THE CONWAY PARADOX:
ITS SOLUTION IN AN EPISTEMIC FRAMEWORK

by

Peter van Emde Boas,
Jeroen Groenendijk & Martin Stokhof

REPRINT FROM: MC TRACT 135
FORMAL METHODS IN THE STUDY
OF LANGUAGE

Edited by J.A.G. Groenendijk
T.M.V. Janssen
M.B.J. Stokhof

THE CONWAY PARADOX: ITS SOLUTION IN AN EPISTEMIC FRAMEWORK

by

Peter van Emde Boas,
Jeroen Groenendijk & Martin Stokhof

1. INTRODUCTION

The aim of this paper is to describe a new application of a formalism, designed originally by the last two authors as part of a theory in which various pragmatic phenomena concerning the information of language users can be handled. Using this framework, we analyse a paradox brought to the attention of the first author by CONWAY et al. (1977). In fact, the paradox involved is much older. A description of the paradox and its history was presented in GARDNER (1977); two variants can be found in LITTLEWOOD (1953). Conway's contribution consists of an impressive generalization of the situations in which the paradox can be shown to arise. We will discuss this generalization in Section 2 of the paper, but for reasons of simplicity our analysis deals only with the original simple case, indicating the explosion of the combinatorial complexity which will arise if our analysis is extended to more complicated cases.

The paradox involves hypothetical incomplete information games, to be played by perfect logicians. In the most simple situation there are two players. Each player has a non-negative number written on his forehead, which his opponent can see but which he cannot see himself. There are no mirrors available and asking the other player for information is not permitted. However, it is known to both players that according to the rules of the game the two numbers are adjacent; so a situation like (3,4) is legal, whereas (6,9) is not. Moreover, each of the players knows that the other player is informed about the rules of the game; this again is mutual knowledge up to every level. The goal of the game is to find out which number is written on one's own forehead. Both players inform each other in alternating turns about whether they know their own number, taking into account what they see and the development of the game so far. As soon as one of the players affirms that he can decide what his number is, the game terminates.

As an example, consider the game (0,1); as soon as the player with number 1 has to answer he can affirm that he knows his number, for seeing a 0, he can conclude that his own number has to be +1 or -1. Since -1 is excluded by the rules of the game, he has complete information as to what number is written on his forehead: he knows that it is +1. Similarly, in the game (1,2) the player with number 2 can answer affirmatively, as soon as the player with number 1 has given a negative answer, for the failure of the player with number 1 to terminate the game at his first turn proves to the other player that he himself does not have a zero, so he must have a 2.

The paradox arises as soon as we start analysing the games (k,k+1) for larger values of k. On the one hand a plausible argumentation can be given which shows that the game will terminate for every value of k, whereas on the other hand a straightforward analysis of a single round during a game such as (3,4) shows that such a round does not produce any useful information at all, implying that the game will never terminate. We present both argumentations in full detail in Section 2.

Our application of the epistemic framework, developed in GROENENDIJK & STOKHOF (1980), will provide for a mathematical model within which the termination proof can be shown to be correct by explicit calculation. The model also supports the possibility of obtaining a non terminating game by restricting the structures used, where these restrictions should correspond to psychological barriers in the human mind. However, it is not our intention to claim that this model explains human behaviour: our main concern is to sharpen the mathematical tools, in order to build formalisms applicable to the more interesting hard problems involving information.

The paper is organized as follows. Section 2 presents the details of the paradox together with its generalization as described by Conway. It is argued that the termination proof is in fact based upon some a priori analysis of the game. Section 3 introduces the general epistemic models introduced by Groenendijk and Stokhof, and indicates the additional restrictions which have to be satisfied by epistemic models in order for them to be useful in the analysis of the paradox. In the models there occurs a modal, possible world component which is used for expressing information about logical and factual relations between states of affairs, information about them, etc. This component is applied in Section 4 where a model for the initial state of the game is obtained. Still the resulting model is not sufficiently general, since it does not allow for the representation of the

extra information conveyed by a "no" answer from one of the players. This missing feature is added by defining an update operator, which transforms the entire structure into a new one. This operator is introduced in Section 5, and it is shown by means of an explicit example that, starting from the initial state as defined in Section 4, after finitely many updates a new state is obtained where the game will terminate. Section 6 contains some concluding remarks.

2. THE PARADOX

2.1. Termination and non-termination proof

Consider a particular instance of the two person game with numbers which are not too small, such as the game (3,4). Given the fact that the two numbers are adjacent, each player can find out that the parity of his number is opposed to the parity of the number which he sees on the forehead of his opponent, so he knows whether his own number is even or odd. Let us name the players with the even and the odd number 'Eve' and 'Ott', respectively. Each player will know at the start of the game which role he or she is playing. Now we can easily indicate why the two person game (3,4) will never terminate. It can be argued that the first two answers in the game will be "no", and moreover that both players can predict this. This shows that virtually no information is exchanged during the first round, so why play this round at all?

The argument is based upon the possibilities which both players can discern at the outset. First consider Ott. He sees his opponent carrying a 4, so he knows he must have a 3 or a 5. He does not have complete information and, if asked whether he knows his number, he can only answer "no". Moreover, in both cases the information about Ott's number which is visible to Eve, will be of no help for *her* to solve *her* problem: if Eve sees a 3 she will hesitate between 2 and 4, whereas seeing a 5 will make her hesitate between 4 and 6. Since Ott knows that these are the only two possibilities, Ott is sure that the first answer given by Eve has to be "no" as well. Note moreover that the fact that Eve actually says "no" does not convey any new information to Ott, since he knew at the outset that this was the only possible answer in the given circumstances. Ott can also figure out that a "no" answer given by himself, before Eve has had to answer, won't help Eve in solving her problem, since he knows that Eve is clever enough to infer that

Ott must say "no", regardless of whether he in fact has a 3 or a 5. Adding all this up, we conclude that, regardless of who begins, Ott is sure that the first two answers in the game will be "no".

Now consider the situation for Eve. By the same argumentation as above (where the values of all numbers have to be decreased by 1), we may infer that Eve knows as well that the first two answers in the game will be "no", regardless who begins. This indicates that there does not happen anything interesting during the first round: there is no exchange of new information, and during the next and all subsequent rounds the situation will be the same - the game does not terminate.

Next we show that the game always terminates by proving the following

THEOREM. *The game* (x,x+1) *is terminated at move* x+1 *by the player having the highest number in case the player with the odd number starts, and at move* x+2 *otherwise.*

PROOF. Note that from the fact that it is given that the two numbers are adjacent, each player is able to infer whether his number is even or odd. We prove the result by induction, keeping track of the parity of x. (Again we denote the player with the even number by 'Eve' and the other player by 'Ott'.)

Base induction proof, x=0.

