Synthese (2008) 160:21–25
DOI 10.1007/s11229-006-9075-3

ORIGINAL PAPER

Trust, risk, and the social contract

Brian Skyrms

Received: 17 May 2006 / Accepted: 12 June 2006 / Published online: 16 September 2006

Abstract The problem of trust is discussed in terms of David Hume's meadow-draining example. This is analyzed in terms of rational choice, evolutionary game theory and a dynamic model of social network formation. The kind of explanation that postulates an innate predisposition to trust is seen to be unnecessary when social network dynamics is taken into account.

Keywords Trust · Evolution · Stag hunt game · Risk dominance · Local interaction · Dynamic network

> *Two neighbors may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each other's 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.*
>
> — David Hume *A Treatise of Human Nature*, Bk. III, Pt.II, Sec. VII.

Social contracts, great and small, depend on trust. Hume's two neighbors are able to sustain an implicit contract even though the failure of one to perform his part causes the cooperative enterprise to fail. If we view Hume's two neighbors through the lens of game theory, the simplest representation is that they are playing a two-person non-zero sum game with the structure of a Stag Hunt. (The name comes from a story in Rousseau with a similar moral.) There are two equilibria: *both cooperate; neither cooperates*. It is not an equilibrium for one to cooperate and the other not, because in such a case each would have an incentive to switch. The equilibrium where both

B. Skyrms (✉)
Department of Logic and Philosophy of Science and Department of Economics,
School of Social Sciences, University of California,
Irvine, CA 92697-5100, USA
e-mail: bskyrms@uci.edu

cooperate is the one in which they are both better off—it is said to be *payoff dominant*—, but each runs the risk that the other may not do his part. The equilibrium where neither cooperates is one in which neither player runs a risk, since the outcome is the same no matter what the other player does. (In Rousseau's story a successful Stag Hunt requires cooperation, while Hare Hunting is a solitary occupation.) Mutual benefit is pitted against the risk that the other (or others) may not honor the implicit contract. This is the prototypical problem of the social contract.

How bad is the problem? That depends on two things: first—as Hume points out—the beliefs about what the other player does; second—the magnitudes of the benefit of cooperation and of the risk of being abandoned by one's potential partner.

Suppose that the effort in doing one's part is E, the benefit to each of a drained meadow is B, and the value of the status quo is D. Then the meadow draining game has payoffs (of form row's payoff, column's payoff):

	Work to drain	Don't work
Work to drain	B–E, B–E	D–E, D
Don't work	D, D–E	D, D

If the project is worth the effort, B–E > D, this has the structure of a Stag Hunt game. If I am sure that I can trust my partner to cooperate, I will too.

But suppose that my partner is as likely to work as not. Then my expectation of working is (1/2)(B–E) + (1/2)(D–E) and my expectation for not working is D. Here I am indifferent between working or not if B–D = 2E. A little better benefit or a little less required effort tips the balance in favor of cooperation; a little less benefit or a little increase in the work load tips the balance the other way. In any case, the equilibrium that comes out ahead in this calculation—with equal probabilities for the other cooperating or not—is called the *risk dominant* equilibrium. In easy cases, the risk dominant equilibrium coincides with the payoff dominant equilibrium; in hard cases risk dominance and payoff dominance pull in the opposite direction. Let us direct our attention to hard cases.

For example, suppose that B = 7, E = 3, and D = 3. Then we have a Stag Hunt game in which draining the meadow is the payoff-dominant equilibrium and no one working is the risk dominant equilibrium:

	Work to drain (Stag)	Don't work (Hare)
Work to drain (Stag)	4, 4	0, 3
Don't work (Hare)	3, 0	3, 3

If people have a moderate degree of trust, probability > 0.75, cooperation ensues. But where does the trust come from? One can say that trust is based on prior experience, but that only pushes the question back. If there had not been prior trust, then there would not have been the kind of prior experience that supports trust, but rather the kind that supports distrust.

Perhaps, someone might say, we are evolved to be the kind of species with a predilection for cooperation—with some initial but defeasible predilection for trust in cooperative enterprises built into our nature. The same problem now emerges in even grander evolutionary terms. Should we expect evolutionary dynamics to respect payoff dominance when it conflicts with risk dominance?

