DOXASTIC PARADOX AND REPUTATION EFFECTS IN ITERATED GAMES

Robert C. Koons
Philosophy Dept. / Center for Cognitive Science
University of Texas at Austin
koons@sygmund.cgs.utexas.edu

The "chain-store paradox" of Reinhard Selten is one of a number of scenarios involving the finite repetition of a certain kind of sub-game about which a paradoxical conclusion can be derived. In each of these cases, a backward-induction argument is used to prove that it is futile to try to establish a reputation for cooperative or punitive behavior through appropriate action in the early stages of the game, despite the fact that nearly all agree that it is intuitively reasonable to do so. Selten believed that this is paradoxical in the weak sense: a surprising, unexpected result of game theory. In this paper, I argue that it is paradoxical in the strong sense: a logical antinomy of rational belief or subjective probability, analogous to the paradox of the Liar.

I use formal theories of rational belief and of higher-order probability to demonstrate the existence in such games of a Liar-like antinomy. This involves extending Montague's 1963 generalization of Tarski's theorem [Kaplan and Montague 1960, Montague 1963]. It also involves adding some assumptions about the belief-revision dispositions of the players. Then, I apply recently developed solutions to the Liar paradox by Burge, Gaifman, and Barwise & Etchemendy to the situation presented by Selten's game. The result demonstrates and explains the existence of context-determined cognitive "blindspots" afflicting the players of the game. This necessitates the use of interval-valued probability functions in prescribing the mental states of such players.

Game theorists have discovered several scenarios involving the finite repetition of a non-cooperative game which give rise to a certain kind of "paradox". These include: Selten's "chain-store paradox" [Selten 1978], the problem of the finite series of "Prisoner's Dilemma" games [Luce and Raiffa 1957, Hardin 1982] and the controversy over the game-theoretic justifiability of deterrent punishment by known act-utilitarians [Hodgson 1967, Reagan 1980]. In each of these cases, a "backward-induction" argument is used to prove that it is futile to try to establish a reputation for cooperative or punitive behavior through appropriate action in the early stages of the game, despite the fact that nearly all agree that it is intuitively "reasonable" to do so.

1. Selten's Paradox of Reputation

Selten's "chain-store paradox" arose from the attempt by game-theoretical economists to analyze and evaluate the rationality of predatory behavior by monopolists. The name derives from a standard example, that of a firm which monopolizes retailing in a region through ownership of a chain of stores.

In each sub-game, the corresponding potential competitor has to decide whether to enter into competition with the monopolist. If the potential competitor does enter the market, then the monopolist faces a choice between two alternatives: (1) engage in predatory pricing, driving the competitor out of business, at a great cost to both the competitor and to the monopolist, or (2) reach

an accomodation with the competitor (eg, by buying her out), which yields the competitor a profit and which costs the monopolist less than the first alternative does.

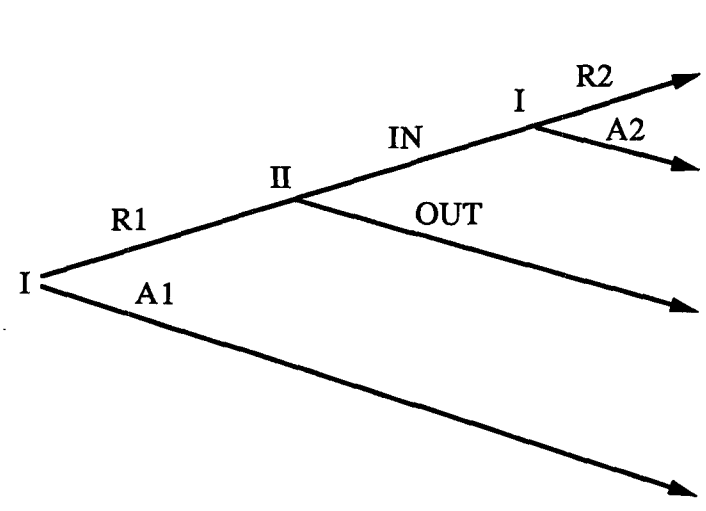
Selten demonstrated the following paradoxical result: if we assume that the exact number of potential competitors is a matter of common belief, then it is irrational for the monopolist ever to engage in predatory behavior. Clearly, it would be irrational for the monopolist to prey on the very last potential competitor, should he enter the market, since there are no further competitors to deter. If the last potential competitor is rational, therefore, he will enter the market no matter what the monopolist has done in the past, so long as he still believes that the monopolist is a rational maximizer of his own interest. For this reason, it would be irrational for the monopolist to prey on the next-to-last potential competitor, since he has no hope of deterring the last competitor. This argument can be repeated indefinitely often, demonstrating (by backwards induction) that it is irrational for the monopolist to prey on any potential competitor.

