האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

RATIONALITY AND COMPREHENSION

by

AVISHAI MARGALIT
and
MENAHEM YAARI

Discussion Paper # 40 February 1994

מרכז לחקר הרציונליות
וההחלטות האינטראקטיביות
CENTER FOR RATIONALITY
AND INTERACTIVE DECISION THEORY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-584135, [972]-2-584136
E-MAIL: ratio@sunrise.huji.ac.il
FAX: [972]-2-513681

RATIONALITY AND COMPREHENSION

by

A. Margalit and M.E. Yaari

ABSTRACT

Devising a Theory of Knowledge for interacting agents has been on many people's minds recently. A near concensus has emerged, that the appropriate framework is a multi-agent version of C.I. Lewis's system S5 or one of S5's standard weakenings. In this essay, it is argued that such a framework cannot possibly be adequate, if it is to capture the intricacies of genunine inter-agent epistemics. Introducing a notion of "comprehension" -- knowledge which is non-sensory yet non-analytic -- may possibly be a remedy.

RATIONALITY AND COMPREHENSION

by

A. Margalit and M.E. Yaari*

1. Introduction

Economists and game theorists have recently been devoting a great deal of attention to laying down a theory of knowledge for decision-making units. (We shall often refer to decision-making units as "agents".) What does the agent know about "the world"? What does the agent know about other agents? And finally, when can we say that a given fact, event or proposition is common knowledge among two or more agents? Some readers may find this kind of exploration rather curious, in view of the fact that "decision-making units" are theoretical entities. It should be noted, however, that the purpose of the exercise is normative rather than descriptive. Basically, the question being asked concerns the structure of knowledge being **required** for the agents, if this or that theory of interaction (e.g., Nash Equilibrium) is to be upheld. A detailed account of the recently developed models of knowledge for interacting agents may be found in Geanakoplos [4].

In this essay, we shall argue that the concept of knowledge that has emerged in these discussions is unreasonably narrow and confining. The framework which we criticize and which, we claim, lies at the basis of today's theories of knowledge for interacting agents, is characterized by the following three principles:

---

* *We are very much indebted to Kenneth J. Arrow, Kenneth G. Binmore and Amartya K. Sen for their comments. We also wish to thank Robert J. Aumann for giving us permission to refer to his unpublished work* [2].

- **Dichotomy.** There are two, and only two, types of knowledge: Factual and Analytic.
- **Factual = Sensory.** Knowledge is factual if, and only if, it is acquired through the receipt of a sensory signal. The agent's factual knowledge at any given state is defined to coincide with the sensory signal received by the agent at that state.
- **Logical Omniscience.** Analytic knowledge is innate, perfect, and complete. All tautologies and contradictions are known for what they are; knowing a tautology is itself a tautology; knowing p implies knowing all p's logical consequences; if p is not a contradiction, then p is known to be possible.

We regard these three percepts as a vestige of a naive Positivist orthodoxy which, for some reason, microeconomists and game theorists have chosen to embrace. Why some of today's finest economic theorists should submit to such an orthodoxy is a fascinating question in the sociology of science. Perhaps it has something to do with the cult of Free Market Economics which tends to view the economic agent as an organism relentlessly engaged in gratification-seeking. Be that as it may, our purpose here is merely to point out that the three principles stated above provide an inadequate basis for the theory of knowledge that is needed in the study of interaction among agents.

2. Interactive Epistemology

Among the many contributions made recently on this topic of knowledge in rational decision-making, there is as far as we know only one where an attempt is made to develop the decision-makers' information structures from basic principles. This is Robert J. Aumann's remarkable (unfortunately as yet unpublished) essay, **Notes on Interactive Epistemology** [2]. Aumann's construction uses the following building blocks:

(a) There is a (finite non-empty) set of **individuals**. Their number is n.

(b) A **language** is given, where the symbols are letters from a (finite or denumerable) alphabet, to which are added n+4 further symbols, viz. V, $\neg$, (,), $k_1, k_2, \ldots, k_n$.