In this situation Ott has complete information (since he sees a 0), whereas the information of Eve is incomplete (she may have a 0 or a 2). So Ott will terminate the game as soon as his turn is up; this is at move 1 in case Ott starts and at move 2 otherwise. This proves the result for x=0.

Induction step, x=2k.

In this situation Ott has the highest number. He knows that his number equals 2k-1 or 2k+1. If the first holds, by induction Eve will terminate the game at move 2k, in case Ott starts, and at move 2k+1 otherwise. As soon as Ott finds out that this has not happened (which situation arises at move 2k+1 or 2k+2, respectively) he can terminate the game, which proves the result for x=2k.

Induction step, x=2k+1.

In this situation Eve has the highest number. The possibility that her number equals 2k is ruled out by the behaviour of Ott at turn 2k+1 or 2k+2, respectively, depending on whether Ott or Eve starts. So at the next move, which is move 2k+2 or 2k+3, respectively, Eve can terminate the game. This proves the result for x=2k+1.

The structure of this proof is illutrated by the diagram below. It shows a graph whose vertices are legal configurations in the game. Two configurations are connected by an edge labeled X when player X cannot discriminate the two positions on the basis of his/her visible information. For example, the games (5,6) and (6,7) are connected by an edge labeled 'Ott', since Ott, seeing a 6, cannot decide whether he has a 5 or a 7.

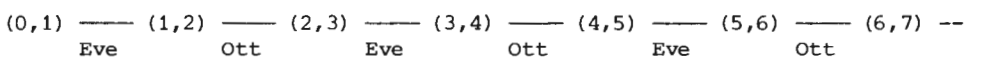


Diagram 1

A graph describing the two person game

Note that each game in the graph as two incoming edges, labeled 'Eve' and 'Ott', respectively; the only exception being the game (0,1). This game has no edge with label 'Ott' which connects it to another game, illustrating the fact that Ott can terminate this game at his first move.

In the induction proof presented above the players are supposed to perform the following "edge-cutting" game: whenever some player says "no", both players take their copy of the above graph and remove from it all nodes which have no incoming edge labeled by the player saying "no", together with all their incoming edges. After this reduction of the graph, both players consider the collection of games which remains and investigate whether the collection of games which are compatible with the real state of the world is reduced to a singleton, in which case the player can terminate the game at his next move.

The main weakness of the termination proof now can be explained as follows: it seems that, in order to terminate the game at all, the players are supposed to stop playing the real game, and start playing the edge-cutting game instead. So an *a priori analysis* of the game is added to the set of rules of the game supposed to be known to both players. Moreover, this knowledge is known to the other player as well, up to each level of epistemic analysis. Presumably the assertion that the players are "logically perfect" has to be interpreted in such a way that they independently have arrived at the solution just given, before the game starts. We consider this an unreasonable assumption.

At the same time it is easy to accuse the non-termination proof of being a prime example of a *proof by intimidation*: the rhetoric question: "Why play this round at all?" obscures the fact that we did not analyze all

possible information of the "A knows that B knows that C knows that ..." type which may play a role.

The present paper aims at developing an epistemic model in which information at all epistemic levels can be represented and which, moreover, obeys the rules of the game. We do not want to build into the model an a priori analysis of the game which tells in advance *which* conceivable position is removed at *which* move in the game. Instead, we want an update operator which removes from the structure those positions which are incompatible with a "no" answer given by a player, but which does this independently of at *which* move in the game this "no" answer is given.

2.2. Generalizations

Littlewood, in LITTLEWOOD (1953), presents two variants of the paradoxical situation described above. He considers cards which are showing two adjacent, non-negative integers on the two sides. The two players are seated opposite to each other. A third player (the umpire) draws a card and puts it between the two players in such a way that each player can only see one of the faces. The player having the highest number wins the round. However, each player has the right to cancel the round, so the first thing the umpire has to do is ask the two players whether they will play or whether they want to go to the next round by asking for a new card. Littlewood claims that by an induction proof it can be shown that all rounds are vetoed by some player.

In the other version the cards are drawn from an urn, containing a single copy of the card (0,1), 10 copies of the card (1,2), 100 copies of the card (2,3), etc. One can prove that under these circumstances each player has a change of one against ten of losing. This latter version brings us back to older paradoxes involving probability notions, which can be solved by basing probability theory on measure theory.

Conway has generalized the paradox in CONWAY et al. (1977) by considering games with more than two players. In this generalization there are $k \geq 2$ players, each carrying a number on his forehead. The players are seated in such a way that each player can read all numbers except his own. Moreover, there is an umpire, who has written a list of m consecutive numbers on a blackboard, one of which is the sum of the numbers on the foreheads of the players. We denote an instance of such a game by $(n_1,n_2,\ldots,n_k \mid p_1,p_2,\ldots,p_m)$, where $n_1,n_2,\ldots,n_k$ are the numbers written on the foreheads of the players A,B,..., respectively, and the numbers $p_1,p_2,\ldots,p_m$ the numbers written on the blackboard. The umpire asks the

players in cyclic order whether they know their number or not, and the game ends on the first "yes" answer.

By analysing a game such as (1,1,1|3,4,5) it can be made clear that in this game the first three answers will be "no", regardless of who begins, so again non-termination is proved by asking what possible use such a round could have. On the other hand, Conway has shown by a nice induction proof, that the edge-cutting variant of this game will terminate for an arbitrary initial position, as long as the number of values on the blackboard m does not exceed the number of players k.

We illustrate the termination by illustrating the edge-cutting variant of the above game (1,1,1|3,4,5) in the diagrams 2 and 3 below. Vertices in the game are all positions sharing the public information, i.e. the values of the numbers written on the blackboard $p_1,\ldots,p_m$. In our example these are the numbers 3,4,5. A node is therefore completely determined by a triple n_1,n_2,n_3 with sum equal to 3, 4, or 5. Two positions only differing with respect to the value of n_1,n_2,n_3, respectively, are connected by an adge labeled A,B,C, respectively, indicating that these two positions can not be discriminated by player A,B,C, respectively. It is possible to embed the resulting graph in three-dimensional space in such a way that the three orthogonal directions correspond to the three edge labels - the diagram gives a plane projection of this embedding: hence the label of an edge is determined by its direction in the diagram, as indicated in the "tripod" shown above the graph.

As before each player removes at his turn those vertices not having (or no longer having) incoming edges labeled by his color; these positions correspond to configurations where he has complete information - a "no" answer denies existence of such a configuration, and the configuration is therefore removed from the graph. Diagram 3 shows for each node the number of the move at which it will be removed in the edge-cutting game. Termination of the game is equivalent with the fact that each node sooner or later gets numbered.

The reader may convince himself that it is necessary for the proof to work that the players' behaviour is competent. During move 9 player C will remove node 211 together with the A-edge connecting it to the considered actual game 111. If A fails to terminate the game at move 10 by answering "yes", the four nodes numbered 10 will be removed and the graph will become empty, representing the situation where the game gets blocked.