2. From a Paradox of Reputation to an Antinomy of Rational Belief

Selten believed that this is paradoxical in the weak sense: a surprising, unexpected result of game theory. I would argue that it is paradoxical in the strong sense: a logical antinomy of rational belief, analogous to the paradox of the Liar. To simplify matters, let's suppose that there are only two potential competitors and that, for whatever reason, the first potential competitor does enter the market. Suppose it were true that it would be irrational for the monopolist to prey on the first competitor should she enter the market. Then preying on the first competitor would convince the second potential competitor that the monopolist is not a rational maximizer of his self-interest and therefore that he may prey on the second competitor. If we suppose that under these conditions the second competitor would be deterred from entering the market, then we have a compelling argument for preying on the first competitor's being in the monopolist's best interests. An intelligent monopolist would be aware of this argument, so preying on the first competitor would not be irrational, contrary to our original assumption.

Therefore, preying on the first competitor would not be an irrational thing for the monopolist to do. However, this means that the monopolist's preying on the first competitor would not be inconsistent with the belief that the monopolist is a rational maximizer. Thus, if the monopolist did prey on the first competitor, the second potential competitor would still be able rationally to believe that the monopolist is a rational maximizer who would certainly not prey on the second competitor. Preying on the first competitor would not, therefore, deter the second potential competitor and so would not be in the monopolist's interest. The monopolist is aware of all the facts we used to reach this conclusion, so it would be irrational for the monopolist to prey on the first competitor.

Thus, the monopolist's preying on the first competitor is irrational if and only if it is not irrational. In order to formalize this paradox, I must be a good deal more explicit about the parameters of the game. The following diagram represents the structure of the game tree.



The U1 column represents the actual utilities of the two players involved, player I (the monopolist) and player II (the competitor). In fact, both players know that U1 represents the actual utilities of both players. Nonetheless, as has been pointed out by Philip Reny [1988] and Cristina Bicchieri [1988,1989], an adequate formulation of this sort of game must not only specify what each of the players believes, but also what each would believe if subsequent observation contradicts the agent's initial set of beliefs. A complete specification of a game situation must also include a theory of rational belief revision (such as Gärdenfors's) and a specification of the relevant parameters of the agent's belief states. The U2 column represents a very special alternative possibility: I stipulate that if the competitor were to observe a fact which contradicts his initial set of beliefs, he would restore consistency by rejecting U1 and adopting U2 instead. To use the language of belief revision theory, the belief in U1 is the least entrenched of the competitor's beliefs, while the proposition that U1 or U2, as well as the propositions specifying the other parameters of the game are extremely well-entrenched. Moreover, I wil stipulate that these facts about the competitor's belief state is known by the monopolist.

The belief state I have described certainly seems to be a possible state in which such a competitor might find himself. To demonstrate that a liar-like antinomy really arises from such situations, I need describe only one possible situation from whose description the paradoxical conclusions can be drawn. The existence of similar situations in which no paradox arises (as, for example, described by [Bicchieri 1988]) in no way contradicts my claim.

Let's use the following abbreviations:

Rxy: x believes (at the outset of the game) that y

R*xy: if R1 were to occur, x would then believe that y

m: the monopolist (player I)

c: the competitor (player II)

J: if c were to play IN, then m would play A2.

K: if m were to play R1, then c would play IN.

U1: m's actual utility function corresponds to U1 above.

The following facts can either be plausibly stipulated or can be deduced immediately from the Selten game tree, assuming that the players are causal decisionmakers.

1. K iff R*cJ

2. R*cJ iff R*cU1

3. R*cU1 iff ¬RcRmK

4. RcRmK iff RmRmK

5. m believes 1 through 4.

Claim 1 simply asserts that if R1 were to be observed, the competitor would play IN iff he believes that playing IN would be result in A2 rather than R2. This is a consequence of the fact that the competitor is an expected-utility maximizer (as described by causal decision theory).

Claim 2 corresponds to the stipulation mentioned above concerning the competitor's belief-revision regime: if forced to give up U1, the competitor would retreat to belief in U2. If, after observing R1, the competitor would continue to believe that U1 accurately represents the monopolist's utility function, then the competitor would believe that playing IN would be followed by A2, not R2. In other words, he would believe J. Alternatively, if the competitor would be forced to give up his belief in U1, he would switch to a belief in U2, in which case he would believe that playing IN would not be followed by A2. Hence, he would not believe J.

Claim 3 is a consequence of the fact that, initially, the competitor does accept U1. Consequently, the competitor would give up his belief in U1 after observing R1 if and only if R1 contradicts the competitor's initial set of beliefs. R1, in turn, would contradict the competitor's initial set of beliefs if and only if the competitor initially believed that the monopolist believed in K (i.e., that R1 would fail to deter). On the one hand, if the competitor believed initially that RmK, then he would deduce, given the fact that the monopolist is a rational decisionmaker, that R1 cannot occur. On the other hand, if the competitor did not believe that RmK, then he must acknowledge R1 to be a genuine possibility, and observing R1 would not necessitate any belief change.

