Discussion Paper No. A-553

Undecidable Statements in Game Theory

Christian Ewerhart

June 1997

Undecidable Statements in Game Theory

Christian Ewerhart[1]

June 1997

Abstract:

This paper points towards formally undecidable statements in non-cooperative game theory. We present a general theory where rational players base their decisions solely on the provable implications of their assumptions. For a version of the centipede game, we show that it is undecidable whether a deviation from the backward induction path is possible under commonly assumed rationality or not. The approach retraces various impossibility results in the definition of rational behavior to the presence of undecidable statements. It is argued that the problem of undecidability can be avoided by assuming that each player has a *private* epistemic model of his opponents.

Keywords: Game Theory, Rationality, Undecidability

[1]University of Bonn, Wirtschaftspolitische Abteilung, Adenauerallee 24–26, 53113 Bonn, Germany. The paper is based on Chapter II of my Ph.D. thesis in the European Doctoral Program. Previous versions of this paper have been presented at Bonn, Leuven-la-Neuve, Iowa City and Stony Brook under the titles "Commonly Assumed Rationality" and "Language-based Reasoning". For helpful comments and discussions, I am grateful to Mamoru Kaneko, Tilman Börgers, Robert Aumann, Geir Asheim, Arnis Vilks, Urs Schweizer, Avner Shaked, Stephen Morris, and Dieter Balkenborg.

I Introduction

What characterizes rational behavior in games? Is it at all possible to give a precise definition of perfect rationality? A series of game-theoretic articles comes to the conclusion that the answer to the latter question is negative. Basu (1990) and Börgers (1994) show that rational behavior cannot be defined in extensive-form games. Samuelson (1992) and Samuelson and Börgers (1992) prove the impossibility of defining what they call cautious rationality in normal-form games. In this paper the above-cited impossibility results are retraced to a common source; it is argued that the presence of *undecidable statements* in game theory is responsible for the undefinability of rationality notions, both for normal and for extensive-form games.

The phenomenon of undecidability was discovered by Kurt Gödel when working on the foundation of mathematical theories in the 1930s. Its existence is due to a subtle distinction that is made by logicians between the mathematical notions of truth and provability. Of course, any theorem that can be proved is true, and consequently, most theoretical work is unaffected by this difference. However, the converse does not always hold, in which case undecidable statements exist, and it will be one of the principle objectives of this paper to argue that this fact is relevant for the definition of rational behavior in games.

Truth and provability may differ in the following sense. There may be candidate theorems that cannot be proved, but which are nevertheless true[2]. The underlying theory is then said to be incomplete. Using this term, the objective of this

[2]The unability to prove a statement could be due to a limited ability to find a proof. In this paper, however, we understand the unability to prove a statement as the non-existence of a proof.

paper is it to give evidence for the proposition that standard game theory is in fact incomplete.

The paper is organized as follows. In section II, we review the logical framework necessary to prove our undecidability results. Everything in this section is well-known. In Section III we give the key definitions of impossible, possible, provably impossible, and provably impossible strategies. We also introduce what we call response correspondences. Then we prove the existence of assumptions that represent commonly assumed rationality. In section IV, we prove the existence of an effective procedure to determine the set of strategies which are (provably) (im-)possible under commonly assumed rationality. Two examples are given, each of which involves a strategy for which it is formally undecidable whether it is possible or not under commonly assumed rationality. Section V concludes by proposing an alternative approach to epistemic logic which allows players to have private epistemic models of their opponents.

II Logical Prerequisites

It will be necessary to introduce a logical framework to formally show the undecidability of certain statements[3]. The framework developed in the sequel is known as Peano Arithmetic which is an axiomatization of the integers. We choose this framework since it is standard in matematical logic. Initially, our exposition will closely follow Boolos (1993, chapter 2).

[3]The effort that must be taken to give precise proofs is considerable. However, a previous, much simpler version of this paper that argued on a more intuitive level was not found to be satisfactory.

Symbols Peano Arithmetic uses logical and non-logical symbols.

- Logical symbols
  - $v_0, v_1, ...$ infinitely many distinct variables,
  - $\perp$ logical falsity,
  - $\rightarrow$ conditional sign ("implies"),
  - $\forall$ universal quantifier ("for all"),
  - $=$ sign of identity,
- Non-logical symbols
  - $\mathbf{0}$ individual constant,
  - $\mathbf{s}$ 1-place function symbol ("successor"),
  - $+$ 2-place function symbol ("addition"),
  - $*$ 2-place function symbol ("multiplication").

Ordered pairs It will be necessary to be very precise about the way in which symbols are concatenated. We assume that for any "objects" a and b there is an object $< a, b >$, the *ordered pair* of a and b. The law of ordered pairs is: if $< a, b >=< c, d >$, then $a = c$ and $b = d$. The *ordered triple* $< a, b, c >$ of a, b, and c is the ordered pair $< a, < b, c >>$.

Terms The *terms* of Peano Arithmetic are defined as follows.

- Every variable is a term,
- $\mathbf{0}$ is a term,
- if t is a term, then $\mathbf{s}t$ is a term,

- if t and t' are terms, then $(t{+}t')$ and $(t{*}t')$ are terms,

- no other string is a term.

We suppose that $\mathbf{s}t$ is the ordered pair of $\mathbf{s}$ and t, and that $(t{+}t')$ and $(t{*}t')$ are the ordered triples of $+$, t, and t' and of $*$, t, and t', respectively. Terms of the form $\mathbf{0}$, $\mathbf{s0}$, $\mathbf{ss0}$, etc. will be referred to as *numerals* and written as $\mathbf{0}$, $\mathbf{1}$, $\mathbf{2}$, etc. respectively.

Formulas F is an *atomic formula* if either F is the symbol $\perp$, or for some terms t and t', F is $t = t'$, where $t = t'$ is understood to be the ordered triple of $=$, t, and t'. Here is a definition of a *formula*.

- Every atomic formula is a formula,

- if F and F' are formulas, then $(F \to F')$ is a formula,

- if v is a variable and F is a formula, then $(\forall_v F)$ is a formula,

- no other strings are formulae.