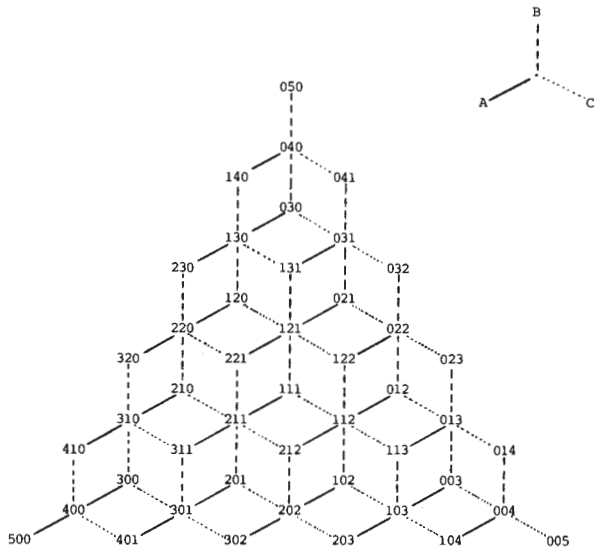


Diagram 2

Graph of possible games (a,b,c| 3,4,5). The label of an edge is determined by its direction.

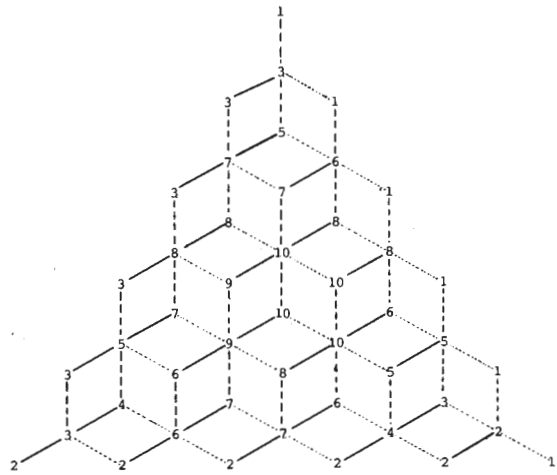


Diagram 3

Edge-cutting game for the graph of Diagram 2. A is the first to move. The numbers indicate at which move the position becomes incompatible with a "no" answer.

3. EPISTEMIC FRAMEWORK

3.1. Preliminaries

An epistemic model should not only encode the state of the actual world, but also the information that individuals in this world have about that state of the world and about the information of other individuals about the world or information of other individuals, etc. Disregarding psychological limits inherent to the human mind, this formulation leads to rather complex, infinite structures. Groenendijk and Stokhof have introduced a set theoretic framework for representing this kind of information, which we will describe briefly. But first we define some mathematical tools.

For A, a (finite or infinite) set, we define inductively the sequence of sets A^i by:

$$A^0 := A, \quad A^{k+1} := P_f(A^k)\backslash\{\emptyset\},$$

where P_f is the finite powerset operation.

A^+ denotes the disjoint union of all sets A^i, i = 0,1,... . This union is called a *graded set*.

If f: A → B is a mapping then we can define a mapping f^+: $A^+ \to B^+$ by defining a sequence of functions f^k: $A^k \to B^k$ inductively by

$$f^0 := f, \quad f^{k+1}(w) := \{f^k(x) \mid x \in w\},$$

and letting f^+ be the disjoint union of the sequence f^i. We call f^+ a *graded mapping*.

Consider the following example. Let A = {0,1} be the set of truth values. We can take for f the operation ¬ (negation). Then the operation $\neg^+$ is defined by

$$\neg^+(x) = \neg(x) \qquad \text{, if } x \in A^0,$$

$$\neg^+(w) = \{\neg^+(x) \mid x \in w\}, \quad \text{otherwise.}$$

So, since A = {0,1}, $\{\{0,1\},\{0\}\} \in A^2$, and $\neg^2(\{\{0,1\},\{0\}\}) = \{\neg^1(\{0,1\}),\neg^1(\{0\})\} = \{\{\neg(0),\neg(1)\},\{\neg(0)\}\} = \{\{1,0\},\{1\}\}$.

3.2. Some sideremarks

Before going on, we will make some brief remarks on the various ways in which graded mappings in more arguments can be defined. What follows is not essential to the paper and may safely be skipped.

The operator + introduced above actually yields a functor from the category of sets and mappings to the category of graded sets and graded mappings. If this functor behaved in a certain way with respect to Cartesian products, this would lead to a simple theory for functions with more than one argument. This turns out not to be the case. There are two ways to extend functions in more arguments. First of all, one can simply apply the functor + to the mapping $f: A\times B \to C$, yielding a graded mapping

$$f^{+}: (A\times B)^{+} \to C^{+}.$$

Note however that $(A\times B)^{+} \neq A^{+} \times B^{+}$, the latter object being the Cartesian product of A^{+} and B^{+} in the category of graded sets. Let us call $A^{+}\times B^{+} := (A\times B)^{\%}$. We can define the graded mapping $f^{\%}$ to be the union of the mappings $f^{\%i}: A^{i}\times B^{i} \to C^{i}$, where $f^{\%i}$ is defined inductively by:

$$f^{\%0} = f, \quad f^{\%i+1}(U,V) := \{f^{\%i}(u,v) \mid u \in U,\ v \in V\}.$$

It is clear from the contents of GROENENDIJK & STOKHOF (1980) that these authors intended to use the construction % for products rather than the functor +. It can also be seen by considering small examples that the functor + does not preserve products (taking the union of products of the component sets as a definition of product in the category of graded sets, as suggested by the definition of %). The connection between the operations + and % for products is illustrated by Diagram 4.

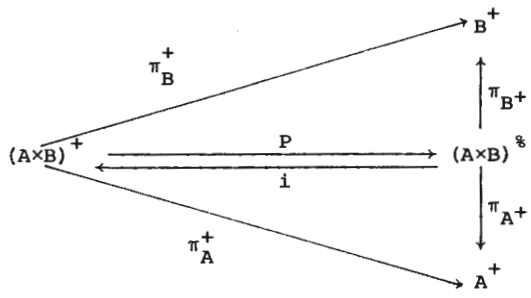


Diagram 4

The embedding i is obtained as $id^{\$}_{A\times B}$, whereas the projection p is obtained from the pair of mappings π^{+}_{A} and π^{+}_{B}, using the fact that $(A\times B)^{\$}$ is the product of A^{+} and B^{+} in the category of graded sets and graded mappings. A straightforward induction proof shows that $p(i) = \underline{id}_{(A\times B)^{\$}}$.

This non-preservation of products by the functors described above, has, as Groenendijk and Stokhof have observed, as one of its consequences that some logical laws concerning the usual logical connectives are no longer valid at higher levels. This matter will not be pursued further in this paper.

3.3. Epistemic models

Returning to the topic of this paper, we are now in a position to define the notion of an *epistemic model*, using the tools defined in 3.1.

