Presidential Address
Pacific Division of the American Philosophical Association
March 2001

The Stag Hunt

Brian Skyrms
U. C. Irvine

I: The Stag Hunt

The Stag Hunt is a story that became a game. The game is a prototype of the social contract. The story is briefly told by Rousseau, in *A Discourse on Inequality*:

> If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple..."

Rousseau's story of the hunt leaves many questions open. What are the values of a hare and of an individual's share of the deer given a successful hunt? What is the probability that the hunt will be successful if all participants remain faithful to the hunt? Might two deer hunters decide to chase the hare?

Let us suppose that the hunters each have just the choice of hunting hare or hunting deer. The chances of getting a hare are independent of what others do. There is no chance of bagging a deer by oneself, but the chances of a successful deer hunt go up sharply with the number of hunters. A deer is much more valuable than a hare. Then we have the kind of interaction that that is now generally known as the Stag Hunt.

Once you have formed this abstract representation of the Stag Hunt game, you can see Stag Hunts in many places. David Hume also has the Stag Hunt. His most famous illustration of a convention has the structure of a two-person Stag Hunt game:

> Two men who pull at the oars of a boat, do it by an agreement or convention, tho' they have never given promises to each other. ...

Both men can either row or not row. If both row, they get the outcome that is best for each - just as in Rousseau's example, when both hunt the stag. If one decides not to row then it makes no difference if the other does or not - they don't get anywhere. The worst outcome for you is if you row and the other doesn't, for then you lose your effort for nothing, just as the worst outcome for you in the Stag Hunt is if you hunt stag by yourself.

We meet the Stag Hunt again in the meadow-draining problem of Hume's *Treatise*:

> Two neighbors may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each others mind, and each may perceive that the immediate consequence of failing in his part is the abandoning of the whole project. But 'tis difficult, and indeed impossible, that a thousand persons shou'd agree in any such action ...

where Hume observes that achieving cooperation in a many-person Stag Hunt is more difficult than achieving cooperation in a two-person stag hunt.[1]

The Stag Hunt does not have the same melodramatic quality as the Prisoner's Dilemma. It raises its own set of issues, which are at least as worthy of serious consideration. Let us focus, for the moment, on a two-person Stag Hunt for comparison to the familiar two-person Prisoner's Dilemma.

If two people cooperate in Prisoner's Dilemma, each is choosing less rather than more. In Prisoner's Dilemma, there is a conflict between individual rationality and mutual benefit.

In the Stag Hunt, what is rational for one player to choose depends on his beliefs about what the other will choose. Both stag hunting and hare hunting are *equilibria*. That is just to say that it is best to hunt stag if the other player hunts stag and it is best to hunt hare if the other player hunts hare. A player who chooses to hunt stag takes a risk that the other will choose not to cooperate in the Stag Hunt. A player who chooses to hunt hare runs no such risk, since his payoff does not depend on the choice of action of the other player, but he foregoes the potential payoff of a successful stag hunt. Here rational players are pulled in one direction by considerations of mutual benefit and in the other by considerations of personal risk.

Suppose that hunting hare has an expected payoff of 3, no matter what the other does. Hunting stag with another has an expected payoff of 4. Hunting Stag alone is doomed to failure and has a payoff of zero. It is clear that a pessimist, who always

expects the worst, would hunt hare. But it is also true with these payoffs that a cautious player, who was so uncertain that he thought the other player was as likely to do one thing as another, would also hunt hare.[2] That is not to say that rational players could not coordinate on the stag hunt equilibrium that gives them both better payoff, but it is to say that they need a measure of trust to do so.

I told the story so that the payoff of hunting hare is absolutely independent of how others act. We could vary this slightly without affecting the underlying theme. Perhaps if you hunt hare, it is even better for you if the other hunts stag for you avoid competition for the hare. If the effect is small we still have an interaction that is much like the Stag Hunt. It displays the same tension between risk and mutual benefit. It raises the same question of trust. This small variation on the Stag Hunt is sometimes also called a Stag Hunt [3] and we will follow this more inclusive usage here.

Compared to the Prisoner's Dilemma, the Stag Hunt has received relatively little discussion in contemporary social philosophy – although there are some notable exceptions.[4] But I think that the Stag Hunt should be a focal point for social contract theory.