The symbols of the language are given the following interpretations: Each letter in the alphabet is a "natural occurrence". The symbol V means "or", the symbol $\neg$ means "it is not true that...", the symbols (and) have the usual meaning of parentheses, and, for i=1,2,...,n, the symbol k_i means "individual i knows that...". The **formulas** of the language are given by the rules that each letter of the alphabet is a formula and that, given two formulas f and g, fVg, $\neg$f, k_if (for i=1,2,...,n) are also formulas. On this basis, a system of modal logic is now constructed, using the axioms of propositional calculus plus 4n further axioms, viz.

$$(k_i f) \Rightarrow f$$
$$(k_i (f \Rightarrow g)) \Rightarrow ((k_i f) \Rightarrow (k_i g))$$
$$(k_i f) \Rightarrow (k_i k_i f)$$
$$(\neg k_i f) \Rightarrow (k_i \neg k_i f)$$

where i=1,...,n, f and g are formulas, and f => g means $(\neg f) V g$. A **thesis** (or a **tautology**) is now defined by the conditions:

every axiom is a thesis;
if f and (f => g) are theses, then g is a thesis;
for i=1,...,n, if f is a thesis, then k_if is a thesis.

The resulting system is a multi-agent version of the modal logic known as **S5**, originally introduced by C.I. Lewis and C.H. Langford [6] in the exploration of possibility and necessity. Aumann now defines a **state of the world** as a list of formulas that contains all the theses and is inference-complete (i.e., if f and f => g are in the list, so is g) such that a formula f is in the list if,

and only if, its negation $\neg f$ is **not** in the list. The **signal** received by agent i at state S is the set of all formulas f such that f is not a thesis and $k_i f$ belongs to the list S.

Agent i's signal at a given state S carries two different kinds of information: First, there are the letters α of the alphabet for which $k_i \alpha$ belongs to S. This is what agent i knows at S about what Aumann calls "natural occurrences". This part of the signal is clearly sensory. Secondly, agent i's signal contains information about what **other agents** know, i.e., it includes formulas of the form $k_j f$ such that $k_i k_j f$ belongs to S, for some other agent j. How agent i can ever come to possess this kind of information turns out to be a very thorny issue. Note, in particular, that as a consequence of the axioms, the implication $k_i k_j f \Rightarrow k_i f$ always holds. i's sensory knowledge of f is a necessary condition for i's ability to know that j knows f. All this will be discussed at greater length in Section 3, below.

Aumann's essay contains a curious distinction between knowledge (without quotation marks) and "knowledge" (with quotation marks). If at some state S f is a fact (i.e., a **formula** which is not a thesis and belongs to S) then agent i either knows f (this is written $k_i f \in S$) or else agent i does not know f ($\neg k_i f \in S$). This is factual knowledge. For this to be at all meaningful, agent i must "know" the **language** in which f and $k_i f$ appear as formulas. Knowledge that is embodied in the structure of the language is normally called **analytic**. Aumann calls it "knowledge" (with quotation marks). Notice that, in Aumann, a great many things are worked into the structure of the language, thus becoming "known" (with quotation marks) to all agents, as part of their analytic tool-kit. For example, the state space, i.e., the complete account of all the possible states that the world can be in, is "known" and therefore analytic. Who are the other agents, how many of them are there, and what are their information structures, all this is "known" and therefore analytic. To say nothing of all theses, which are (tautologically)

"known", thus obviously analytic. Now, from the point of view of decision-making, knowledge and "knowledge" are of course equally applicable. In choosing actions, agents will bring to bear all the information they have, whether factual or analytic. The main difference between the two is that factual information possessed by one agent may not become available to other agents (due to differences in signals received), whereas analytic information is common to all agents, and this is itself commonly known. Some simple consequences of all this will now be explored.

3. Knowledge at the Lowest Common Denominator

In examples of two-agent interactive situations it has recently become fashionable to refer to the two agents as Alice and Bob. Onto this bandwagon we are delighted to jump: Alice and Bob are travelling in a car. Alice is driving and Bob is at her side, in the passenger's seat. They approach an intersection with a traffic light. It's a special kind of traffic light: Instead of being made up of three separate lighting elements – one red, one amber and one green – this particular traffic light is a high-tech job, with a single lighting element that changes colors. When Alice and Bob reach the intersection the light is red, so Alice stops the car. She can do this because she can see the red light for what it is (namely red), i.e., she is color-cognizant. Bob, on the other hand, is color-blind. The light changes to green. Alice knows (sees) that now the light is green. Bob is color-blind, so he does not know this. So far, so good. Now, what does Bob know about what Alice knows? Bob certainly knows that Alice knows the color of the light, i.e. Bob knows that Alice is not color-blind, and this knowledge is **analytic**. Indeed, Alice's being color-cognizant (i.e., her information structure) is part of the "language". If Bob had any doubt about whether or not Alice can see colors, he would have to move to a new world, and a new language, with the number of agents increased to include "Alice_1" and "Alice_2" in place of what used to be simply "Alice",