If Σ is a finite alphabet, then we let Σ^{*} denote the set of finite strings over Σ, letting ε denote the empty string. The length of a string s is denoted by $|s|$.

DEFINITION. A *general epistemic model* is a quintuple $\langle L,\Sigma,W,A,V\rangle$, where

L represents a *language*, the elements of which are to be interpreted,

Σ is a finite alphabet, the elements of which are called *conscious entities* or *persons*,

W is a set of *possible worlds* (its role will become clear in the next section),

A is some *domain* of interpretation for the elements of L,

V is the *interpretation function*; it is a mapping $V: L\times\Sigma^{*}\times W \to A^{+}$, such that $V(f,s,w) \in A^{|s|}$.

The intended meaning of the valuation function V is expressed as follows:

$V(f,\varepsilon,w) = a$ means: in world w the interpretation of f equals *a*

$V(f,Zs,w) = \{q_1,\ldots,q_r\}$ means: in world w person Z has the information that one out of the r possibilities expressed by $V(f,s,w) = q_i$ is the case, but Z does not have the information which one of these possibilities is in fact the case.

As usual in semantic frameworks, the mapping V is required to obey the so-called Fregean principle of compositionality, which expresses that the meaning of a compound expression is a function of the meanings of its constituent parts (see for example VAN EMDE BOAS & JANSSEN (1979) for a discussion of this principle). The framework as it was originally proposed in GROENENDIJK & STOKHOF (1980), obeys this principle. In the present paper compositionality is not under discussion, since the language considered consists of just two atomic expressions.

For the remainder of this paper we stipulate the following:

$L = \{\underline{X},\underline{Y}\}$ (representing the numbers of players X and Y, respectively),

$\Sigma = \{X,Y\}$ (representing the players X and Y, respectively),

$A = \mathbb{N}$ (the set of natural numbers including 0).

As an example consider the assertion expressed by the formula $V(\underline{X},YX,w) = \{\{1,3\},\{3,5\}\}$. This assertion states that in world w, player Y has the information that X is hesitating about his own number; according to Y, X is either doubting between 1 and 3 or doubting between 3 and 5, but Y does not know which of these two possibilities is in fact the case. This assertion describes a situation which arises in the two person game when X actually has the number 3 on his head. In this situation Y will hesitate whether his number is 2 or 4 and accordingly he will attribute to X corresponding hesitations about his own number: hesitation between 1 and 3 in case Y has a 2, and hesitation between 3 and 5 in case Y has a 4.

3.4. Restricting epistemic models

The kind of epistemic models covered by the definition given above are still much too general. E.g., it is not required at all that the information of various persons is connected in a reasonable way. Nor is it required that the information reflects knowledge of the rules of the two person game. These requirements can be enforced by adding further conditions which the valuation function V has to satisfy. The first condition expresses that if a person X has certain information, he also has the information that he has this information. Moreover, it is known at each level in the epistemic framework that all persons fulfill this requirement. In order to express this so-called *optimal information principle*, we need a further operator defined on the set A^+.

Let $\uparrow_i$ be the operation $A^i \to A^{i+1}$ defined by $\uparrow(U) := \{U\}$. This operation may be extended to a mapping from $\bigcup_{j=i}^{\infty} A^j \to A^+$ in the usual way. Note

that the operation obtained in this way, which we denote by $\uparrow_i^+$, does not preserve the grading of the set A^+; in fact, it increases its grade by one. Note also that for $i < j \leq k$ both $\uparrow_i^+$ and $\uparrow_j^+$ are definedon A^k, but that their effect is different. For example, $\uparrow_1(\{0,1\}) = \{\{0,1\}\}$ and $\uparrow_0(\{0,1\}) = \{\{0\},\{1\}\}$.

The *optimal information principle* now can be expressed as follows: for all $w \in W$, $f \in L$, $Z \in \Sigma$ and s_1 and $s_2 \in \Sigma^*$ it holds that

$$V(f,s_1ZZs_2,w) = \uparrow_i^+(V(f,s_1Zs_2,w)),$$

where $i = |s_2| + 1$. So from $V(\underline{X},Y,w) = \{0,2\}$ we may infer that $V(\underline{X},YY,w) = \{\{0,2\}\}$. Similarly, from $V(\underline{X},XY,w) = \{\{1,3\},\{3,5\}\}$ we obtain $V(\underline{X},XYY,w) = \{\{\{1,3\}\},\{\{3,5\}\}\}$ and $V(\underline{X},XXY,w) = \{\{\{1,3\},\{3,5\}\}\}$. As a result, by assuming the *optimal information principle*, we can specify V completely restricting ourselves to values of V with respect to strings s without iterated symbols. For our two-element alphabet Σ this implies that we only have to look at alternating strings. We denote this set of strings by Σ_a^*; it may be defined by

$$\Sigma_a^* := (\varepsilon+Y)(XY)^*(\varepsilon+X),$$

where we have used the terminology of regular expressions. In the sequel we shall only consider strings from Σ_a^*.

Our next condition represents the rule of the two person game, that the two numbers $\underline{X}$ and $\underline{Y}$ are adjacent and the fact that this is known to both players at all epistemic levels. This leads to what will be called the *adjacency conditions*. The definition requires another operator S^+. Let $S: \mathbb{N} \to \mathbb{N}^1$ be defined by $S(0) := \{1\}$, $S(k+1) := \{k,k+2\}$. So S maps each positive number on the pair consisting of its neighbours and maps the number 0 on the singleton consisting of its only positive neighbour 1. We extend S to a mapping $S^+: \mathbb{N}^+ \to \mathbb{N}^+$ in the usual way. Again the mapping S^+ increases the grading by one.

The *adjacency conditions* are going to express the following facts about our two person game:

(0) The actual state is a legal position of the game.

(i) Each player sees (and consequently knows) the number of the other player.

(ii) Each player knows his number to be a neighbour of the number of his

opponent.

(iii) These facts (i) and (ii) are known by each player at all epistemic levels.

The mathematical formulation of the *adjacency conditions* reads as follows: for each world $w \in W$, and each alternating string s not ending on Y, and t not ending on X, the following relations hold:

(0) $V(\underline{X},\varepsilon,w) = k$, $V(\underline{Y},\varepsilon,w) = \ell$ for adjacent, non-negative k and ℓ.

(1) $V(\underline{X},sY,w) = \uparrow_0^+(V(\underline{X},s,w))$ "Y knows $\underline{X}$"

(2) $V(\underline{Y},sY,w) = S^+(V(\underline{X},s,w))$ "Y knows $\underline{Y}$ to be a neighbour of $\underline{X}$"

(3) $V(\underline{Y},tX,w) = \uparrow_0^+(V(\underline{Y},t,w))$ "X knows $\underline{Y}$"

(4) $V(\underline{X},tX,w) = S^+(V(\underline{Y},t,w))$ "X knows $\underline{X}$ to be a neighbour of $\underline{Y}$".