Claims 4 and 5 concern the monopolist's information. Claim 4 states that the monopolist and the competitor have exactly symmetrical information about whether or not the monopolist has a rational basis for believing K. This is a consequence of two stipulations: (1) the coompetitor knows what information is available to the monopolist, and (2) what reason requires of the monopolist, given that body of information, is known equally well by both the monopolist and the competitor. Finally, according to claim 5, the monopolist accepts the first four claims.

From claims 1 through 3, we can deduce that K iff ¬RcRmK. By claim 4, it follows that K iff ¬RmRmK. Finally, assuming claim 5 and assuming that the monopolist is aware of the deductions we have just completed, we can conclude that Rm[K iff ¬RmRmK]. However, as I have argued elsewhere, any proposition of the form Rm[K ↔ ¬RmRmK] is inconsistent with the axioms and rules of a very plausible theory of rational belief [Koons 1990, 1989]:

R1. R¬Rφ → ¬Rφ

R2. Rφ, where φ is a logical axiom

R3. $R(\varphi \rightarrow \psi) \rightarrow (R\varphi \rightarrow R\psi)$

R4. $R\varphi \rightarrow RR\varphi$

R5. From φ, infer $R\varphi$

In an article on the surprise quiz paradox, Doris Olin [1983] discussed the principle I call R1. She argued:

> It can never be reasonable to believe a proposition of the form 'p and I am not now justified in believing p'. For if a person A is justified in believing a proposition, then he is not (epistemically) blameworthy for believing it. But if A is justified in believing that he is not justified in believing p, then he would be at fault in believing p. Hence, if A is justified in believing that he is not justified in believing p, then he is not justified in believing p.

If one has overwhelmingly good reason for believing that acceptance of p is not ultimately justifiable in one's present epistemic situation, then that fact must undermine any reasons one has for accepting p itself. To believe that p is not ultimately justifiable in one's present epistemic situation is to believe that it is inconsistent or otherwise not cotenable with data which is, by one's own lights, weightier than the data (if any) which supports or seems to support p. This realization should undermine one's confidence in any data supporting p.

Principles R2 and R3 together entail that any logical consequence of what is rationally believed may be rationally believed. If you are persuaded by what Henry Kyburg has said against 'conjunctivitis' [Kyburg 1970], then read 'Rφ' as saying that φ belongs to the corpus of subjectively certain propositions in the relevant situation. Even Kyburg admits that the conjunction of two subjectively certain propositions is itself subjectively certain.

Principle R4 is a principle of positive introspection: it states that if a proposition p is rationally justifiable, than to believe that p is rationally justifiable is itself rationally justifiable. In a recent analysis of mutual belief, I have argued [Koons 1989] that there is a notion of "virtual belief" for which this principle holds.

Finally, principle R5 is analagous to a rule of Necessitation in modal logic. It states that if a proposition is a theorem of the doxastic logic, then that theorem itself is rationally believed.

The inconsistency can be proved as follows:

1. $R(K \leftrightarrow \neg RRK)$	Assumption
2. RK	Assumption
3. $R\neg RRK$	1,2, R2,R3 {1,2}
4. $\neg RRK$	3, R1 {1,2}
5. $RK \rightarrow \neg RRK$	Conditional proof, {1}
6. $R(K \leftrightarrow \neg RRK) \rightarrow [RK \rightarrow \neg RRK]$	Conditional proof
7. $R[R(K \leftrightarrow \neg RRK) \rightarrow [RK \rightarrow \neg RRK]]$	6, R5
8. $RR(K \leftrightarrow \neg RRK)$	1, R4 {1}
9. $R(RK \rightarrow \neg RRK)$	7, 8, R3 {1}
10. RRK	Assumption
11. $R\neg RRK$	9, 10, R3 {1,10}
12. $\neg RRK$	11, R1 {1}*
13. $R(K \leftrightarrow \neg RRK) \rightarrow \neg RRK$	Conditional proof
14. $R[R(K \leftrightarrow \neg RRK) \rightarrow \neg RRK]$	13, R5
15. $R\neg RRK$	8, 14, R3 {1}

16. RK	1, 15, R2, R3 {1}
17. RRK	16, R4 {1}*

Christina Bicchieri [1988] attempts to solve this paradox by using the theory of belief revision developed by Isaac Levi [1977,1979] and Peter Gärdenfors [1978,1984] to constrain the reaction of the second competitor to the monopolist's act of retaliation. Bicchieri points out that the competitor cannot update his beliefs by Bayesian conditionalization, since the prior probability of the monopolist's retaliating was zero (i.e., it wasn't a "serious possibility"). According to the Levi-Gardenfors theory of belief revision, the competitor should give up those beliefs which had the least epistemic importance to him. If the theory of the game (including the competitor's beliefs) and the rules of belief revision are initially matters of common belief between the two players, then Bicchieri [argues that the competitor should give up his assumption that "the players always play what they choose to play at all nodes."[Bicchieri 1988, p. 392] The competitor should assume that the retaliation was the result of an unintentionally "trembling hand", a failure by the monopolist to carry out his true intentions. The competitor will suppose such errors to be rather rare exceptions, and consequently he will not be deterred from entering. Therefore, the monopolist should not retaliate, and the backward induction argument is apparently vindicated.