$Alice_1$ being color cognizant and $Alice_2$ being color blind. He would then have to proceed from some prior probability distribution on who it is that he is driving with ($Alice_1$ or $Alice_2$), which he would have updated, using Bayes' Rule, to take account of the evidence that, whoever it is that he is driving with, this person did in fact stop at the intersection. ("Did she stop because she saw a red light?") So, to keep matters simple, let us go back to just one Alice, good old color-seeing Alice, and to Bob knowing this. Alice knows that now the light is green and Bob does not. Can Bob know that Alice knows that the light is now green? Absolutely not! While he does know (analytically) that Alice can see colors, he cannot know (factually) that right now she sees green. Why? Because the event "Alice sees green" is contained in the event "the light is green" and so if Bob, who is color blind, cannot know the latter, he certainly cannot know the former. (In symbols, let g be the formula "the light is green". $K_A g \Rightarrow g$ is a thesis, so $K_B(K_A g \Rightarrow g)$ is true. It therefore follows from the second axiom that $K_B K_A g \Rightarrow K_B g$ and since the consequent is false, so is the antecedent.) Well, you might say, Bob obviously cannot know that Alice knows that the light is green, because Bob, being color blind, simply does not know what "green" **is**. This argument sounds good but it is invalid. For Bob knows **exactly** what "green" is. Recall that Bob knows Alice's information structure. In fact, he knows Alice's information structure **analytically**, as part of his knowledge of the language. Now Alice can see colors and Bob knows this. He knows exactly which visual signal Alice receives at every state of the world and, in particular, he knows exactly at which states she receives the signal "green". But the set of states at which she receives the signal "green" is precisely the set **defining** the color green. This, after all, is the meaning of Alice's being color cognizant: She can recognize green for what it is. Hence Bob knows exactly what "green" means, yet he cannot know that Alice knows that the light is green. Why? Bob cannot acquire the knowledge that Alice knows that the light is green, because knowledge is only acquired through the receipt of a sensory signal.

The problem seems to lie in the fact that people do not have sensory access to what other people know, which, under the sensory/analytic dichotomy, would appear to rule out any possibility for knowledge to be conveyed between individuals. Yet, much of what people know is, in fact, knowledge that had been conveyed to them from other people. Alice cannot arrange for Bob to know what she knows, no matter how hard she might try. Certainly, no oral pronouncement can ever make Bob know that she sees a green light. If she were to utter the words "I see a green light", all Bob would know would be "Alice says she sees a green light". This newly acquired knowledge may affect Bob's **beliefs** concerning Alice actually seeing a green light, but not his knowledge of it. Bob's beliefs depend, among other things, on Bob's theory regarding Alice's **motives** in making her pronouncement, whereas Bob's knowledge should not.

Things get worse as Alice and Bob prepare to move. (Imagine a dozen or so cars lined up behind Alice and Bob, with the drivers impatiently hooting their horns.) As soon as Alice sees the light changing to green, she can of course put the car in gear and proceed to cross the intersection. But Alice wants Bob to know what she is doing. She wants him to know that the policy "move when the light is green" has been adopted (or is being considered). To her dismay, she discovers, however, that Bob can never know this policy. For, in order for Bob to know any policy of action, it must be the case that this policy of action is **measurable** with respect to Bob's information structure. The policy "move when the light is green" does not satisfy this requirement. There are in fact only two policies which Alice can adopt with Bob's knowledge, namely "move, no matter what color the light" and "stay put, no matter what color the light". And if both Alice and Bob subscribe to the rule that peril must be avoided, then they will agree that "stay put, no matter what" is the policy to be adopted. Just imagine the frustration of those drivers in the cars behind.