The two mentioned games, Prisoner's Dilemma and the Stag Hunt, are not unrelated. Considerations raised by both Hobbes and Hume can show that a seeming Prisoner's Dilemma is really a Stag Hunt. Suppose that Prisoner's Dilemma is repeated. Then your actions on one play may affect your partner's actions on other plays, and

considerations of reputation may assume an importance that they cannot have if there is no repetition. Such considerations form the substance of Hobbes' reply to the Foole. Hobbes does not believe that the Foole has made a mistake concerning the nature of rational decision. Rather, he accuses the Foole of a shortsighted mis-specification of the relevant game:

> He, therefore, that breaketh his Covenant, and consequently declareth that he think that he may with reason do so, cannot be received into any society that unite themselves for Peace and Defense, but by the error of them that receive him.[5]

According to Hobbes, the Foole's mistake is to ignore the future.

David Hume invokes the same considerations in a more general setting:

> Hence I learn to do a service to another, without bearing him any real kindness; because I foresee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me and with others. [6]

Hobbes and Hume are invoking the *shadow of the future.*

How can we analyze the shadow of the future? We can use the theory of indefinitely repeated games. Suppose that the probability that the Prisoner's Dilemma is will be repeated another time is constant. In the repeated game, the Foole has the strategy *Always Defect.* Hobbes argues that if someone defects, others will never cooperate with

him. Those who initially cooperate, but who retaliate as Hobbes suggests against defectors, have a *Trigger* strategy.

If we suppose that *Always Defect* and *Trigger* are the only strategies available in the repeated game and that the probability of another trial is .6, then the Shadow of the Future transforms the two-person Prisoner's Dilemma:

	Cooperate	Defect
Cooperate	3	1
Defect	4	2

into the two-person Stag Hunt: [7]

	Trigger	All Defect
Trigger	7.5	4
All Defect	7	5

This is an exact version of the informal arguments of Hume and Hobbes.[8]

But for the argument to be effective against a fool, he must believe that the others with whom he interacts are not fools. Those who play it safe will choose *Always Defect.*

Rawls' maximin player is Hobbes' Foole.[9] The Shadow of the Future has not solved the problem of cooperation in the Prisoner's Dilemma; it has transformed it into the problem of cooperation in the Stag Hunt.

In a larger sense, the whole problem of adopting or modifying the social contract for mutual benefit can be seen as a Stag Hunt. For a social contract theory to make sense, the state of nature must be an equilibrium. Otherwise there would not be the problem of transcending it. And the state where the social contract has been adopted must also be an equilibrium. Otherwise, the social contract would not be viable. Suppose that you can either *devote energy to instituting the new social contract* or not. If everyone takes the first course the social contract equilibrium is achieved; if everyone takes the second course the state of nature equilibrium results. But the second course carries no risk, while the first does. This is all quite nicely illustrated in miniature by the meadow-draining problem of Hume.

The problem of reforming the social contract has the same structure. Here, following Binmore, we can then take the relevant "state of nature" to be the *status quo,* and the relevant social contract to be the projected reform. The problem of instituting, or improving, the social contract can be thought of as the problem of moving from riskless Hunt Hare equilibrium to the risky but rewarding Stag Hunt equilibrium.

II: Game Dynamics

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 that Hobbes and Hume had available, but still less than we need for fully adequate answers. The news from the frontiers of game theory is rather pessimistic about the transition from hare hunting to stag hunting.

The modern embodiment of Hobbes' approach is rational choice based game theory. It tells us that what a rational player will do in the Stag Hunt depends on what he thinks the other will do. It agrees with Hume's contention that a thousand person stag-hunt would be more difficult to achieve than a two-person stag hunt, because –assuming that everyone must cooperate for a successful outcome to the hunt – the problem of trust is multiplied. But if we ask how people can get from a Hare Hunt equilibrium to a Stag hunt equilibrium, it does not have much to offer. From the standpoint of rational choice, for the Hare Hunters to decide to be Stag Hunters, each must *change her beliefs* about what the other will do. But rational choice based game theory as usually conceived, has nothing to say about how or why such a change of mind might take place.