Again, $(F \to F')$ is the ordered triple of $\to$, F, and F'. Similarly, $(\forall_v F)$ is the ordered triple of $\forall$, v, and F.

Free variables and sentences The variable v is said to be *free* in formula F if there is a finite sequence $h_0, ...h_r$ such that h_0 is an atomic formula $t = t'$ and v occurs in either t or t', h_r is equal to F, and for all $i < r$, either for some formula F', $h_{i+1} = (h_i \to F')$ or $(F' \to h_i)$, or for some variable u different from v, $h_{i+1} = \forall_u h_i$. A formula with no free variables is said to be a *sentence*.

Notational conventions The negation of F is defined as $(F \to \bot)$ and written as $\neg F$. The inequality $t \neq t'$ abbreviates $\neg t = t'$. We suppose that the other familiar logical symbols $\wedge$, $\vee$, $\leftrightarrow$, and $\exists$ are defined in any one of the usual ways, and we omit parentheses when it is reasonable to do so. Finally, we shall use the symbols $\bigwedge$ and $\bigvee$ for finite conjunctions and disjunctions, respectively.

Axioms There are logical and non-logical axioms. The *logical axioms* are the tautologies and any formula of the following list (cf. Monk (1976)):

(L1) $\forall_v(e \to f) \to (\forall_v e \to \forall_v f)$,

(L2) $e \to \forall_v e$, if v does not occur in e,

(L3) $\exists_v(v = t)$, if v does not occur in t,

(L4) $t = t' \to (e \to f)$ if e and f are atomic formulas and f is obtained from e by replacing an occurrence of t in e by t'.

Here, e and f denote formulas, v a variable, and t as well as t' terms. The *non-logical axioms* of Peano Arithmetic are the following six formulas

(N1) $\mathbf{0} \neq \mathbf{s}x$,

(N2) $\mathbf{s}x = \mathbf{s}y \to x = y$,

(N3) $x{+}\mathbf{0} = x$,

(N4) $x{+}\mathbf{s}y = \mathbf{s}(x{+}y)$,

(N5) $x{*}\mathbf{0} = \mathbf{0}$, and

(N6) $x{*}\mathbf{s}y = x{*}y{+}x$,

and the induction axioms, which are the (infinitely many) formulas

(I) $(\forall_x(x = \mathbf{0} \to F) \wedge \forall_y[\forall_x(x = y \to F) \to \forall_x(x = \mathbf{s}y \to F)]) \to F,$

where F is any formula, x is any variable, and y is any variable not occuring in F and different from x.

Deduction rules G is said to be a consequence by *modes ponens* of $(F \to G)$ and F and $\forall_v F$ is said to be a consequence by *generalization* of F. A *proof* of a formula F is a finite sequence of formulas, each entry of which is either an axiom or a consequence of modus ponens or generalization of earlier formulae in the sequence, and whose last entry is F. The formula F is a *theorem* or provable if there is a proof of F. If F is a theorem, we write $PA \vdash F$.

Semantics A sentence of Peano Arithmetic is called *true* if it is true when its variables range over the natural numbers $0, 1, 2, ...$ and $\mathbf{0}$, and $\mathbf{s}$, $+$ and $*$ denote zero and the successor, addition, and multiplication functions. Every theorem of PA is true.

Gödel numbers We associate with each symbol, term, and formula of PA a natural number, called its Gödel number. To the symbols $\perp$, $\to$, $\forall$, $=$, $\mathbf{0}$, $\mathbf{s}$, $+$, $*$, we assign the numbers $1, 3, 5, 7, 9, 11, 13$, and 15. To the variable v_i, we assign the number $2i + 17$. To define Gödel numbers for terms and formulas, it suffices to stipulate a rule for ordered pairs: if the objects x and y (whether symbols or ordered pairs) have Gödel numbers i and j, then the ordered pair $< x, y >$ shall have the Gödel number

$$\pi(i, j) = 2((i + j)^2 + i + 1). \tag{1}$$

It can be shown that this Gödel numbering does not assign the same number to different objects (cf. Boolos (1993, p. 36)). If F is a formula, then $\lceil F \rceil$ denotes

the Gödel number of F.

Formal provability Let $Bew(x)$ denote the standard "provability" predicate of Peano Arithmetic (for the lengthy construction of the formula $Bew(x)$, and for the proofs of its properties stated below, we refer the reader to Boolos (1993, Chapter 2)). If S is a sentence, then $Bew(\ulcorner S \urcorner)$ is a sentence asserting that S is a theorem of Peano Arithmetic. If S, T are sentences, then:

(P0) The sentence $Bew(\ulcorner S \urcorner)$ is true iff $PA \vdash S$,

(P1) if $PA \vdash S$ then $PA \vdash Bew(\ulcorner S \urcorner)$,

(P2) $PA \vdash Bew(\ulcorner S \to T \urcorner) \to Bew(\ulcorner S \urcorner) \to Bew(\ulcorner T \urcorner)$,

(P3) $PA \vdash Bew(\ulcorner S \urcorner) \to Bew(\ulcorner Bew(\ulcorner S \urcorner) \urcorner)$,

(P4) if PA is consistent[4], then $PA \nvdash \neg Bew(\ulcorner \bot \urcorner)$.

Property 4 is Gödel's Second Incompleteness Theorem.

Constant sentences We call a sentence *constant* if it is an element of the smallest set of sentences containing $\bot$ and containing $(S \to T)$ and $Bew(\ulcorner S \urcorner)$ whenever it contains S and T.