This is, of course, none other than Aumann's [1] celebrated **Agreeing to Disagree** result: Suppose that Alice, after receiving a signal, uses this signal to update her prior probability distribution (i.e., her prior **beliefs**) over some sample space. If Bob is to know Alice's posterior distribution, then Alice must take care not to condition on anything that Bob cannot know. She must, in other words, restrict the evidence to be used for updating her beliefs to publicly known events. So, if Alice and Bob share a common prior and if both use the same decision rule (Bayes' updating rule, say) then restricting the evidence to publicly known events will obviously result in Alice and Bob reaching the same posterior distribution. They cannot agree to disagree. (Note that this coincidence of posteriors is here a consequence merely of Bob's knowing Alice's posterior distribution. This is because we are dealing with the special case where one agent's – Alice's – information structure is a **refinement** of the other agent's – Bob's – information structure. In the more general case, with no such relationship between the information structures, to obtain the coincidence of posteriors one must postulate that these posteriors are **common knowledge** among the agents. The argument, however, is basically the same: In order for the posteriors to be common knowledge, agents must restrict their updating to publicly known events, so agreement on the posteriors follows from agreement on the priors and on the updating rule.)

Where does all this leave us? It leaves us with the awkward conclusion that, under the sensory/analytic dichotomy, the only knowledge that agents can **share** is knowledge at the **lowest common denominator**. If you want me to know that you know something, then you must take care not to know too much. If Alice wants Bob to know what she knows, then she must take care not to know that the light has turned green. This awkward conclusion follows, we maintain, from an uncritical embrace of the principle that knowledge is either sensory or else analytic. How can Bob know that Alice knows that the light is green? Obviously, he cannot

know this analytically, as part of the "language", so this knowledge must in fact be sensory. Bob's information on what Alice knows must be obtained through the receipt of a sensory signal, and this can only happen when that information is consistent with Bob's capacity for receiving sensory signals.

These remarks apply also to the much-discussed notion of common knowledge of rationality. If you and I are rational, then for our rationality to be common knowledge between us, we must restrict our behavior in such a way that an observer who can only see public events will know that we are rational. Our rationality must, so to speak, be evident at the lowest common denominator.

All this points to a theory of knowledge which seems to us too confining. Now the obvious retort to this criticism is to say that for a theory to have explanatory power, it must be confining. But this, we maintain, is more a case of embracing a dogma then of seeking reasonable explanatory power. We would argue, following Hintikka [5], that the possibility of knowledge which is neither analytic nor sensory must be allowed for. Agents must be allowed to possess knowledge which, on the one hand, is knowledge about the world (and in this sense "factual") but, on the other hand, does not depend on the receipt of a sensory signal. We will maintain, for example, that knowing what states the world can possibly be in – i.e., knowledge of the "state space" – is necessarily knowledge of this kind. (In our opinion, writing the state-space into the language, thereby making its knowledge analytic, is a ploy that cannot seriously be defended.) We propose that the word **comprehension** be used to designate this kind of knowledge which is neither analytic nor sensory. Thus, Bob, who is color blind, should nevertheless be able to **comprehend** the fact that Alice knows that the light is green. Similarly, agent A should be able to **comprehend** that agent B is rational without B's rationality necessarily being measurable with respect to the information structure that arises from writing down the sensory signal received by A at every state of the world.

4. The Hangman's Paradox

Think of a group of boy-scouts (say n of them) forming a single file, i.e., standing in a straight line one behind the other, each (except the first) facing the back of the boy directly in front. Of course, the boy-scouts in the line all wear caps on their heads. Among these caps, all but one are white and the remaining one – exactly one cap – is red. None of the boys is near-sighted or color-blind, so each boy can see all the caps worn by the boys who are ahead of him in line, and recognize their colors. He can't see his own cap, nor of course the caps of those behind him. This is, so to speak, the physics of the situation, and we shall assume that all the boys know it. How did they come to know it? In what sense do they know it? How thoroughly do they know it (do they know, for example, the underlying physical and biological laws)? These questions are all relevant and difficult, but we are here following Aumann and saying that the boy-scouts have a common "language", in which their knowledge of what we have called the physics of the situation is somehow embedded. It should be noted that through knowing the physics of the situation, each boy also knows something about what the other boys know. For example, the boy who is fifth in line knows what it means (physically) to be, say, eighth in line, and in particular what being eighth in line allows one to see. The fifth boy thus knows something about what the eighth boy knows. Basically, this knowledge rests upon **counter-factual** reasoning: If I **were** eighth in line, this is what I **would** see, so says the boy who is fifth. Counter-factual knowledge generally plays a major role in economic and game-theoretic reasoning, and this case is no exception. Note, however, that under the sensory/analytic dichotomy, all counter-factual knowledge is necessarily analytic since the **counter**-factual cannot be factual.