Let us turn to the tradition of Hume. Hume emphasized that social norms can evolve slowly:

> "Nor is the rule regarding the stability of possession the less derived from human conventions, that it arises gradually, and acquires force by a slow progression..."

We can reframe our problem in terms of the most thoroughly studied model of cultural evolution, the replicator dynamics. If we ask in this framework, how one can get from the Hunt Hare equilibrium to the Hunt Stag equilibrium, the answer is that you can't! In the vicinity of the state where all Hunt Hare, Hunting Hare has the greatest payoff. If you are close to it, the dynamics carries you back to it. This reasoning holds good over a large class of adaptive dynamics. The transition from non-cooperation to cooperation seems impossible.

Perhaps the restriction to deterministic dynamics is the problem. We may just need to add some chance variation. We could add some chance shocks to the replicator dynamics[10] or look at a finite population where people have some chance of doing the wrong thing, or just experimenting to see what will happen.[11] If we wait long enough, chance variation will bounce the population out of hare hunting and into stag hunting.

But in the same way, chance variation can bounce the population out of stag hunting into hare hunting. Can we say anything more than that the population bounces

between these two states? Yes, we know how to analyze this system[12] but the news is not good. When the chance variation is small, the population spends almost all its time in a state where everyone hunts hare.[13] It seems that all we have achieved so far is to show how the social contract might degenerate spontaneously into the state of nature.

Social contracts do sometime spontaneously dissolve. But social contracts also form. And there is experimental evidence that people will hunt stag even when it is a risk to do so.[14] This suggests the need for a richer theory.

III: Local Interaction

The foregoing discussion proceeded in terms of models designed for random encounters in large populations. But cooperation in the Stag Hunt may well have originated in small populations with non-random encounters. How should we think about this setting?

Perhaps, instead of interacting at random with other members of the population, we interact with our neighbors. Can local interaction make a difference? We know that it can, from investigations the dynamics of other games played with neighbors. Philosophers have contributed these developments and in this regard I would like to mention Jason Alexander, Peter Danielson, Patrick Grim, Bill Harms, Rainer Hegselmann, Gary Mar, and Elliott Sober. In some cases local interactions make a spectacular difference in the outcome of the evolutionary (or learning) process.

Does local interaction make a difference in the Stag Hunt? Can it explain the institution of the social contract? The news gets worse. Glenn Ellison (1993) investigates the dynamics of the Stag Hunt played with neighbors, where the players are arranged on a circle. He finds limiting behavior not much different than that in the large population with random encounters. With a small chance of error, the population spends most of its time hunting hare. The difference in the dynamics of the two cases is that given local interaction, the population approaches its long run behavior much more rapidly. The moral for us, if any, is that in small groups with local interaction the degeneration of the social contract into the state of nature can occur with great rapidity.[15]

There is, however, a small glimmer of light from the following fable.[16] There is a central figure in the group, who interacts (pairwise) with all others and with whom they interact - you might think of him as the Boatman in honor of Hume or the Huntsman in honor of Rousseau.

(figure here)

Every round, players revise their strategies by choosing the best response to their partners' prior play, with the exception that the Huntsman revises his strategy much less often. We add a small chance of spontaneously changing strategy for each player. If everyone hunts stag and the Huntsman spontaneously switches to hunting hare, in two

rounds probably everyone will hunt hare. But conversely, if everyone is hunting hare and the Huntsman spontaneously switches to hunting stag, in two rounds probably everyone will hunt stag. In the long run the population spends approximately half its time hunting hare and half its time hunting stag. The story does not speak to all our concerns, but at least it shows that the social contract need not have negligible probability in the long run and that the *structure* of local interaction can make a difference[17].

IV: Dynamics of Interaction Structure

Still, it is hard not to feel that there must be something fundamental that is missing from our analysis. I would like to suggest that what is missing is an account of the evolution of the structure of interactions. Game theory takes the interaction structure as fixed. But in real life individuals adjust with whom they interact on the basis of past experience. This as a fundamental aspect of social behavior that is completely absent from the theory of games.