But what if the least entrenched, least epistemically important belief is not the playing-what-one-chooses principle, but rather U1 (or more specifically, the belief in the truth of U1 as opposed to U2)? Surely that is a possible state for the competitor to be in, and surely the monopolist could be aware of that fact. All that is required to show that there is a genuine paradox here is to produce a single hypothetical situation from which the contradiction can be generated.

3. A Probabilistic Version of the Paradox

It is possible to construct a probabilistic version of the paradox. Our probability theory must contain something like Harper's Popper functions, constraining how an agent updates his probabilities when confronted by evidence with a zero prior probability (this corresponds to the role played by Gärdenfors's belief revision theory in the doxastic version). To generate a paradox, we need to assume the following six claims:

1. $Pc(U_1/R_1) = 1$ iff $Pc(Pm(K) \leq (a-1)/a) > 0$

2. $Pc(U_1/R_1) = 0$ iff $Pc(Pm(K) \leq (a-1)/a) = 0$

3. $Pm(K/\ Pc(U_1/R_1) = 1) = 1$

4. $Pm(K\ /\ Pc(U_1/R_1) = 0) = 0$

5. $Pc(1\ \&\ 2) = 1$

6. $Pm(K/\varphi) = Pc(K/\varphi)$, for all φ

The value of (a-1)/a is the crucial value for the probability of the subjunctive conditional K. If the monopolist assigns a probability of greater than (a-1)/a to K, then he should choose A1. Otherwise, he may rationally choose R1. It can be shown that statements 1 through 6 are in conflict with a fundamental principle of higher-order probability theory, the so-called "Miller's principle". [Miller 1966]

Van Fraassen [1984] has produced a Dutch Book argument in favor of Miller's principle. The principle has also been endorsed by Haim Gaifman [1986] and Brian Skyrms [1980]. The principle can be stated in its general form thus:

(Miller) If $P(P(\varphi) \geq x \ \& \ \psi) \neq 0$, then $P(\varphi / P(\varphi) \geq x \ \& \ \psi) \geq x$, where ψ is any conjunction of probabilistic formulae.[1]

Miller's principle corresponds closely to the axiom schema R1 presented above. I will sketch briefly the Dutch book argument for Miller. Assume that $P(P\varphi \geq x) > 0$. Suppose for contradiction that $P(\varphi / P\varphi \geq x)$ is less than x. Then the agent is vulnerable to a dutch book. He is willing to do simultaneously the following: (1) bet for $P\varphi \geq x$, at some positive odds, (2) place a conditional bet against φ on the condition that $P\varphi \geq x$ at the odds x. If $P\varphi < x$, then the agent loses the first bet and the second conditional bet is called off. If the agent wins the first bet, then he is willing to bet for φ at the odds x (since, ex hypothesi, $P\varphi \geq x$), i.e., he is willing to buy back his first bet at a net loss. If the proportion between the stakes of the two original bets is chosen correctly by the bookie, the agent is sure to suffer a loss.

I will now establish that statements 1 through 6 are inconsistent with Miller's principle. Since we are assuming that $Pm(K/\varphi) = Pc(K/\varphi)$, for all φ, we can replace 'Pm' with 'Pc' in statements 3 and 4. For simplicity's sake, I will simply use 'P'. Given the competitor's certainty of statements 1 and 2, we can replace in 3 and 4 the proposition $P(U_1/R_1) = 0$ by the proposition $P(P(K) \leq 1/a) = 0$, and we can replace the proposition $P(U_1/R_1) = 1$ by the proposition $P(P(K) \leq 1/a) > 0$. Let Q abbreviate $P(P(K) \leq (a-1)/a)$. Thus, statements 3 and 4 can be re-stated as:

3. $P(K/Q > 0) = 1$

4. $P(K/Q = 0) = 0$

Let's begin with statement 3. Using the standard rules of probability, we can calculate that:

$$P(K/Q > 0) = P(K/P(K) \leq 1/a \ \& \ Q > 0) \cdot P(P(K) \leq 1/a / Q > 0)$$
$$+ P(K/P(K) > 1/a \ \& \ Q > 0) \cdot P(P(K) > 1/a / Q > 0)$$

By Miller's principle, this sum must be less than or equal to $1/a \cdot P(P(K) \leq 1/a / Q > 0) + [\,1 - P(P(K) \leq 1/a / Q > 0)]$. Furthermore, by Miller's principle, $P(P(K) \leq 1/a / Q > 0)$ must be greater than zero. Hence, this sum must be less than 1, contrary to statement 3.