We shall now consider a world with n possible states, defined by who's wearing the red cap. Formally, a state-space Ω may be defined by writing $\Omega=\{1,2,\ldots,n\}$, where the i-th state ($i\in\Omega$) is

identified by the assertion "the red cap is on the head of the boy who is i-th in line". We assume, with Aumann, that all the boys know Ω, and that this knowledge is analytic.

Let $s_i(j)$ be the sensory signal which the i-th boy (i.e., the boy who is i-th in line) receives when the state of the world is j (i.e., when the j-th boy is the one wearing the red cap). We have:

$$s_i(j) = \begin{cases} j & \text{if } j<i \\ \neg(1V2V\ldots Vi-1) & \text{if } j\geq i \end{cases}$$

The values which the function $s_i(\ )$ can take are sentences (or propositions). When $j<i$, $s_i(j)$ is the sentence "the red cap is on j's head", i.e., when $j<i$, the signal received by the i-th boy tells him exactly what the state of the world is. When $j\geq i$, i's signal tells him only that the red cap is neither on 1's head, nor on 2's head,..., nor on (i-1)'s head. Note that when $j\geq i$ the values of $s_i(j)$ are constant, i.e., i receives the same signal at all j, so long as $j\geq i$. Let $\Pi_i(j)$ be the set of all states at which the signal received by i would be the same as that received at j. In other words, $\Pi_i(j)$ is the set of those states in Ω which i cannot distinguish from state j:

$$\Pi_i(j) = \{j'\in\Omega \mid s_i(j')=s_i(j)\}$$

$$= \begin{cases} \{j\} & \text{if } j<i \\ \{i,i+1,\ldots,n\} & \text{if } j\geq i \end{cases}$$

This definition gives rise to the following important remark:

In defining the set $\Pi_i(j)$, one uses a condition of the form $s_i(j')=s_i(j)$. Now the values of the function $s_i(\ )$ are **sentences**, so in order to define $\Pi_i(\ )$, one must first settle the question of what it means for two sentences to be equal. In a formal language, sentences are equal when they are logically equivalent, and

logical equivalence is well-defined within the language. Normally, however, people don't receive signals dressed as formal propositions, nor do they normally have a Robert Aumann on hand to transform for them all facts and all knowledge of facts into formulas in some pre-constructed formal language. People normally receive signals in the form of sentences in some **natural** language. More precisely, the signals which an individual receives, if they are not themselves sentences in some natural language, are **processed** by the receiving individual into sentences in some natural language. In either case, the assertion $s_i(j')=s_i(j)$ is now to be understood as saying that individual i regards (or understands, or comprehends) the two natural-language sentences $s_i(j')$ and $s_i(j)$ as equivalent to each other, i.e., that from individual i's point-of-view the sentences $s_i(j')$ and $s_i(j)$ "say the same thing". At this point, one immediately faces a whole host of issues having to do with the status and possible indeterminacy of meaning in natural languages and with how information which is formulated in a natural language is comprehended and conveyed. These are precisely the questions which Donald Davidson [3] and others (see, e.g., Lewis [7]) have been discussing under the heading of "Radical Interpretation". Thus, going from the function $s_i(\ )$ to the function $\Pi_i(\ )$ is far from a trivial matter.

The set $\Pi_i(j)$ tells us what the i-th boy knows about the state of the world when the true state is j. Specifically, when the true state is j, what the i-th boy knows is that the world is in one of the states in $\Pi_i(j)$. Unlike $s_i(j)$ which is a sensory signal and hence represents **factual** knowledge, $\Pi_i(j)$ spells out knowledge which in general is not factual (or not **merely** factual). In particular, when $j \geq i$ we have $\Pi_i(j) = \{i,i+1,\ldots,n\}$ which says that i knows the true state to be one of the states $i,i+1,\ldots,n$. In order to arrive at this knowledge, the i-th boy must reason as follows: (1) I can see (factual knowledge) that none of the boys ahead of me wears the red cap, i.e., that the true state is not in the set $\{1,2,\ldots,i-1\}$. (2) I know (analytically) that the state-