Together with the *optimal information principle* the five equations above allow us to compute for every actual configuration in the game the values of the valuation function for $\underline{X}$ and $\underline{Y}$ with respect to every string s. In Diagram 5 we illustrate this for the configuration where $\underline{Y} = 3$ and $\underline{X} = 2$. In this diagram we let F(f,s) denote V(f,s,w), since w is fixed.

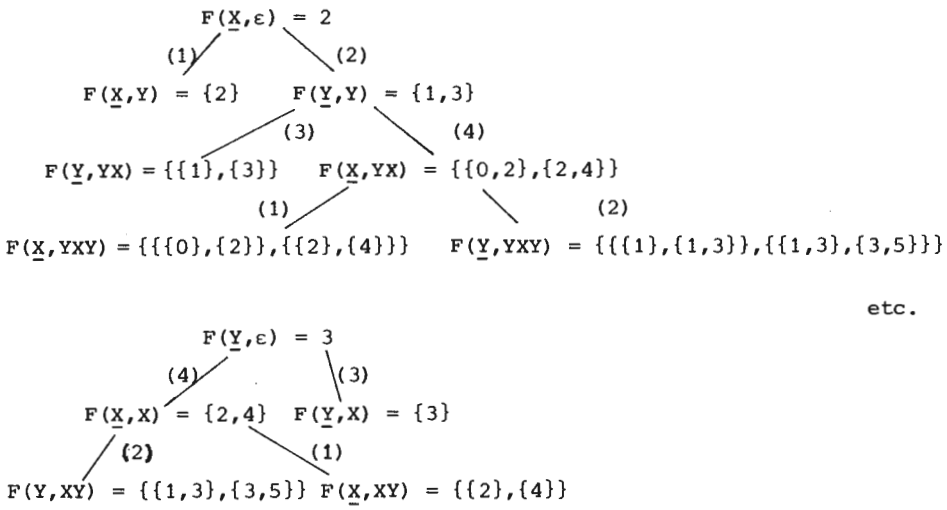


Diagram 5

Information computed in accordance with the adjacency conditions

Note that for an index s which ends on Y the values at the deepest level in the set F(Y,s) are almost never singletons. This relates to the fact that Y is uncertain about the value of his own number. Note, however,

the exception in the example given above: in the computation of F(Y, YXY) there occurs a single singleton {1} at the deepest level. This singleton indicates a configuration of complete information which is going to be denied by a "no" answer from Y.

Another remark we can make at this point is that the epistemic model, although it properly encodes the information the participants may have in some configuration of the game, does not account for connections between the possibilities arising at different levels. For example, the pair {2,4} in the expansion of F(X,YX) comes from the number 3 in F(Y,Y) and not from the number 1 in the latter set. Our model so far does not yet represent this part of the information which we shall need in order to complete our analysis of the paradox.

4. MODAL EPISTEMIC FRAMEWORK

4.1. The role of the possible world component

The model as described in Section 3 represents a large part of the information of the participants in the game. However, certain connections between pieces of information are not accounted for. Generally speaking, such connections represent information about logical and factual relations between states of affairs, information about them, etc. A representation of this kind of information is an essential part of a theory in which pragmatic phenomena concerning the information of language-users are to be handled. An example of information about a factual relation between possible situations in our two person game is the following.

In the situation where X is hesitating whether X equals 2 or 4, he infers that at the same time Y must be hesitating either whether Y equals 1 or 3, or whether Y equals 3 or 5. The first possibility is connected to the value 2 for X, whereas the second corresponds to the possibility that X = 4. The information which the participants have about these connections has not been accounted for in our model so far.

In order to represent this kind of information we will use the possible world component in our general epistemic model. Given some state in the game, which will be called the *actual world* hereafter, each player can create new possible worlds for himself where he has made *hypothetical choices* between the possibilities available to him. Moreover, he can imagine within such a world the other player making similar choices, etc. up to

every level of our analysis. Since for our particular game there exist at each level at most two possibilities between which a player can choose, this leads to a structure which has the form of a (pair of) binary tree(s). The nodes of such a tree are labeled by possible worlds, described by their V-values, and its edges are labeled by strings which indicate which player created this hypothetical situation. The logic of the game is represented by the fact that, up to some particular level, the V-values are obtained by hypothesis formation (i.e. explicitly assumed by the involved conscious entity), whereas below this level the V-values are again computed using the optimal information principle and the adjacency rules. Implicitly we enlarge the collection of legally possible worlds through addition of these so-called s-extensions, to be described in more detail shortly. First, we define some more mathematical tools.

4.2. Formal implementation

Let A be some set and let q be a member of A^+. We say that q is a 1-singleton iff q is a singleton, and we say that q is a k+1 singleton for $k > 0$ iff q is a singleton whose only member is a k-singleton. We denote this property by k-sgl(q). If q is a k-singleton then its only element at level k is denoted $[q]^k$. So if $q = \{r\}$ then $[q]^k = [r]^{k-1}$.

Now let w_0 be a possible world in an epistemic model satisfying the adjacency conditions such that $V(\underline{X},\varepsilon,w_0)$ and $V(\underline{Y},\varepsilon,w_0)$ are two adjacent non-negative numbers y and y+1. With respect to string X we have $V(\underline{X},X,w_0) = \{y,y+2\}$ and $V(\underline{Y},X,w_0) = \{y+1\}$. The values of $V(f,s,w_0)$ for s starting with X are computed from these values in accordance with the adjacency conditions.

We can introduce two possible words w_1 and w_2 such that

(i) $V(\underline{X},\varepsilon,w_1) = V(\underline{X},\varepsilon,w_2) = V(\underline{X},\varepsilon,w_0)$ and similarly for $\underline{Y}$;

(ii) $V(\underline{Y},X,w_1) = V(\underline{Y},X,w_2) = V(\underline{Y},X,w_0)$;

(iii) $V(\underline{X},X,w_1) = \{y\}$, $V(\underline{X},X,w_2) = \{y+2\}$;

(iv) for other strings starting with X the values of V are computed in accordance with the adjacency conditions starting from (ii) and (iii).

The worlds w_1 and w_2 are called the elementary X-extensions of w_0. Note that we do not require anything about the values of V in the extensions with respect to strings starting with Y, but for definiteness we preserve the values at w_0. The worlds w_1 and w_2 are hypothetical situations in the mind of X and information available to Y is completely unrelated to these

worlds, so it makes no difference at all what is postulated concerning the values of V with respect to strings starting with Y.

Assume that we have already defined the elementary s extensions for s starting with X of length ≤ j. Let s' = XYXY... be string of length j+1, s the string of length j resulting by removing the last element of s', and let w be one of the extensions of w_0 with respect to s. By induction hypothesis the following conditions are fulfilled:

(a) for s" starting with X and length j"≤ j it is the case that V(f,s",w) is a j"-singleton q such that $[q]^{j"}$, its only element at level j", occurs $\underbrace{\text{as an element in an element in an element .....}}_{\text{j"-times}}$ in V(f,s",w_0).