In the case of statement 4, the probability calculus entails that:

$$P(K/Q = 0) \ \geq \ P(K/P(K) > 1/a \ \& \ Q = 0) \cdot P(P(K) > 1/a / Q = 0)$$

By Miller's principle, $P(K/P(K) > 1/a \ \& \ Q = 0) > (a-1)/a$, and $P(P(K) > 1/a / Q = 0) = 1$. Hence, their product must be greater than (a-1)/a, contrary to statement 4.

At leat one of the two conditional probabilities, $P(K/Q > 0)$ and $P(K/Q = 0)$ must be well-defined. Yet, we have shown that neither statement 3 nor statement 4 can be true, contrary to what we deduced from the game tree and our theory of rational decisionmaking.

4. From Doxastic Blindspots to Doxastic Paradox

What I have established is that the monopolist suffers from what Roy Sorensen [1988] has called a doxastic "blindspot" with respect to the theory of the game. A proposition p is a doxastic blindspot for person x if p is possibly true but it is impossible for x to be in a position of rationally

---

[1] Equations or inequalities involving only numerical constants and probability-terms.

believing p. Sorensen has analyzed the very same iterated super-games discussed here, but without, I think, correctly applying his notion of a blindspot. Sorensen has claimed that the competitor is in a blindspot with respect to the monopolist's rationality after the monopolist has played R1. [Sorensen 1988, pp. 355-361; Sorensen 1986] Sorensen has not recognized that, before play has begun, the monopolist is already in a blindspot with respect to the conditions of the game. The theory of rational belief entials that the monopolist cannot recognize the truth of the biconditional 'K $\leftrightarrow \neg$RmRmK'.

There is a strong analogy between this doxastic paradox and the semantical paradox of the Liar. In the case of the liar, a biconditional of the form 'true('α') $\leftrightarrow \neg\alpha$' is provable in arithmetic and is inconsistent with a very plausible theory of truth (namely, one containing the Tarski schema). In the case of the present doxastic paradox, epistemological considerations support the possible truth of a proposition of the form 'R(R(α) $\leftrightarrow \neg\alpha$)', which is inconsistent with a very plausible theory of rational belief (one containing the schemata R1-R4 and rule R5). It would be worthwhile, therefore, to consider how well recent work on the Liar paradox transfers to the present problem.[2]

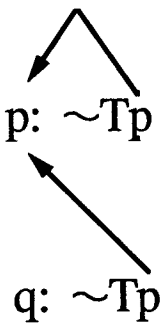
In this paper, I will discuss only one such approach: the theory of context-sensitivity developed by Charles Parsons [Parsons 1983], Burge [Burge 1979], Gaifman [Gaifman 1988], and Barwise and Etchemendy [Barwise and Etchemendy 1989]. In this approach, there is a single, univocal belief predicate, but one whose extension varies from context to context. In effect, the predicate is contextually relativized to a level in a hierarchy. By using Kripke's construction in 'Outline of a Theory of Truth', [Kripke 1975] it is possible to permit significant self-application of the belief predicate. Gaifman's work on pointer semantics provides in effect an algorithm for evaluating such indexically-relativized tokens.

One of the principal motivations for the context-sensitive approach is the phenomenon of the strengthened or extended liar. For example, Gaifman discusses the following dialogue:

Max: What I am now saying is nonsense.

Moritz: Yes, what you just said was nonsense.

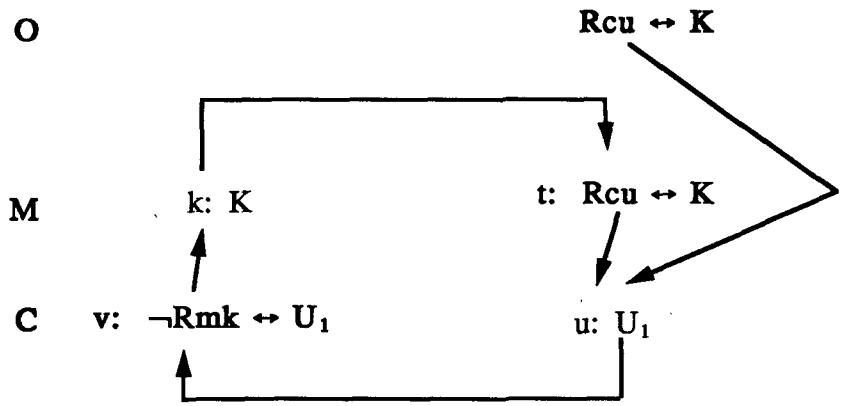
Gaifman argues that what Max said is nonsense, but that Moritz spoke the truth, despite the fact that both uttered tokens of exactly the same type. The truth-value of a token depends not only on its type (and not only on those familiar sorts of context-sensitivity as indexicality and demonstratives), but also on the token's location in a directed network of such tokens. For example, in the following network, token p is viciously circular, while token q is not.