Let me show you how the analysis of the Stag Hunt is changed if individuals can learn with whom to interact. What follows is the result of joint work with Robin Pemantle.[18] Suppose that we have a small group of agents, some disposed to hunt stag and some disposed to hunt hare. They start out interacting at random, but when agents interact they are reinforced for interaction with the same agent by the payoff that they receive from the interaction. Reinforcement modifies the probability that one agent will choose to interact with another.

How do these interaction probabilities evolve? It can be shown, both by simulation and analytically, that stag hunters learn to interact with other stag hunters. This, perhaps, should come as no great surprise. It is not quite so obvious that hare hunters will end up interacting with other hare hunters. Hare hunters do not care with whom they interact. Nevertheless it is so.[19] Learning dynamics leads to a structure of interaction probabilities quite different from that of a random pairing model. In this environment, stag hunters prosper.

Now we can add the further consideration of players revising their strategies. Suppose that once in a while a player looks around the little group, sees who is doing best, and imitates that strategy. If interaction structure were fixed at random pairing, we would be back were we started and the most likely outcome would be that everyone would end up hunting hare. But if structure is fluid and the learning dynamics for structure is fast relative to the strategy revision dynamics, stag hunters will find each other and then imitation will slowly convert the hare hunters to stag hunters. This conclusion is robust to the addition of a little chance. Here, we finally have a model that can explain the institution of a modest social contract.

In between the extremes, the limiting outcome in a small group may be either all hare hunters or all stag hunters. Which outcome one gets depends somewhat on the vicissitudes of chance in the early stages of the evolution of the group. But it is also

strongly influenced by the relative speeds of structure and strategy dynamics. Rapid structural adaptation favors the stag hunters. Our social contract may form in some circumstances but not in others.

We can consider different strategy revision dynamics. Suppose, as before, that structure dynamics is fast relative to strategy revision, but that the individuals never look around the group for successful models, but rather choose the best response to the strategy that they have encountered on the last play. Then the ultimate outcome will be that the group is divided into two stable classes, the stag hunters and the hare hunters, which never interact with one another. A different kind of strategy revision leads to a different social contract. There are other interesting possibilities to explore.

Conclusion

We should pay more attention to the Stag Hunt. There is a lot to think about. Real stag hunts are complex interactions between more than two people. So are the social contracts - large and small - that we have used the Stag Hunt to represent. In our analysis of two-person Stag Hunts we finally focused on the coevolution of strategy and interaction structure. I believe that this is a key to the larger question of the emergence of social structure.

References:

Alexander, Jason and Brian Skyrms (1999) "Bargaining with Neighbors: Is Justice Contagious?" *Journal of Philosophy* 96 588-598.

Axelrod, Robert (1984) *The Evolution of Cooperation* New York: Basic Books.

Binmore, Ken (1993) *Playing Fair: Game Theory and the Social Contract I.* Cambridge: MIT Press.

Binmore, Ken (1998) *Just Playing: Game Theory and the Social Contract II.* Cambridge:MIT Press.

Danielson, Peter (1992) *Artificial Morality* London:Routledge.

Ellison, Glenn (1993) "Learning, Local Interaction and Coordination" *Econometrica* 61, 1047-1071.

Epstein, Joshua and Axtell, Robert (1996) *Growing Artificial Societies: Social Science From the Bottom Up.* Cambridge:MIT Press.

Foster, Dean P, and H. Peyton Young (1990) "Stochastic Evolutionary Game Dynamics" *Theoretical Population Biology* 28:219-32.

Grim, Patrick, Gary Mar, Paul St. Denis (1998) *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling* Cambridge: MIT Press.

Hampton, Jean. (1997) *Hobbes and the Social Contract Tradition.* N.Y.: Cambridge.

Harms, William (2000) "The Evolution of Cooperation in Hostile Environments" *Journal of Consciousness Studies* 7, 308-313.

Harsanyi, J. and R. Selten (1988) *A General Theory of Equilibrium Selection in Games* Cambridge: MIT Press.

Hegselmann, Rainer (1996)"Social Dilemmas in Lineland and Flatland" In Wim B. G. Liebrand and David. Messick (eds) *Frontiers in Social Dilemmas Research.* Berlin: Springer, 337-362.