The answer usually delivered up by contemporary evolutionary game theory is "No." In the long run, one should expect to see the risk-dominant equilibrium almost all the time. This is the central result of Kandori, Malaith, and Rob (1993)[1] (See also Young, 1998). The idea is that there is an underlying dynamics of differential reproduction perturbed by some very small probability of mutation. Sooner or later—perhaps a lot later—a lot of mutations move the (finite) population from the basin of attraction of one equilibrium to that of the other, and then the underlying differential reproduction quickly takes it to that equilibrium. Sooner or later, a lot of mutations take the population to the basin of attraction of the other equilibrium, and differential reproduction takes the population to the second equilibrium. If mutations are unlikely and the probability of mutations is independent across individuals, the probability of mutations taking you from the basin of attraction of the cooperative equilibrium to that of the non-cooperative one is much larger than the probability of mutations taking you in the opposite direction. Therefore the population spends most of its time not cooperating.

On the face of it, the reasoning seems remarkably robust. The underlying dynamics need not be differential reproduction; it could be anything with the same basin of attraction—anything that moves in the direction of greatest payoff. The stochastic shocks to the system might be interpreted as experimentation, or as some kind of exogenous noise with a similar probability structure (See Foster & Young, 1990). The only problem seems to be that the expected waiting time for all those mutations, experiments or whatever to happen all at once may be astronomical. It is almost like a theory of evolution driven by the probability of miracles.

Ellison (1993) provides an answer from perhaps an unexpected quarter. He considers a local interaction model where individuals are located on a circle, and interact with their immediate neighbors. It is still true that with small probability of mutations, the population spends almost all its time not cooperating, but here the expected waiting time is short. If mutation delivers two contiguous defectors in a population of cooperators, the defectors will rapidly spread and take over the population. In a later paper Ellison (2000) shows that defecting also takes over quickly—though not quite so quickly—in a two dimensional local interaction model where individuals play with their neighbors to the North, East, South and West.[2] Young (1998) gives further support to the selection of risk-dominant equilibria in games played on local interaction structures. We do not yet have a good model for the emergence of sufficient trust to allow the selection of the mutually beneficial equilibrium over the risk-dominant equilibrium in a coordination game.

Why not? It is because the foregoing analysis holds the interaction structure *fixed*. We should consider the possibility that individuals learn *with whom* to interact as well as *how* to act. Interaction structures will then be dynamic entities, and strategy and

[1] But compare Robson and Vega-Redondo (1996). They point out that Kandori–Mailath–Rob assume a special type of matching (round-robin matching) to get their results, and show how the correlations generated by random matching in a finite population can lead to the payoff dominant equilibrium in the Stag Hunt game. Later in this paper we consider correlations in matching that arise not by chance, but by the choices of the agents involved.

[2] However, in contrast to the original Kandori–Mailath–Rob and Foster–Young models, the underlying deterministic dynamics here *does* matter. Ellison uses a best-response dynamics. If you switch to an imitate-the-best dynamics, for example, you get quite different results (See Skyrms, 2004, Chap. 3).

structure will co-evolve. Such a model is proposed in Skyrms and Pemantle (2000), and pursued in Pemantle and Skyrms (2004a, b), Skyrms (2004) and Skyrms and Pemantle (forthcoming). In this model a small group of individuals start interacting at random. Interactions are modeled as games, and agent's types determine their strategies in the games. The interaction structure evolves by reinforcement learning, with the magnitude of reinforcements being the payoffs from interactions. If you have a good payoff from an interaction with someone, you are more likely to interact with them again.

Consider our Stag Hunt game in this context, where Stag hunting (cooperating) is payoff dominant and Hare hunting (not-cooperating) is risk-dominant. Using a standard model of reinforcement learning (Roth & Erev, 1995), Stag hunters rapidly learn to interact with one another. In our prototypical Stag Hunt, Stag hunters then get a payoff of 4 while Hare hunters get a payoff of 3. Risk-dominance now has no teeth, because a little learning has taken the risk out of Stag hunting.

Since Stag hunters are now doing better than Hare hunters, Hare hunters may eventually notice this and imitate them—or in a biological context Stag hunters can out—reproduce Hare hunters. In either case Stag hunting takes over the population. All that is required is that the interaction structure is sufficiently fluid so that Stag hunters can find each other quickly.