[2]See for instance, Asher and Kamp1989 and Asher 1988.

Elsewhere I have worked out a theory of context-sensitivity for the rational-belief predicate.[Koons 1991a & 1991b] Here I will briefly sketch its application to the chain-store paradox. There are five relevant belief-tokens, two in the 'language of thought' of the monopolist, two in that of the competitor, and one in that of a detached observer. As before, let 'K' represent the subjunctive sentence: if R1 were to occur (the monopolist were to retaliate against the first competitor, then IN would occur (the second competitor would enter the market). There is a token k in the monopolist's language of thought of type 'K', representing the monopolist's **possible** belief in K. Let 'U_1' represent the true proposition that the monopolist has the sort of utility function which does not license retaliation independently of any supposed deterrent effect. There is a token u of this type in the competitor's language, representing his possible belief in U_1. The monopolist accepts a token t of the type: Rcu $\leftrightarrow$ K, that is, if the competitor subsequently accepts U_1 (that the monopolist's utility function does not license retaliation independently of deterrence), then K is true (R1 would not deter). If R1 were to occur, the competitor would accept a token v of the type: $\neg$Rmk $\leftrightarrow$ U_1, that is, if the monopolist does not believe that R1 would not deter, then he must have utility function U_1.

The situation can be pictured as follows (tokens occurring in boldface are members of the corresponding agent's data set):



The arrows represent dependency relations among the tokens. Token k is dependent on the antecedent of the token t, since the acceptability of k depends (in the monopolist's epistemic situation) on whether 'K' can be deduced from the biconditional 'Rcu $\leftrightarrow$ K'. Likewise, token u is dependent on the component 'Rmk' in token v, since the acceptability of 'U_1' for the competitor depends on the acceptance or rejection of 'Rmk'. The component 'Rcu' of t is dependent on token u, since it is an evaluation of u, and, likewise, the component 'Rmk' of v is dependent on token k.

Thus, in this case, the dependency relations constitute a closed loop. A Gaifman-like algorithm would assign a value of GAP to the tokens k and u. In this case, this would mean that the players can neither believe nor disbelieve these tokens. If we were working with a probabilistic representation of the player's beliefs, this consequence could be represented by means of interval-valued probability functions, in which the intervals straddle the crucial values. If the players respond to this sort of cognitive blindspot conservatively, by adopting a maximin principle, the resulting strategies would be: (A1; if IN, then A2) and (if R1, then OUT). These strategies are in neither a Nash nor even a correlated equilibrium. They are not rationalizable, even though the players are fully rational, because the liar-generated cognitive blindspot has made full mutual knowledge impossible, despite the optimal availability of information.

Because of the context-dependent nature of thought the token t does not in fact express a truth, although the superficially similar token t', entertained by detached observers, would express a true proposition:

t': Rcu ↔ K

Through observation of games played by similar agents, the monopolist may come to realize that t does not, in its actual context, express a true proposition. Unfortunately, due to the context-sensitivity of thought, he is unable to grasp the thought expressed by t'. Thus, he is unable rationally to infer from the fact that the competitor does not believe token u that R1 would in fact deter the competitor. The detached observer has access to information (expressed by t') which is in principle unavailable to the engaged participants and is able rationally to infer that deterrence would be successful. Thus, even when both participants and observers have access to exactly the same sources of information, a distinction between *rational from the participant's point of view* and *optimal from the observer's point of view* will remain.

By applying this context-sensitive solution theory to the paradoxes of rational credibility which appear in the analysis of games of reputation, I been able to locate and to explain certain "cognitive blindspots" inherent in these games. As a result of these cognitive blindspots, players in these games are unable to predict how the other players might act, despite the fact that all the information relevant to determing how they will act is a matter of common knowledge, in public view.

5. Consequences

In respect of the field of game theory, the arguments presented here give support to the view that the Nash equilibrium solution concept is too narrow (Bernheim 1984, Pearce 1984, Aumann 1987, Brandenburger & Dekel 1987). However, these arguments also suggest that the most popular candidate for replacing Nash equilibrium, namely, rationalizability or, equivalently, the correlated equilibrium, is also too narrow. The paradox-generated cognitive blindspots undermine any basis for confidence that the players will choose mutually coherent strategies (as in a Nash or correlated equilibrium), even under optimal conditions of mutual transparency. For example, Aumann [1987] shows that mutual knowledge of rationality implies that the players' strategies constitute a correlated equilibrium. Aumann's theorem is unassailable, but the analysis presented here throws into doubt its general applicability, even under ideal conditions. Mutual knowledge can be blocked by doxastic blindspots alone, even under conditions of perfect mutual transparency. By exploiting the details of the doxastic logics developed here, it should be possible to define a more general solution.