(b) for f = either $\underline{X}$ or $\underline{Y}$ (depending on the parity of j) it is the case that V(f,s',w) is a j+1-singleton, whereas for the other it is a j-singleton q with $[q]^j$ possibly being a pair.

(c) For strings t longer than j+1 the values of V(f,t,w) are computed in accordance with the adjacency conditions starting from the values mentioned sub (a) and (b).

The s'-extensions of w are constructed as follows:

(i) for strings up to length j the values are equal to those in w; the same holds for s' and the expression $\underline{X}$ or $\underline{Y}$, whichever yields a j+1-singleton as mentioned sub (b).

(ii) for string s' and the remaining expression $\underline{X}$ or $\underline{Y}$ the value is a j+1-singleton q' with $[q']^{j+1}$ obtained by making a choice among the members of the pair mentioned sub (b).

(iii) For longer strings the values are obtained by computation in accordance with the adjacency conditions starting from the values obtained sub (i) and (ii).

The collection of s'-extensions of w_0 is obtained by performing the above construction for each s-extension of w_0. Since each s-extension yields at most two s'-extensions the system of s-extensions for strings starting with X results in a binary tree structure. The binary tree, called the *X-tree*, represents the information available to X at the initial state of the game, together with all possible hypothetical situations which X can conceive and which might have led to the situation as it is observed. The structure of the tree makes explicit the connection between hypotheses at various levels.

A similar construction can be performed for indices starting with Y.

In Diagram 6 below we give an example of a part of the (infinite) Y-tree labelling all nodes with partial information about the values of V at these nodes; only the most relevant part of the information is presented, from which the other values can be computed easily. A pair of such trees, an X-tree and Y-tree, models the initial state of the game.

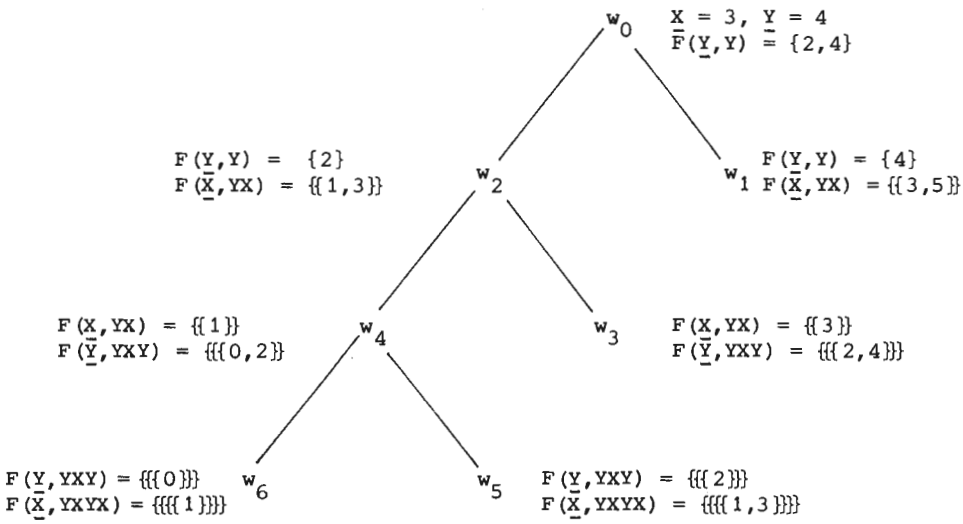


Diagram 6

Fragment of the Y-tree for the initial state of the game (3,4)

5. UPDATING THE STATE

5.1. What updating comes to

Consider the representation of the initial state of the game where $\underline{X} = 2$ and $\underline{Y} = 3$. It follows that $V(\underline{X},X,w_0) = \{2,4\}$, i.e. X is uncertain about his number. Similarly it follows that $V(\underline{Y},Y,w_0) = \{1,3\}$, i.e. Y is uncertain as well. So both players will answer "no" when asked whether they know what their number is. Further, it holds that $V(\underline{Y},XY,w_0) = \{\{1,3\},\{3,5\}\}$. This means that X knows that Y is hesitating between two possible values,

although X, at his turn, is hesitating about which pair. So X knows in advance that Y will answer "no". The same holds for Y.

In order to have any progress in the game it is necessary that the players use the information conveyed by a "no"-answer being given for updating their information about the state of the game. If the players don't use this information nothing will change and the game will last forever. But how is the information conveyed by a "no"-answer to be used? Once X or Y has answered "no", it may be assumed that both players know that this answer has been given, and that they know that the other will know so as well, etc. The information must be used for ruling out hypothetical extensions of the actual world in which the player who has given the "no"-answer has the kind of complete information which he just denied to have. Note that the s-extensions of the actual world constructed in the preceding section are hypothetical situations in which the players have more information than they have in the actual world - they were constructed in that way. In some of these a player has complete information. Often this fact is the direct outcome of a choice between alternatives. But there are some worlds in which this is not the case. In these worlds the fact of complete information is not simply chosen from the alternatives, or to put it differently, it is not enforced by extending the choice that created the world upto the corresponding level.

Consider world w_6 in Diagram 6 in the preceding section. In this world choices have been made upto level 3. In this situation Y knows that X knows that Y knows the following remarkable fact: "X knows that $\underline{X} = 1$", and this instance of complete information was not created by choice-expanding upto level 4. It is the existence of such a world which is denied by the fact that, after X says "no", Y knows that X knows that Y knows that X has said "no". So w_6 no longer should be considered to be a possible world. Moreover, the possibilities higher up in the tree which led to its creation in the tree of extensions should be removed as well. This task has to be performed by an update operator which we shall now define.

5.2. The update operator

The actual world w is called a *world with complete information for Y* iff $V(\underline{Y},Y,w)$ is a singleton. Similarly for X. Let s be a string of length k ending with X, and let w' be some s-extension of world w. We say that w' is a *world with complete information for Y* iff $V(\underline{Y},sY,w')$ is a k+1-singleton.

Similarly, if s ends with Y and $V(\underline{X},sX,w')$ is a k+1-singleton then w' is a *world with complete information for X*.

In the game the answer given by a player will be "yes" if the actual world is a world of complete information for that player, and "no" otherwise. Consider the binary tree representing the information of Y, consisting of some world labelling the root (called the actual world) together with all s-extensions for strings s starting with Y. In order to represent the configuration which occurs after X says "no", we introduce the update operator \$X, which modifies the tree in the following way:

(i) all words in the tree which are worlds with complete information for X are removed, together with all their descendants;

(ii) if some world w" at level k (the level of the root being 0) is removed from the tree, the information present in this world is k-extracted from the information in all worlds on the path from the actual world to w";

(iii) the resulting tree with updated information forms a new tree consisting of an actual world at the root together with its s-extensions for indices s starting with Y.