Hobbes, Thomas. (1668) *Leviathan.* ed. & tr. E. Curley (1994) Indianapolis: Hackett.

Hume, David. (1739) *A Treatise of Human Nature.* ed. Selby-Bigge (1978) Oxford: Clarendon.

Jackson, Matthew and Alison Watts (working paper) "On the Formation of Interaction Networks in Social Coordination Games".

Jervis, Robert (1978) "Cooperation under the Security Dilemma" *World Politics* 30: 167-214.

Jiborn, Magnus. (1999) *Voluntary Coercion.* Lund: Lund University.

Kandori, Michihiro, George Mailath, and Rafael Rob (1993) "Learning, Mutation and Long-Run Equilibria in Games" *Econometrica* 61:29-56.

Luce, Duncan and Howard Raiffa (1957) *Games and Decisions* New York:Wiley.

Rankin, Frederick W., John B. Van Huyck and Raymond Battalio (2000) "Strategic Similarity and Emergent Conventions: Evidence from Similar Stag Hunt Games" *Games and Economic Behavior* 32, 315-337.

Sen, Amartya (1967) "Isolation, Assurance, and the Social Rate of Discount" *Quarterly Journal of Economics*81:112-124.

Skyrms, Brian (1998) "The Shadow of the Future" In *Rational Commitment and Social Justice: Essays for Gregory Kavka.* Ed. Jules Coleman and Christopher Morris. Cambridge: Cambridge University Press, 12-22.

Skyrms, Brian and Pemantle, Robin (2000) "A Dynamic Model of Social Network Formation" *Proceedings of the National Academy of Sciences of the USA* 97, 9340-9346.

Skyrms, Brian and Vanderschraaf, Peter (1997) "Game Theory" in *The Handbook of Practical Logic.* Ed. Philippe Smets. Dordrecht:Kluwer.

Sober, Elliott and David Sloan Wilson (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior* Cambridge: Harvard University Press.

Vanderschraaf, Peter (1998) "The Informal Game Theory in Hume's Account of Convention" *Economics and Philosophy* 14:251-257.

Vanderschraaf, Peter (2001) *Learning and Coordination: Inductive Deliberation, Equilibrium and Convention.* London: Routledge.

Van Huyck, J. B., R. C. Battalio and R. O. Beil (1990) " Tacit Coordination Games, Strategic Uncertainty and Coordination Failure" *The American Economic Review* 80:234-248.

Young, H. Peyton (1998) *Individual Strategy and Social Structure* Princeton: Princeton University Press.

NOTES:

[1] For evidence that this is true in laboratory experiments, see Van Huyck, Battalio and Beil (1990).

[2] Hunting hare is said to be the *risk dominant* equilibrium.

[3] Sometimes it is called an Assurance Game, following Sen (1967).

[4] Ed Curley, Jean Hampton, Magnus Jiborn, and Peter Vanderschraaf.

[5] Hobbes, *Leviathan*, xv,5, 205.

[6] Hume, *Treatise*, 521.

[7] If the probability of repetition is less than .5, the repeated game is still a Prisoner's Dilemma. If the probability of repetition is high enough, the stag hunting equilibrium becomes risk dominant.

[8] And of Curley's remarks in his introduction to his edition of the Leviathan, p. xxviii.

[9] Rawls recommends that agents choose a social contract according to the maximin principle which would have each agent play it safe by maximizing her minimum gain. If agents followed this advice here, they would choose *Always Defect.*

[10] Foster and Young (1990).

[11] Kandori, Mailath and Rob (1993).

[12] Kandori, Mailath and Rob (1993), Young (1998).

[13] In the risk dominant equilibrium.

[14] See Van Huyck, Battalio and Beil (1990).

[15] Peyton Young (1998) finds the same story true much more generally for local interaction on structures different from the circle.

[16] adapted from an example of Jackson and Watts.

[17] If the huntsman were not patient, then the population would spend a smaller proportion of its time hunting stag, but that proportion would still not be negligible and it would still be true that structure makes a difference. That is the form of the example in Jackson and Watts.

[18] for details and analysis, see Skyrms and Pemantle (2000).

[19] Skyrms and Pemantle (2000).