The diagonal lemma We shall make use of the following theorem, taken from Boolos (1993, p. 53). Suppose that $y_1, ..., y_n, z_1, ..., z_m$ are distinct variables and that $P_1(y_1, ..., y_n, z), ..., P_n(y_1, ..., y_n, z)$ are formulas of the language of PA in which all free variables are among $y_1, ..., y_n, z$. (z abbreviates $z_1, ...z_m$.) Then there exist formulas $\phi_1(z), ..., \phi_n(z)$ of the language of PA in which all free variables

[4]In mathematical logic, it is common practice to treat the consistency of formal number theory as an unproved assumption. Cf. Mendelson (1964, p. 107).

are among z, such that

$$\begin{array}{rcl} PA & \vdash & \phi_1(z) \leftrightarrow P_1(\lceil\phi_1(z)\rceil, ..., \lceil\phi_n(z)\rceil, z) \\ & \vdots & \\ PA & \vdash & \phi_n(z) \leftrightarrow P_n(\lceil\phi_1(z)\rceil, ..., \lceil\phi_1(z)\rceil, z). \end{array} \tag{2}$$

III Strategies and Logic

III.1 Impossible strategies

Consider a game for N players with finite strategy set S_i for player i, where the preferences shall remain unspecified for the moment. We assume that the elements of S_i are numerals $\mathbf{s_i}$. To formalize the players' strategic choices, fix once and for all pairwise distinct variables $s_1, ..., s_N$. We shall say that a formula A is an *assumption* if the free variables in A are among the variables $s_1, ..., s_N$. For an assumption A_i on player i and for a strategy $\mathbf{s_i} \in S_i$, let $imp_i(\mathbf{s_i}, A_i)$ be the sentence

$$\forall_{s_1} ... \forall_{s_N} (A_i \rightarrow s_i \neq \mathbf{s_i}). \tag{3}$$

In words this says that player i's assumption A_i implies that strategy $\mathbf{s_i}$ is not chosen by player i. We shall say that the strategy $\mathbf{s_i}$ is *impossible under assumption* A_i if $imp_i(\mathbf{s_i}, A_i)$ is true; we shall call $\mathbf{s_i}$ *possible* otherwise. A strategy $\mathbf{s_i}$ will be called *provably impossible* if $imp_i(\mathbf{s_i}, A_i)$ is provable. Similarly, the strategy $\mathbf{s_i}$ is said to be *provably possible* if $\neg imp_i(\mathbf{s_i}, A_i)$ is a theorem.

Proposition 1 *Fix some assumption A_i. Then the following holds:*

1. *Every strategy provably impossible under A_i is impossible under A_i;*

2. *Every strategy provably possible under A_i is possible under A_i.*

3. *No strategy can be both provably possible and provably impossible under the same* A_i.

Proof: Assertions 1 and 2 follow from the fact that every theorem of PA is true. Assertion 3 is an immediate consequence of 1 and 2. q.e.d.

We shall call an assumption *inconsistent* if it is refutable, i.e. if $PA \vdash \neg A_i$.

Proposition 2 *Under the empty assumption* $\top$*, all strategies are provably possible. Under any inconsistent assumption* A_i*, all strategies are provably impossible.*

Proof: If $A_i = \top$ then

$$PA \vdash imp_i(\mathbf{s_i}, A_i) \leftrightarrow \bot \ . \tag{4}$$

Similarly, if $\vdash \neg A_i$ then

$$PA \vdash imp_i(\mathbf{s_i}, A_i) \leftrightarrow \top. \tag{5}$$

q.e.d.

We extend the definitions given above to strategy profiles. Fix a player i and an assumption A_{-i}. Still, A_{-i} is a formula, but here the interpretation is that it is an assumption on the opponents of player i, while A_i denotes an assumption on player i. Write $imp_{-i}(\mathbf{s_{-i}}, A_{-i})$ for the sentence

$$\bigwedge_{j \neq i} imp_j(\mathbf{s_j}, A_{-i}) \tag{6}$$

As above, we say that a strategy profile $\mathbf{s_{-i}}$ is impossible, possible, provably impossible, provably possible if $imp_{-i}(\mathbf{s_{-i}}, A_{-i})$ is true, false, provable, and disprovable, respectively.

III.2 Response Correspondences

Fix a player i. Let $B_i(\cdot)$ be a correspondence that specifies, for any $R_{-i} \subseteq S_{-i}$, a nonempty set $B_i(R_{-i}) \subseteq S_i$. We shall refer to such a $B_i(\cdot)$ as a *response correspondence* for player i. We give three examples. Assume now that the utilities of the players are given.

Bayesian normal-form rationality Let $B_i^b(R_{-i})$ denote the set of strategies of player i which maximize expected utility for some probability distribution on R_{-i} if $R_{-i} \neq \emptyset$, and on S_{-i} if $R_{-i} = \emptyset$.

Cautious normal-form rationality Let $B_i^c(R_{-i})$ be the set of player i's best replies to Bayesian beliefs with support equal to R_{-i} if $R_{-i} \neq \emptyset$, and equal to S_{-i} if $R_{-i} = \emptyset$. Thus, given any set of profiles $R_{-i} \neq \emptyset$, the strategies in $B_i^c(R_{-i})$ are utility maximizing with respect to some subjective probability distribution on S_{-i} that gives strictly positive probability to all profiles in R_{-i} and probability zero to all other profiles.

Rationality in the extensive form For a given set R_{-i} of profiles, let $B_i^e(R_{-i})$ be the set of best replies to some non-archimedian full support probability distribution p where $p(\tilde{\mathbf{s}}_{-\mathbf{i}})/p(\mathbf{s}_{-\mathbf{i}})$ is infinitesimal for all $\mathbf{s}_{-\mathbf{i}} \in R_{-i}$ and $\tilde{\mathbf{s}}_{-\mathbf{i}} \notin R_{-i}$. This is the set of best replies to a non-standard probability distribution which gives infinitely more weight to any profile in R_{-i} than to any strategy not in R_{-i}. In words, a strategy is rational if it is a best reply given that a rational strategy of the other player is given infinitely more weight than any non-rational strategy.