The operation of *k-extraction* used in clause (ii) above is defined as follows: let w" be a world which is removed at level k and let w' be some ancestor at level $k_1 < k$. Then w' is replaced by a new world w^* such that

$$V(f,s,w^*) = V(f,s,w') \qquad \text{if s is of length} < k,$$
$$V(f,s,w^*) = V(f,s,w') \setminus_k V(f,s,w'') \qquad \text{otherwise,}$$

where the operator $\setminus_k$ is defined by:

$$A\setminus_1 B := A\setminus B, \quad A\setminus_{j+1} B := \{a\setminus_j b \mid a \in A, b \in B\} \quad \text{for } j \geq 1.$$

A similar definition can be given for updating the Y-tree after Y has said "no", yielding an operator \$Y. Analogous definitions are required in order to explain how the operators \$X and \$Y modify the X-tree. Note that the actual world occurs in both trees: in order to have it updated properly the values of $V(f,s,w_0)$ are modified according to the definition for the X-tree for indices starting with X and according to the definition for the Y-tree for indices starting with Y.

We now have developed all tools needed for calculating the termination of our game. The calculation consists of two stages:

stage 1: By computing the values in accordance with the adjacency conditions a world describing the initial state of the game is defined. This world w_0 becomes the root of both an X-tree and a Y-tree which are constructed according to the methods described in Section 4.

stage 2: If it is X's turn to answer, we inspect whether the actual world is a world with complete information for X. If so, the game terminates; otherwise the operator $X is performed on both the X- and the Y-tree. Similarly, if it is Y's turn to answer. Next stage 2 is repeated.

We illustrate by an example that the calculation, starting from the situation described by Diagram 6 shown at the end of the preceding section, terminates after three answers, assuming that it is X who begins.

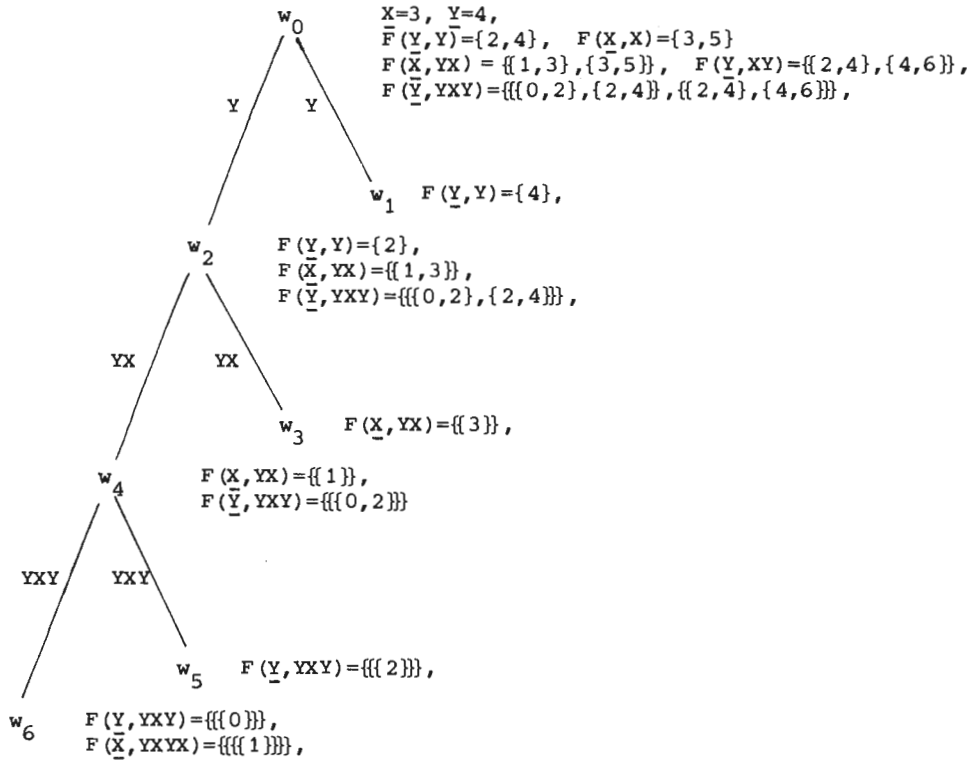


Diagram 7

Initial state: X says "no"; w_6 is a world with complete information for X; the information presented in w_6 is 3-extracted from the tree.

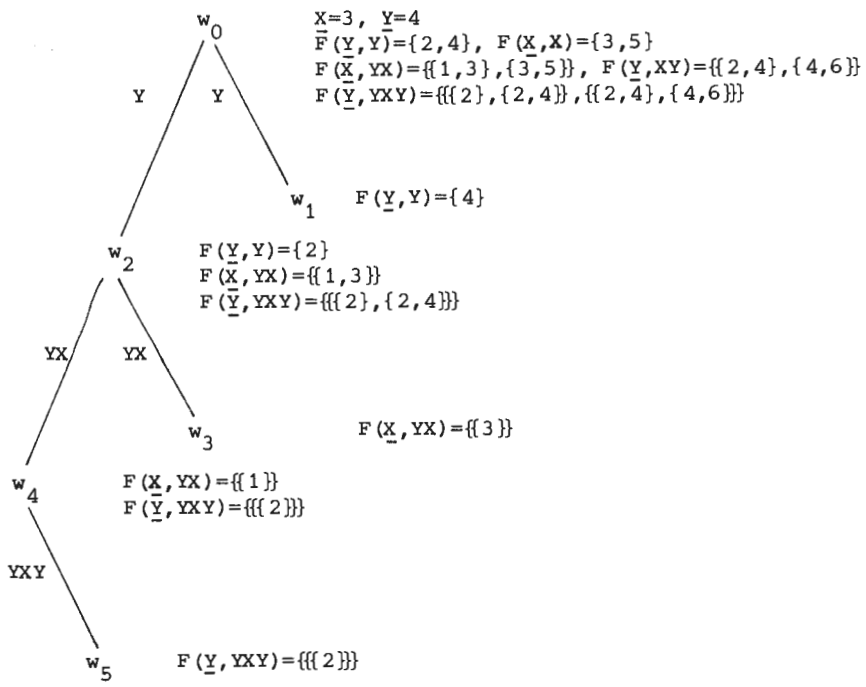


Diagram 8

Stage after X's "no" answer; Y says "no"; w_4 is a world with complete information for Y; its information is 2-extracted from the tree.