If there is, for the competitor in the chain-store game, a **finite, nonzero** probability of U2, then there is a Nash (and, therefore, also a correlated) equilibrium solution, as was demonstrated by Kreps, Milgrom, Roberts and Wilson [1982]. In this case, we cannot derive an antinomy in any simple way from the description of the game situation, as we were able to do when U2 had a zero probability. Nonetheless, by calling into question the possibility of mutual knowledge in such situations, this investigation does cast doubt upon the propriety of the Kreps et al. solution. As we have seen, we cannot simply assume that under ideal conditions agents enjoy perfect mutual knowledge of what reason demands of each player. Some of the propositions needed to enjoy such knowledge may be cognitively inaccessible for contextual reasons. In order to justify a solution like that of Kreps et al., we must describe a process of reasoning by which the agents would be able to reach the appropriate conclusions. Game theory must take into account the dynamics of deliberation and not be content with the static analysis of equilibrium theory.

The cognitive blindspots postulated by the context-dependent solution theory have a wider set of implications, implications concerning the relation between institutionalist social theory and the rational agent model, and, in ethics, the relation between deontic, rule-based ethical theories and consequentialism. Traditionally, theories based on the rational agent model (including much of mainstream economics) and theories which describe social reality in terms of rules, practices, and institutions (what we might call "social science proper") have been taken to be hostile competitors or, at least, as unrelated and incommensurable approaches. The understanding of paradox-generated cognitive blindspots holds some promise of achieving a reconciliation of these two approaches by explaining how the phenomenon of rule-following can emerge in a society of utility-maximizing rational agents.

In these paradoxical games of reputation, the players are unable, as a result of these cognitive blindspots, to predict how another playes might act, even given some particular utility function for that player. Consequently, the players are also unable to learn anything about the others' utilities from their actually observed behavior. Thus, the essence of reputation does not consist in conveying information (or misinformation) about one's actual state of mind to others via one's observable behavior. Instead, players afflicted by these cognitive blindspots must "learn" from the observed behaviors of others in a way quite different from the standard Bayesian model of hypothesis confirmation.

My conjecture is that rational agents must engage in some sort of as-if reasoning: they must (for want of a better alternative) operate on the basis of deliberate misrepresentations of the situation which, unlike the accurate representation, are not afflicted by cognitive blindspots. For example, they might pretend that there are two possibilities for the monopolist's utility function, one according to which retaliation is not costly at all, and another according to which retaliation is so costly as to be prohibited even if it did deter. They must do this in order to avoid being disabled by blindspots, despite the fact that they know perfectly well that the monopolist's utility function is of neither of these two kinds! An interesting question which arises here is: what sort of properties make such fictional hypotheses salient choices for this purpose?

By using this sort of as-if modelling of the situation, the players will be able to "learn" (in some attenuated sense) from the observed behavior of others. For example, if the monopolist does in fact retaliate, they will take this as "confirming" the pseudo-hypothesis that his utility function is such as to make retaliaton optimal, even without deterrent effects. In this sense, the monopolist can be characterized as following a rule (namely, the rule of always retaliating against market entries) which is distinct from the rule of always maximizing one's actual utility. Moreover, if the monopolist recognizes that the other players will engage in this sort of as-if modelling of the situation, then it can be reasonable for him to conform to this rule, in order to increase the quasi-probability of the appropriate fiction. At this point, there is a fairly rich sense in which the monopolist is following a rule distinct from the rule of individual-utility-maximization. The monopolist is not merely employing a convenient rule of thumb: he is conforming to a salient regularity because such conformity is itself an integral part of his plan.

A similar line of reasoning could be used to explain the possibility of various kinds of "indirect" consequentialisms, such as rule-utilitarianism, and to explain why such indirect forms of consequentialism do not simply collapse into their direct counterparts. It has been argued, for example, that it would be simply inconsistent for a utilitarian to follow any rule other than: maximize utility! An analysis in terms of paradox-driven cognitive blindspots could be used to show that this argument is simply wrongheaded: in the pursuit of a valuable reputation, it would be possible for a consistent utilitarian to follow (in a strict sense) rules other than that of utility maximization.

References

Asher, N. and Kamp, H. 1989. Self-Reference, Attitudes, and Paradox. *Property, Types and Meaning*, G. Chierchi, B. Partee, and R. Turner (eds.), Dordrecht: Kluwer Academic, pp. 85-158.