III.3 Rationality

For $R_{-i} \subseteq S_{-i}$ and for an assumption A_{-i}, define $Prm_i(R_{-i}, A_{-i})$ as

$$\bigwedge_{\mathbf{s_{-i}} \in R_{-i}} \neg Bew(\lceil imp_{-i}(\mathbf{s_{-i}}, A_{-i})\rceil) \ \wedge \bigwedge_{\mathbf{s_{-i}} \in S_{-i} \setminus R_{-i}} Bew(\lceil imp_{-i}(\mathbf{s_{-i}}, A_{-i})\rceil). \quad (7)$$

This sentence is true if R_{-i} is the set of profiles which are not provably impossible. Let $B(\cdot)$ be some response correspondence. Define $Rat_i^{R_{-i}}(A_{-i}, B)$ as the formula

$$Prm_i(R_{-i}, A_{-i}) \longrightarrow \bigvee_{\mathbf{s_i} \in B_i(R_{-i})} s_i = \mathbf{s_i}.$$

The formula has the following interpretation: if R_{-i} is the set of strategy profiles that are not provably impossible under assumption A_{-i}, then player i chooses a strategy from $B_i(R_{-i})$. We shall write $Rat_i^{S_i}(A_{-i}, B)$ ("player i is rational w.r.t. his assumption A_{-i}") for the formula

$$\bigwedge_{R_{-i}} Rat_i^{R_{-i}}(A_{-i}, B).$$

To simplify notation, this formula will from now on be written as $Rat_i(A_{-i})$. Note that $Rat_i(A_{-i})$ is itself an assumption. It is the assumption that player i is rational and bases his decisions on the assumption A_{-i}.

III.4 Commonly Assumed Rationality

Theorem 1 *There exist assumptions car_1,..., car_N such that for $i = 1, ..., N$,*

$$PA \vdash car_i \leftrightarrow Rat_i(\bigwedge_{j \neq i} car_j). \quad (8)$$

The formula car_i expresses that player i is rational and bases his decisions on the assumption $\bigwedge_{j \neq i} car_j$. Although there are differences in interpretation, the defined formula should be seen as an analogue to "common knowledge of rationality".

Proof: We shall apply the diagonal lemma (cf. Section II). It will suffice to define appropriate formulas $P_i(y, s)$, for $i = 1,...N$, where y and s abbreviate $y_1,...y_N$ and $s_1,...,s_N$, respectively. Fix some i. The key step of the proof is the construction of terms $t_{j,\mathbf{s_j}}(x)$, for $j \neq i$ with the property

$$t_{j,\mathbf{s_j}}(\lceil A_{-i} \rceil) = \lceil imp_j(\mathbf{s_j}, A_{-i}) \rceil, \tag{9}$$

for all assumptions A_{-i}, where on the left-hand-side, $\mathbf{0}$ and $\mathbf{s}$, $+$ and $*$ denote zero, and the successor, addition and multiplication function. Fix some strategy $\mathbf{s_j}$. Then the term $t_{j,\mathbf{s_j}}$ transforms the Gödel number $\lceil A_{-i} \rceil$ of an assumption A_{-i} into the Gödel number $t_{j,\mathbf{s_j}}(\lceil A_{-i} \rceil)$ of a sentence to the effect that the strategy $\mathbf{s_j}$ is impossible under A_{-i}. To find such a term, recall that $imp_j(\mathbf{s_j}, A_{-i})$ is defined as

$$\forall_{s_1}...\forall_{s_N}(A_{-i} \rightarrow s_j \neq \mathbf{s_j}), \tag{10}$$

and that this is a more readable variant of the more explicit representation

$$< \forall, s_1, ..., < \forall, s_N, <\rightarrow, A_{-i}, s_j \neq \mathbf{s_j} >> ... > . \tag{11}$$

Hence, the number of this formula may be calculated, via the π function defined in Section II, from the Gödel numbers of the symbols $\forall$, s_1,..., s_N, $\rightarrow$, from the number of the formula $s_j \neq \mathbf{s_j}$, and from $\lceil A_{-i} \rceil$. As the π function is expressible as a term of Peano Arithmetic and since the substitution of a variable by a term maps terms to terms, the function $t_{j,\mathbf{s_j}}(x)$ may in fact be represented by a term of PA. In a similar fashion, construct a term $\bar{t}_i(y)$ with the property

$$\bar{t}_i(\lceil A_1 \rceil, ..., \lceil A_N \rceil) = \lceil \bigwedge_{j \neq i} A_j \rceil. \tag{12}$$

Define $t'_{-i,\mathbf{s_{-i}}}(x)$ as $\bar{t}_i(t_{1,\mathbf{s_1}}(x), ..., t_{N,\mathbf{s_N}}(x))$. The next step in the construction of the formulas P_i is a redefinition of the formula $Rat_i(A_{-i})$, where A_{-i} is replaced

by a variable x. For any R_{-i}, write $\tilde{Rat}_i^{R_{-i}}(x)$ to abbreviate

$$\left(\bigwedge_{\mathbf{s_{-i}} \in R_{-i}} \neg Bew(t'_{-i,\mathbf{s_{-i}}})(x) \ \wedge \bigwedge_{\mathbf{s_{-i}} \in S_{-i} \setminus R_{-i}} Bew(t'_{-i,\mathbf{s_{-i}}}(x)) \right) \rightarrow \bigvee_{\mathbf{s_i} \in B_i(R_{-i})} s_i = \mathbf{s_i}. \tag{13}$$

Finally, let $\tilde{Rat}_i(x)$ denote the formula

$$\bigwedge_{R_{-i}} \tilde{Rat}_i^{R_{-i}}(x). \tag{14}$$

It is clear from the construction that

$$PA \vdash Rat_i(A_{-i}) \leftrightarrow \tilde{Rat}_i(\lceil A_{-i} \rceil) \tag{15}$$

for all assumptions A_{-i}. Using again the term $\bar{t}_i(y)$, the formula $P_i(y,s)$ can be chosen as $\tilde{Rat}_i(\bar{t}_i(y))$. (Note that P_i does not possess free variables apart from y, s.) The theorem now follows from the diagonal lemma. q.e.d.

