault!) instead a descriptive or predictive science facilitating substantive claims about future social states. At least that's a worry.

No work in philosophy shorter than the autobiography of Tristram Shandy can hope to answer all questions, and good work in philosophy raises more questions than it answers. As a snapshot of the recent state-of-the-art in evolutionary game theory, and the philosophical applications thereof, *The Stag Hunt* manages to surpass its predecessor in both scope and content. If only George Lucas had been so successful.

References

D'Arms, Justin. 1996. Sex, Fairness, and the Theory of Games. *Journal of Philosophy* 93(12): 615–627.

D'Arms, Justin. 2000. When Evolutionary Game Theory Explains Morality, What Does It Explain? *Journal of Consciousness Studies* 7: 296–299.

D'Arms, Justin, Robert Batterman, and Krzyzstof Górny. 1998. Game Theoretic Explanations and the Evolution of Justice. *Philosophy of Science* 65: 76–102.

Kitcher, Philip. 1999. Games Social Animals Play: Commentary on Brian Skyrms' *Evolution of the Social Contract*. *Philosophy and Phenomenological Research* 59(1): 221–228.

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

—. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.