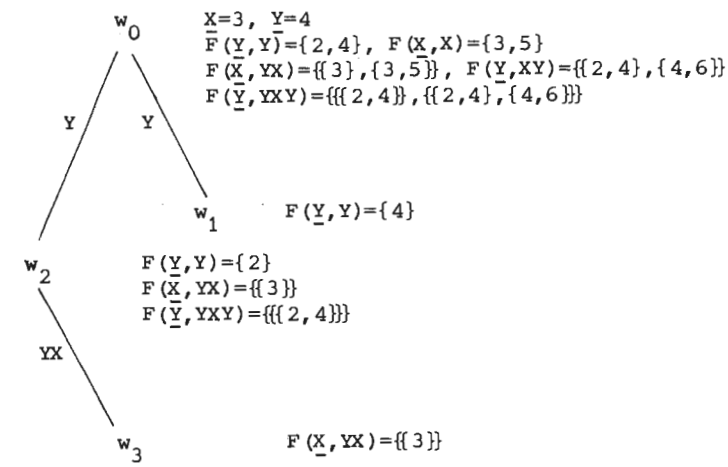
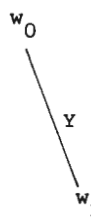


Diagram 9

Stage after Y's "no" answer; X says "no"; w_2 is a world with complete information for X; its information is 1-extracted from the tree.



$\underline{X}=3$, $\underline{Y}=4$
$\overline{F}(\underline{Y},Y)=\{4\}$, $F(\underline{X},X)=\{3,5\}$
$F(\underline{X},YX)=\{\{3\},\{3,5\}\}$, $F(\underline{Y},XY)=\{\{2,4\},\{4,6\}\}$
$F(\underline{Y},YXY)=\{\{\{4\},\{4,6\}\}\}$

$F(\underline{Y},Y)=\{4\}$

Diagram 10

Stage after X's second "no" answer; w_0 is a world with complete information for Y, so Y says "yes" and the game terminates.

Note that in Diagram 10 the update on $F(\underline{Y},YXY)$ in the actual world is the combined result of a 1-extraction of the information at world w_2 in Diagram 9, together with a 3-extraction of a world with complete information two levels below w_3 (which is not shown in the diagram). This illustrates that indeed the entire tree has to be updated at infinitely many places at once, in order for the computation to work out correctly. If we restrict ourselves to V-values with respect to strings of bounded length, the "active" part of the tree, which we have to keep track of, will be finite.

6. CONCLUSION

As shown in the preceding two sections, the mathematical model developed in this paper has the required property: the termination of the game in the simple situation can be derived by an explicit calculation which does not involve an a priori analysis of the entire game. On the other hand the machinery involved is rather cumbersome: a complete formal definition of the tree structures involved would probably require several pages densely filled with formulas, and a formal proof that the computation works as it should, will take many more pages without presenting any new insight. A possible way of proving such a claim might be to show that after k moves, after the first answer of the player with the highest number, all numbers less than k have disappeared from the trees, yielding a new situation which is isomorfic with the initial situation under the mapping $m \to m-k$. This claim can be proved by induction by showing that it is correct for a single move (disregarding the first move in the game in case this is a move by the player with the lowest number). The proof of this induction step will require a nice recursive description of the trees. Note that in each tree

there are infinitely many worlds with complete information since each node is ancestor of infinitely many worlds of this type at arbitrary distances. Therefore, the computation stages described in the previous sections actually require infinitely many steps, and at first glance, it is not at all clear that the resulting stage is always well-defined. It is conceivable that techniques for proving correctness of programs working on recursive data structures can be applied here.

If we consider the generalization of the formalism required for modelling the three-person game described by Conway, the combinatorial complexity increases strikingly: whereas the analysis given above only involves the linearly ordered chain of alternating strings, X, XY, XYX, ... , and Y, YX, YXY, ... , relevant strings in the three-person game itself form a tree, since there are two relevant ways of extending a string. For each path in this tree of strings a ternary tree of hypothetical extensions of the actual world has to be constructed. There will be some generalization of the adjacency conditions which have to be used for computing the initial structures. The update operator for processing a "no" answer probably will be more or less the same as the one presented in Section 5.

Our analysis disregards the question whether the termination of the game obtained corresponds to real human behaviour. One might argue that the model is "non-human". Consider again the tree as presented in Diagram 4 and consider world w_6. In this world, Y knows that $\underline{Y} = 2$, but on the other hand Y knows also that X is certain that Y knows that $\underline{Y} = 0$, but in fact $\underline{Y} = 4$! In this world the players not only use false hypotheses, but also hypotheses which they know by observation to be inconsistent with the real situation. In fact, they are required to disregard the real situation completely, i.e. they are required to act "as if" and to forget that they act "as if". After all it may therefore be the case that, from a psychological point of view, the non-termination argument corresponds to the real human situation, in particular for games (y,y+1), where y is sufficiently large (larger than 4 might already suffice). A similar conclusion might be obtained based upon complexity arguments. In order to terminate the game our analysis for the game (y,y+1) requires the players to develop the possible world trees up to level y at least. If one assumes that the human mind is incapable of dealing with information about information about information ... , at a level higher than three or four, these parts of the tree become inaccessible for human analysis and, consequently, the removal of worlds with complete information, which is necessary for the termination of the game, will never occur -

these worlds are too complex to be considered at all.

Clearly, the above remarks concerning human behaviour are highly speculative. However, the limit 3 or 4 is said to be reasonable by various colleagues during discussions held after talks given about the analysis presented. The reader is invited to amuse (or abuse?) his visitors at some future party be experimenting with the game, using his guests as victims. Such a test would at best affirm the existence of a limit value for y beyond which the game becomes non-terminating, without providing us with a precise explanation why this limit exists. Further psychological investigations will be needed in order to determine whether our model explains real behaviour or not.

From the above observations it now becomes clear how the paradox should be resolved; the conscious entities considered in the non-termination and termination proofs, respectively, are of different nature: humans versus robots.

REFERENCES

CONWAY, J.H., M.S. PATERSON & U.S.S.R. MOSCOW, 'A headache-causing problem', in: J.K. Lenstra et all. (eds), *Een pak met een korte broek; Papers presented to H.W. Lenstra, jr. on the occasion of the publication of his "Euclidische Getallenlichamen"*, Private publication, 1977, Amsterdam.

VAN EMDE BOAS, P. & T.M.V. JANSSEN, 1979, 'The impact of Frege's principle of compositionality for the semantics of programming and natural languages', in: D. Alexander (ed.), *"Begriffsschrift", Jenaer Frege-Konferenz, 7-11 Mai 1979*, Wissenschaftliche Beitrage der Friedrich-Schiller-Universität Jena, 1979, pp.110-129.

GARDNER, M., 1977, 'The "Jump proof" and its similarity to the toppling of a row of dominoes', Mathematical Games section, *Scientific American* 236, 128-135.

GROENENDIJK, J. & M. STOKHOF, 1980, 'A pragmatic analysis of specificity', in: F. Heny (ed.), *Ambiguities in intensional Contexts*, Reidel Publ. Co., Dordrecht 1980.

LITTLEWOOD, J.E., 1953, *A mathematicians miscellany*, Methuen & Co. Ltd., London.