IV The Possibility of Strategies

IV.1 A reformulation

Fix once and for all a vector of assumptions car_1,...,car_N. To simplify notation, write $imp_i(\mathbf{s_i})$ for $imp_i(\mathbf{s_i}, car_i)$. Thus, $imp_i(\mathbf{s_i})$ is the sentence

$$\forall_{s_1} ... \forall_{s_N} (car_i \rightarrow s_i \neq \mathbf{s_i}). \tag{16}$$

If $imp_i(\mathbf{s_i})$ is true, false, provable, or refutable then $\mathbf{s_i}$ will be said to be impossible, possible, provably impossible, and provably possible under commonly assumed rationality, respectively. Write $Prm_i(R_{-i})$ for

$$Prm_i(R_{-i}, \bigwedge_{j \neq i} car_j). \tag{17}$$

Theorem 2 *For every player i, and for every strategy $\mathbf{s_i} \in S_i$,*

$$PA \vdash imp_i(\mathbf{s_i}) \leftrightarrow \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} Prm_i(R_{-i}). \tag{18}$$

Thus, it is a theorem of PA that a strategy $\mathbf{s_i}$ is impossible under commonly assumed rationality iff $\mathbf{s_i}$ is not a best reply given the set of provably impossible profiles for the opponents.

Proof: By Theorem 1, car_i is equivalent to the formula

$$\bigwedge_{R_{-i} \subseteq S_{-i}} (Prm_i(R_{-i}) \rightarrow \bigvee_{\mathbf{s_i} \in B_i(R_{-i})} s_i = \mathbf{s_i}). \tag{19}$$

Fix some $\mathbf{s_i} \in S_i$. Let $F(\mathbf{s_i}, R_{-i})$ be the formula

$$Prm_i(R_{-i}) \wedge \bigwedge_{\substack{\mathbf{s_i'} \in B_i(R_{-i}) \\ \mathbf{s_i'} \neq \mathbf{s_i}}} s_i \neq \mathbf{s_i'}. \tag{20}$$

Then car_i is equivalent to

$$\bigwedge_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \in B_i(R_{-i})}} (F(\mathbf{s_i}, R_{-i}) \rightarrow s_i = \mathbf{s_i}) \quad \wedge \bigwedge_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} \neg F(\mathbf{s_i}, R_{-i}). \tag{21}$$

Hence, the formula

$$car_i \rightarrow s_i \neq \mathbf{s_i} \tag{22}$$

is equivalent to

$$s_i \neq \mathbf{s_i} \ \vee \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} F(\mathbf{s_i}, R_{-i}). \tag{23}$$

Using the fact that $Prm_i(R_{-i})$ does not contain free variables, it follows from the definition of $imp_i(\mathbf{s_i})$ that

$$PA \vdash imp_i(\mathbf{s_i}) \leftrightarrow \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} Prm_i(R_{-i}) \tag{24}$$

proving Theorem 2. q.e.d.

Corollary 1 *There is always a strategy that is possible under commonly assumed rationality. In particular, car_i is a consistent assumption.*

Proof: By Theorem 2,

$$PA \vdash imp_i(\mathbf{s_i}) \leftrightarrow \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} Prm_i(R_{-i}). \tag{25}$$

Hence,

$$PA \vdash \bigwedge_{\mathbf{s_i} \in S_i} imp_i(\mathbf{s_i}) \leftrightarrow \bigwedge_{\mathbf{s_i} \in S_i} \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} Prm_i(R_{-i}). \tag{26}$$

Since

$$PA \vdash \neg(Prm_i(R_{-i}) \wedge Prm_i(R'_{-i})) \tag{27}$$

for $R_{-i} \neq R'_{-i}$ and since $B_i(R_{-i}) \neq \emptyset$,

$$PA \vdash \bigwedge_{\mathbf{s_i} \in S_i} imp_i(\mathbf{s_i}) \leftrightarrow \perp . \tag{28}$$

Since every theorem of PA is true, there is at least one $\mathbf{s_i} \in S_i$ such that $imp_i(\mathbf{s_i})$ is false. This proves the first assertion. The second then follows from Proposition 2. q.e.d.

IV.2 Resolving self-reference

Theorem 3 *There exists a collection $(C_{i,\mathbf{s_i}})$ of constant sentences such that*

$$PA \vdash imp_i(\mathbf{s_i}) \leftrightarrow C_{i,\mathbf{s_i}} \tag{29}$$

for all $\mathbf{s_i} \in S_i$ and all i. The collection can be determined by a finite procedure.

Proof: The construction is based on an iterated application of the fixed point theorem of de Jongh and Sambin. We start from the characterization of the sentences $imp_i(\mathbf{s_i})$ given in Theorem 2:

$$PA \vdash imp_i(\mathbf{s_i}) \leftrightarrow \bigvee_{\substack{R_{-i} \subseteq S_{-i} \\ \mathbf{s_i} \notin B_i(R_{-i})}} Prm_i(R_{-i}) \tag{30}$$

Call this system of equivalences Θ. Note that Θ consists of as many equivalences as there are pairs $(i, \mathbf{s_i})$. Let P denote the set of pairs $p = (i, \mathbf{s_i})$. Let K be the cardinality of P. We are going to define inductively systems Θ_k and sets P_k, for $0 \leq k \leq K$, with the following properties:

1. the system Θ_k consists of K equivalences each of which characterizes some sentence $imp_i(\mathbf{s_i})$;

2. each equivalence in Θ_k is non-recursive in the sense that $imp_i(\mathbf{s_i})$ does not appear on the right-hand-side of the equivalence;

3. the set P_k contains k pairs;

4. a sentence $imp_i(\mathbf{s_i})$ that corresponds to some pair in P_k does not appear on the right-hand-side of any equivalence in Θ_k that corresponds to a pair not in P_k.