The choice of reinforcement learning as the dynamics of interaction requires little of the individuals involved. They don't need to think strategically about the situation at all, and they don't need to observe others' actions or payoffs. More sophisticated and knowledgeable Stag hunters who are free to associate as they please might find each other *right away*. The model of reinforcement learning is therefore something of a worst case analysis.

The foregoing "Learning-to-Network" model was designed for a small group in which agents can identify one another and keep track of reinforcements associated with an individual. This may not be plausible in large populations—but large populations may be made up of smaller subgroups. Suppose that a large population consists of number of small groups—demes—within which interactions take place and interaction structure is adjusted by learning. Strategies may be adjusted by imitation or may also be adjusted by reinforcement. The learning rates may vary from deme to deme. If learning with whom to interact is fast relative to strategy revision, Stag hunting predominates. In demes where interaction structure is rigid and Stag hunters are repeatedly let down by Hare hunters but strategy revision is fast, people learn to hunt Hare.

Now suppose that people can move from one locality to another—perhaps at some small cost, know a little about the norms of other demes, and can do just a little bit of strategic thinking. Stag hunters stuck in a non-cooperative deme can now move to Stag-hunting demes. Stag hunters associate with one another on a larger scale.[3] This effect alone could guarantee the eventual success of Stag hunting even if within-deme dynamics were only driven by mutation and random interaction (See Oechssler, 1999; Ely, 2002; Dieckmann, 1999). But one would have to wait until good luck got the process started.[4] Learning-to-Network within demes jump-starts cooperation, which can then spread to a larger population through mobility between demes.

[3] Hare hunters have no reason to move, and may be left by themselves in poor Hare hunting villages or gradually be converted to Stag hunting, depending on how far then can see and how they revise strategies (See Skyrms, 2004, Chap. 7).

[4] If set up just right, these models can be very fast, but alternative versions can be very slow.

What did we need to explain the possibility of cooperation in the Stag Hunt game? We did not need to assume that evolution had to somehow build in a disposition for trust. It is enough that agents come equipped with a modest capacity for learning.

References

Alexander, J. M. (2003). Random Boolean networks and evolutionary game theory. *Philosophy of Science, 70*, 1289–1304.

Dieckmann, T. (1999). The evolution of conventions with mobile players. *Journal of Economic Behavior and Organization, 38*, 93–111.

Ellison, G. (1993). Learning, local interaction, and coordination. *Econometrica, 61*, 1047–1071.

Ellison, G. (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Review of Economic Studies, 67*, 17–45.

Ely, J. (2002). Local conventions. *Advances in Theoretical Economics, 2*, http://www.bepress.com/bejte/advances/vol2/iss1/art1

Foster, D., & Young, H. P. (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology, 38*, 219–222.

Kandori, M., Mailath, G., & Rob, R. (1993). Learning, mutation and long run equilibria in games. *Econometrica, 61*, 29–56.

Oechssler, J. (1999). Competition among conventions. *Mathematical and Computational Organization Theory, 5*, 31–44.

Pemantle, R., & Skyrms, B. (2004a). Network formation by reinforcement learning: the long and the medium run. *Mathematical Social Sciences, 48*, 315–327.

Pemantle, R., & Skyrms, B. (2004b). Time to absorption in discounted reinforcement models. *Stochastic Processes and Their Applications, 109*, 1–12.

Robson, A. J., & Vega-Redondo, F. (1996). Efficient equilibrium selection in evolutionary games with random matching. *Journal of Economic Theory, 70*, 65–92.

Roth, A., & Erev, I. (1995). Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, 8*, 164–212.

Rousseau, J. (1984). *A discourse on inequality*. Trans. M. Cranston. New York: Penguin Books.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. New York: Cambridge University Press.

Skyrms, B., & Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the USA, 97*, 9340–9346.

Skyrms, B., & Pemantle, R. (forthcoming) Learning to network. In E. Eells, & J. Fetzer (Eds.), *Probability in science*. Chicago, IL: Open Court Publishing.

Vanderschraaf, P., & Alexander, J. M. (2005) Follow the leader: local interactions with influence neighborhoods. *Philosophy of Science* 72, 86–113.

Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*, 57–84.

Young, H. P. (1998). *Individual strategy and social structure*. Princeton, NJ: Princeton University Press.