Asher, N. 1988. Reasoning about Belief and Knowledge with Self-Reference and Time. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, J. M. Y. Vardi (ed.). Los Altos, Calif.: Morgan Kaufmann, pp. 61-82.

Aumann, R. 1987. Correlated Equilibria as an Expression of Bayesian Rationality. *Econometrica* 55:1-18.

Barwise, J. and Etchemendy, J. 1989. *The Liar*, Oxford: Oxford University Press.

Bernheim, D. 1984. Rationalizable Strategic Behavior. *Econometrica* 52:1007-1028.

Bicchieri, C. 1988. Common Knowledge and Backward Induction: A Solution to the Paradox. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, J. M. Y. Vardi (ed.). Los Altos, Calif.: Morgan Kaufmann, pp. 381-393.

Bicchieri, C. 1989. Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge. *Erkenntnis* 30: 69-85.

Burge, T. 1979. Semantical Paradox. *Journal of Philosophy* 76:169-198.

Gaifman, H. 1986. A Theory of Higher Order Probabilities. *Theoretical Aspects of Reasoning about Knowledge*, J. Y. Halpern (ed.), Los Altos, Calif: Morgan Kaufmann, pp. 275-292.

Gaifman, H. 1988. Operational Pointer Semantics: Solutions to the Self-Referential Puzzles I. *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, J. M. Y. Vardi (ed.). Los Altos, Calif.: Morgan Kaufmann, pp. 43-60.

Gärdenfors, P. 1978. Conditionals and Changes of Belief. *Acta Philosophica Fennica* 30: 381-404.

Gärdenfors, P. 1984. Epistemic Importance and Minimal Changes of Belief. *Australasian Journal of Philosophy* 62: 136-157.

Hodgson, D. H. 1967. *The Consequences of Utilitarianism* , Oxford: Clarendon Press, pp. 38-50, 86-88.

Kaplan, D. and Montague, R. 1960. A Paradox Regained. *Notre Dame Journal of Formal Logic* 1: 79-90.

Koons, R. 1989. A Representational Account of Mutual Belief. *Synthese* 81:21-45.

Koons, R. 1990. Doxastic Paradox without Self-Reference. *Australasian Journal of Philosophy* 68:168-177

Koons, R. 1991. Doxic Paradox: A Situational Approach. *Situation Theory and Its Applications, II,* J. Barwise, J. M. Gawron, G. Plotkin, and S. Tutiya (eds.). Stanford: Center for the Study of Language and Information, pp. 161-178.

Koons, R 1991. *Paradoxes of Belief and Strategic Rationality*, New York: Cambridge University Press.

Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. 1982. Rational Cooperation in the Repeated Prisoner's Dilemma. *Journal of Economic Theory* 27:245-52.

Kripke, S. 1975. Outline of a Theory of Truth. *The Journal of Philosophy* 72:690-716.

Kyburg, H. 1970. Conjunctivitis. *Induction, Acceptance and Rational Belief*, M. Swain (ed.), Dordrecht: Reidel, pp. 55-82.

Levi, I. 1977. Subjunctives, Dispositions, and Chances. *Synthese* 34: 423-455.

Levi, I. 1979. Serious Possibility. *Essays in Honour of Jaakko Hintikka*, Dordrecht: Reidel, pp. 219-236.

Luce, R. D. and Raiffa, H. 1957. *Games and Decisions* , New York: Wiley, pp. 100-102

Miller, D. 1966. A Paradox of Information. *British Journal for the Philosophy of Science* 17:59-61.

Montague, R. 1963. Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability. *Acta Philosophica Fennica* 16:153-167.

Olin, D. 1983. The Prediction Paradox Resolved. *Philosophical Studies* 44: 229.

Parsons, C. 1974. The Liar Paradox. *Journal of Philosophical Logic* 3:381-412.

Pearce, D. 1984. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica* 52:1029-1050.

Regan, D. 1980. *Utilitarianism and Cooperation,* Oxford: Clarendon Press, pp. 69-80.

Reny, P. 1988. Rationality, Common Knowledge and the Theory of Games. Mimeo, Department of Economics, University of Western Ontario.

Selten, R. 1978. The Chain-store Paradox. *Theory and Decisions* 9: 127-159.

Skyrms, B. 1980. Higher Order Degrees of Belief. *Prospects for Pragmatism*, D. H. Mellor (ed.), Cambridge: Cambridge University Press, pp. 109-37.

Sorensen, R. A. 1986. Blindspotting and Choice Variations of the Prediction Paradox. *American Philosophical Quarterly* 23: 337-52.

Sorensen, R. A. 1988. *Blindspots*, Oxford: Clarendon Press.

van Fraassen, B. 1984. Belief and the Will. *Journal of Philosophy* 81:231-256.