Let $P_0 = \emptyset$ and $\Theta_0 = \Theta$. Note that conditions 1-4 are trivially satisfied for $k = 0$. To construct P_k and Θ_k from P_{k-1} and Θ_{k-1} for some $1 \leq k \leq K$, proceed as follows. Choose first some element $p_k = (i, \mathbf{s_i})$ of $P \backslash P_{k-1}$. Let

$$P_k = P_{k-1} \cup \{p_k\}.$$

The equivalences corresponding to pairs in P_k are transferred unchanged from Θ_{k-1} to Θ_k. The remaining equivalences are replaced as follows. Consider a

pair $p \in P \backslash P_k$. Substitute all occurences of $imp_i(\mathbf{s_i})$ on the right-hand-side of the equivalence corresponding to p by the right-hand-side of the equivalence corresponding to p_k. Call the resulting right-hand-side c. The equivalence now reads

$$PA \vdash imp_{i'}(\mathbf{s_{i'}}) \leftrightarrow c. \tag{31}$$

Lift the sentence c of PA to a sentence C in the modal logic GL as follows: Replace any appearance of a sentence $imp_j(\mathbf{s_j})$ by a sentence letter $q_{j,\mathbf{s_j}}$, and any appearance of $Bew(\cdot)$ by the box operator $\Box(\cdot)$. Then C is modalized in $q_{i,\mathbf{s_i}}$. By the fixed point theorem of de Jongh and Sambin (cf. Boolos (1993, Ch. 8)), there exists a sentence H of GL, containing only sentence letters contained in C, not containing the sentence letter $q_{i,\mathbf{s_i}}$ such that

$$GL \vdash \boxdot(q_{i,\mathbf{s_i}} \leftrightarrow C) \leftrightarrow \boxdot(q_{i,\mathbf{s_i}} \leftrightarrow H), \tag{32}$$

where $\boxdot S$ is the sentence $\Box S \wedge S$. Now let $*$ be a realization such that

$$imp_j(\mathbf{s_j}) = q^*_{j,\mathbf{s_j}} \tag{33}$$

for all pairs $(j, \mathbf{s_j})$. Then

$$PA \vdash (q_{j,\mathbf{s_j}} \leftrightarrow C)^*, \tag{34}$$

therefore

$$PA \vdash (\boxdot q_{j,\mathbf{s_j}} \leftrightarrow C)^*. \tag{35}$$

By Theorem 2 of Chapter 3 in Boolos (1993),

$$PA \vdash (\boxdot(q_{j,\mathbf{s_j}} \leftrightarrow C) \leftrightarrow \boxdot(q_{j,\mathbf{s_j}} \leftrightarrow H))^*, \tag{36}$$

and therefore

$$PA \vdash \boxdot(q_{j,\mathbf{s_j}} \leftrightarrow H)^*, \tag{37}$$

whence

$$PA \vdash (q_{j,\mathbf{s_j}} \leftrightarrow H)^*. \tag{38}$$

Thus, there is a sentence $h = H^*$ such that

$$PA \vdash imp_j(\mathbf{s_j}) \leftrightarrow h, \tag{39}$$

and such that the sentences $imp_{i'}(\mathbf{s_{i'}})$ corresponding to pairs in P_k do not appear in h. The equivalence in Θ_{k-1} corresponding to p is replaced by h. This defines inductively systems Θ_k. The final system Θ_K has the property that it is recursively solvable in the sense that the right-hand-side of an equivalence corresponding to some pair in P_k contains solely sentences $imp_{i'}(\mathbf{s_{i'}})$ corresponding to pairs in $P \backslash P_k$. This proves the theorem. q.e.d.

Corollary 2 *There exist effective methods for deciding whether a given strategy is possible, provably possible, impossible, and provably impossible under commonly assumed rationality, respectively.*

Proof: By Theorem 3, the problem to decide whether a strategy is possible reduces to the problem to decide the truth of a given constant sentence. But this can be done effectively (cf. Boolos (1993), p. 94). To decide whether a given strategy is provably (im-)possible note that a sentence S is a theorem if and only if $Bew(\lceil S \rceil)$ is true. Thus, the problem reduces to the one just solved. q.e.d.

IV.3 Consistent pairs

Consider the following game taken from Börgers and Samuelson (1992).

	0	**1**
0	1,1	1,0
1	1,0	0,1

FIGURE 1

The response scheme for cautious rationality is given by

$$B_1^c(\emptyset) = B_1^c(\{\mathbf{1}\}) = B_1^c(\{\mathbf{0},\mathbf{1}\}) = \{\mathbf{0}\} \tag{40}$$

$$B_1^c(\{\mathbf{0}\}) = \{\mathbf{0},\mathbf{1}\} \tag{41}$$

for player 1 and by

$$B_2^c(\{\mathbf{0}\}) = \{\mathbf{0}\} \tag{42}$$

$$B_2^c(\{\mathbf{1}\}) = \{\mathbf{1}\} \tag{43}$$

$$B_2^c(\emptyset) = B_2^c(\{\mathbf{0},\mathbf{1}\}) = \{\mathbf{0},\mathbf{1}\} \tag{44}$$

for player 2.

Proposition 3 *For the game in Figure 1, commonly assumed cautious rationality implies the following: For player 1, strategy* $\mathbf{0}$ *is provably possible, and strategy* $\mathbf{1}$ *is provably impossible. For player 2, strategy* $\mathbf{0}$ *is provably possible, and strategy* $\mathbf{1}$ *is impossible, but not provably impossible.*

Proof: Theorem 2 characterizes the formulas $imp_1(\mathbf{0})$, $imp_1(\mathbf{1})$, $imp_2(\mathbf{0})$, and $imp_2(\mathbf{0})$ as the solution of the following system of equivalences:

$$PA \vdash imp_1(\mathbf{0}) \leftrightarrow \bot \tag{45}$$

$$PA \vdash imp_1(\mathbf{1}) \leftrightarrow Bew(\ulcorner imp_2(\mathbf{0}) \urcorner) \vee \neg Bew(\ulcorner imp_2(\mathbf{1}) \urcorner) \tag{46}$$

$$PA \vdash imp_2(\mathbf{0}) \leftrightarrow Bew(\ulcorner imp_1(\mathbf{0}) \urcorner) \wedge \neg Bew(\ulcorner imp_1(\mathbf{1}) \urcorner) \tag{47}$$

$$PA \vdash imp_2(\mathbf{1}) \leftrightarrow \neg Bew(\ulcorner imp_1(\mathbf{0}) \urcorner) \wedge Bew(\ulcorner imp_1(\mathbf{1}) \urcorner) \tag{48}$$

We follow the lines of the proof of theorem 3 and look instead for a solution of a system of equivalences in GL. Formally, we replace the formula $imp_i(\mathbf{s_i})$ by the sentence letter $q_{i,\mathbf{s_i}}$ and the predicate $Bew(\lceil\cdot\rceil)$ by the box operator $\Box$, respectively.

$$GL \vdash q_{1,\mathbf{0}} \quad \leftrightarrow \quad \bot \tag{49}$$

$$GL \vdash q_{1,\mathbf{1}} \quad \leftrightarrow \quad \Box q_{2,\mathbf{0}} \vee \neg\Box q_{2,\mathbf{1}} \tag{50}$$

$$GL \vdash q_{2,\mathbf{0}} \quad \leftrightarrow \quad \Box q_{1,\mathbf{0}} \wedge \neg\Box q_{1,\mathbf{1}} \tag{51}$$

$$GL \vdash q_{2,\mathbf{1}} \quad \leftrightarrow \quad \neg\Box q_{1,\mathbf{0}} \wedge \Box q_{1,\mathbf{1}} \tag{52}$$

Now use (49) to eliminate $q_{1,\mathbf{0}}$ in (51) and (52) and the resulting expressions to eliminate $q_{2,\mathbf{0}}$ and $q_{2,\mathbf{1}}$ in (50). This yields

$$GL \vdash q_{1,\mathbf{1}} \leftrightarrow \Box(\Box \bot \wedge \neg\Box q_{1,\mathbf{1}}) \vee \neg\Box(\neg\Box \bot \wedge \Box q_{1,\mathbf{1}}), \tag{53}$$

which is a recursive equivalence for the sentence letter $q_{1,\mathbf{1}}$. A solution is $q_{1,\mathbf{1}} = \top$. Thus, from (51), $q_{2,\mathbf{0}} = \bot$, and similarly from (52), $q_{2,\mathbf{1}} = \neg\Box \bot$. This induces the following solution of the above system in PA:

$$PA \vdash imp_1(\mathbf{0}) \quad \leftrightarrow \quad \bot \tag{54}$$

$$PA \vdash imp_1(\mathbf{1}) \quad \leftrightarrow \quad \top \tag{55}$$

$$PA \vdash imp_2(\mathbf{0}) \quad \leftrightarrow \quad \bot \tag{56}$$

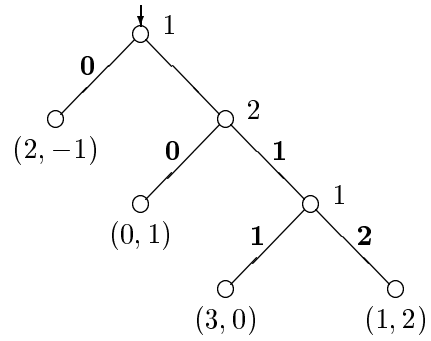
$$PA \vdash imp_2(\mathbf{1}) \quad \leftrightarrow \quad \neg Bew(\lceil\bot\rceil) \tag{57}$$

This proves the proposition. q.e.d.

Corollary 3 *For the game in Figure 1, it is formally undecidable whether strategy* $\mathbf{1}$ *is possible for player 2 under commonly assumed cautious rationality.*

IV.4 Centipede

Consider the following version of a centipede game.



	0	**1**
0	$2, -1$	$2, -1$
1	$0, 1$	$3, 0$
2	$0, 1$	$1, 2$

FIGURE 2

The response scheme for extensive-form games assumes the following form for player 1:

$$B_1^e(\{\mathbf{0}\}) = \{\mathbf{0}\} \tag{58}$$

$$B_1^e(\{\mathbf{1}\}) = \{\mathbf{1}\} \tag{59}$$

$$B_1^e(\emptyset) = B_1^e(\{\mathbf{0},\mathbf{1}\}) = \{\mathbf{0},\mathbf{1}\} \tag{60}$$

Similarly for player 2:

$$B_2^e(\{\mathbf{1}\}) = B_2^e(\{\mathbf{0},\mathbf{1}\}) = \{\mathbf{0}\} \tag{61}$$

$$B_2^e(\{\mathbf{2}\}) = B_2^e(\{\mathbf{0},\mathbf{2}\}) = \{\mathbf{1}\} \tag{62}$$

$$B_2^e(\emptyset) = B_2^e(\{\mathbf{0}\}) = B_2^e(\{\mathbf{1},\mathbf{2}\}) = B_2^e(\{\mathbf{0},\mathbf{1},\mathbf{2}\}) = \{\mathbf{0},\mathbf{1}\} \tag{63}$$

Proposition 4 *For the game in Figure 2, commonly assumed extensive-form rationality implies the following: For player 1, strategies* **0** *and* **1** *are provably possible, while strategy* **2** *is provably impossible. For player 2, strategy* **0** *is provably possible, and strategy* **1** *is impossible, but not provably impossible.*

Proof: By Theorem 2, we have to determine the solutions to the following system of equivalences in PA:

$$PA \vdash imp_1(\mathbf{0}) \quad \leftrightarrow \quad Bew(\lceil imp_2(\mathbf{0})\rceil) \wedge \neg Bew(\lceil imp_2(\mathbf{1})\rceil) \tag{64}$$

$$PA \vdash imp_1(\mathbf{1}) \quad \leftrightarrow \quad \neg Bew(\lceil imp_2(\mathbf{0})\rceil) \wedge Bew(\lceil imp_2(\mathbf{1})\rceil) \tag{65}$$

$$PA \vdash imp_1(\mathbf{2}) \quad \leftrightarrow \quad \top \tag{66}$$

$$PA \vdash imp_2(\mathbf{0}) \quad \leftrightarrow \quad Bew(\lceil imp_1(\mathbf{1})\rceil) \wedge \neg Bew(\lceil imp_1(\mathbf{2})\rceil) \tag{67}$$

$$PA \vdash imp_2(\mathbf{1}) \quad \leftrightarrow \quad \neg Bew(\lceil imp_1(\mathbf{1})\rceil) \wedge Bew(\lceil imp_1(\mathbf{2})\rceil) \tag{68}$$

As above, substitute the formula $imp_i(\mathbf{s_i})$ by the sentence letter $q_{i,\mathbf{s_i}}$ and the predicate $Bew(\lceil\cdot\rceil)$ by the box operator $\Box$, respectively.

$$GL \vdash q_{1,\mathbf{0}} \quad \leftrightarrow \quad \Box q_{2,\mathbf{0}} \wedge \neg\Box q_{2,\mathbf{1}} \tag{69}$$

$$GL \vdash q_{1,\mathbf{1}} \quad \leftrightarrow \quad \neg\Box q_{2,\mathbf{0}} \wedge \Box q_{2,\mathbf{1}} \tag{70}$$

$$GL \vdash q_{1,\mathbf{2}} \quad \leftrightarrow \quad \top \tag{71}$$

$$GL \vdash q_{2,\mathbf{0}} \quad \leftrightarrow \quad \Box q_{1,\mathbf{1}} \wedge \neg\Box q_{1,\mathbf{2}} \tag{72}$$

$$GL \vdash q_{2,\mathbf{1}} \quad \leftrightarrow \quad \neg\Box q_{1,\mathbf{1}} \wedge \Box q_{1,\mathbf{2}} \tag{73}$$

Replacing the right-hand-side of (71) for $q_{1,\mathbf{2}}$ in the equivalences (72) and (73) and using the resulting expression for $q_{2,\mathbf{0}}$ and $q_{2,\mathbf{1}}$ in (69) and (70) yields

$$GL \vdash q_{1,\mathbf{0}} \quad \leftrightarrow \quad \Box \bot \wedge \neg\Box\neg\Box q_{1,\mathbf{1}} \tag{74}$$

$$GL \vdash q_{1,\mathbf{1}} \quad \leftrightarrow \quad \neg\Box \bot \wedge \Box\neg\Box q_{1,\mathbf{1}} \tag{75}$$

$$GL \vdash q_{1,\mathbf{2}} \quad \leftrightarrow \quad \top \tag{76}$$

$$GL \vdash q_{2,\mathbf{0}} \quad \leftrightarrow \quad \bot \tag{77}$$

$$GL \vdash q_{2,\mathbf{1}} \quad \leftrightarrow \quad \neg\Box q_{1,\mathbf{1}} \tag{78}$$

Equivalence (75) is a recursive characterization of $q_{1,\mathbf{1}}$, solved e.g. by $q_{1,\mathbf{1}} = \bot$. This implies $q_{1,\mathbf{0}} = \bot$ and $q_{2,\mathbf{1}} = \neg\Box q_{1,\mathbf{1}}$. For the original system, the following

solution results:

$$PA \vdash imp_1(\mathbf{0}) \quad \leftrightarrow \quad \bot \tag{79}$$

$$PA \vdash imp_1(\mathbf{1}) \quad \leftrightarrow \quad \bot \tag{80}$$

$$PA \vdash imp_1(\mathbf{2}) \quad \leftrightarrow \quad \top \tag{81}$$

$$PA \vdash imp_2(\mathbf{0}) \quad \leftrightarrow \quad \bot \tag{82}$$

$$PA \vdash imp_2(\mathbf{1}) \quad \leftrightarrow \quad \neg Bew(\lceil \bot \rceil) \tag{83}$$

This proves the proposition. q.e.d.

Corollary 4 *For the game in Figure 2, it is formally undecidable whether strategy* $\mathbf{1}$ *is possible for player 2 under commonly assumed extensive-form rationality.*

V Conclusion

The paper discussed the existence of undecidable statements in non-cooperative game theory. Using tools from mathematical logic, we developed a model of rational behavior that distinguishes between formal truth and provability. A simple version of the centipede game was used to illustrate the fact that for some strategies it may be undecidable whether they can result from perfectly rational behavior or not.

The result is provocative for the following reason: The standard epistemic model of knowledge (Aumann (1976)) implicitly presupposes the decidability of statements. We believe that the critical feature of Aumann's model is the common epistemic model, which is shared among the players. To avoid the undecidability

issue, we propose that each player should be given a *private* epistemic model. By a private epistemic model we mean a model that describes what a player assumes, how he comes to conclusions and how his actions are influenced by his conclusions. Our point of view implies that the description of a player contains in particular his assumptions on the other player. If a player assumes that his rival is rational, then his assumptions should also specify the assumptions that he envisages the other player to have. The undecidability issue is mitigated for private epistemic models as a player's assumptions do not necessarily coincide with the assumptions that the other player expects him to have.

References

Asheim, G., and M. Dufwenberg (1996): "Admissibility and Common Knowledge," Discussion Paper, CentER.

Aumann, R. (1976): "Agree to Disagree," *Annals of Statistics*, **4**, 1236–1239.

Basu, K. (1990): "On the Non-Existence of a Rationality Definition for Extensive Games," *International Journal of Game Theory*, **19**, 33–44.

Blume, L., A. Brandenburger, and E. Dekel (1991): "Lexicographic Probabilities and Choice under Uncertainty," *Econometrica*, **59**, 61–79.

Boolos, G. (1993): *The Logic of Provability.* Cambridge, Massachusetts: Cambridge University Press.

Börgers, T. (1994): "Consistent Pairs in Extensive Games," Discussion Paper, University College, London.

Börgers, T., and L. Samuelson (1992): " 'Cautious' Utility Maximization and Iterated Weak Dominance," *International Journal of Game Theory*, **21**, 13–25.

Ewerhart, C. (1997): "Rationality and the Definition of Consistent Pairs," forthcoming in the *International Journal of Game Theory*.

Monk, J.D. (1976): *Mathematical Logic*, Springer: New York.

Samuelson, L. (1992): "Dominated Strategies and Common Knowledge," *Games and Economic Behavior*, **4**, 284–313.