space is Ω, which is given by the set $\{1,2,\ldots,n\}$. (3) From the fact that the true state must be in Ω but is not in $\{1,2,\ldots,i-1\}$ I **deduce** (analytically) that the true state is in $\{i,i+1,\ldots,n\}$. Thus, in order to arrive at $\Pi_i(j)$, the i-th boy uses both factual and analytic knowledge, plus logical deductions which are also defined to be part of analytic knowledge. Now the logical inference which the i-th boy uses to arrive at $\Pi_i(j)$ is a very primitive one, and it is hard to object to the assumption – and an assumption it is – that he is capable of performing such inferences. Recall, however, that under Aumann's theory, agents are assumed capable of correctly carrying out **all** logical reasoning, no matter how complex. Once the axioms are in the language, so are all the theorems. This is the so-called Logical Omniscience, mentioned briefly in Section 1, above. The questions arising in this connection are extremely difficult: Is a capacity for omniscient logical reasoning a pre-condition for rationality? If it is, then nobody is rational. If it isn't, then one must decide what would be the minimal capacity for logical reasoning required for rationality. For it is not even meaningful, let alone useful, to speak of rationality for an agent who does not possess **some** capacity for logical inference. And what about computational omniscience, which is a consequence of logical omniscience? Does inability to solve a given problem in polynomial time imply that full rationality is unattainable? We do not pretend to have a theory of knowledge for rational decision making that gives satisfactory answers to these questions. Our claim is merely that going for logical omniscience, as the existing theory does, is not likely to be the right way.

The function $\Pi_i(\ )$ leads us immediately to the i-th boy scout's information structure. Specifically, since $\Pi_i(j)$ tells us what i knows when the state is j, letting j run through the entire set Ω would give us a description of what i knows at any given state, which is i's information structure. Accordingly, letting Π_i be defined by $\Pi_i = \{\Pi_i(1),\Pi_i(2),\ldots,\Pi_i(n)\}$, we get

$$\Pi_i = \{\{1\},\{2\},\ldots,\{i-1\},\{i,i+1,\ldots,n\}\},$$

for i=1,2,...,n. Π_i, which is indeed referred to as i's **information structure**, is clearly a **partition** of the state-space Ω.

In the previous Section there were two agents, Alice and Bob, with information structures such that Alice could always know what Bob knows, but Bob could not always know what Alice knows. Alice's information partition was a **refinement** of Bob's. Here the situation is similar, albeit with possibly more than two agents. Each boy scout in the line knows more than the boys ahead of him and less than the boys behind him. For any two boy scouts in the file, the information partition of the one who is behind is a refinement of the information partition of the one who is ahead. In particular for the first and last boys we have:

$$\Pi_1 = \{1,2,\ldots,n\}$$

and

$$\Pi_n = \{\{1\},\{2\},\ldots,\{n\}\}.$$

The first in line knows nothing (about who's wearing the red cap) and the last in line knows everything, in the sense that no matter who wears the red cap, n knows it.

We are now, at long last, ready to introduce a version of the so-called Hangman's Paradox. (See Sorenson [8] for a related discussion.) Suppose that the boy in the t-th place in the line is in fact the one wearing the red cap, i.e., let the state of the world which actually obtains be denoted t ($t\in\Omega$). Consider the condition

(*) $$\{t\} \notin \Pi_t \ .$$

What (*) says is: He who wears the red cap doesn't know that he is wearing it.

Suppose that (*) becomes known to those poor boy scouts standing there in line. That is, assume that the boys **know** (*) to be true. Obviously, this knowledge cannot be sensory so we have to assume that the boys know (*) **analytically**, in the same way that they know Ω and the Π_i's. (Note that (*) being known analytically makes (*) **common knowledge** among the boys.) We shall say nothing about how the boys could have come to know (*), even though obviously this is a question of considerable importance.

As soon as (*) is known, the boy who is 17th in line raises his hand and, when given permission to speak, says: "Making us stand in line like this for hours on end is nothing but a cruel hoax. There's no one here with a red cap on his head!" This proclamation obviously causes some alarm in certain quarters, so no. 17 is ordered to explain. "Well", he says, "the red cap – assuming there is one – cannot possibly be on n's head, because if it were, n would know it. Now, n-1 cannot possibly have the red cap on his head, because if he did, he would see all the boys ahead of him wearing white caps so he would know that the red cap is either on his own head or on n's. But n-1 knows that the red cap cannot be on n's head (if it were, n would know it), therefore it must be he, n-1, who is wearing it. Thus, if he himself had a red cap on his head he would know it. Continuing recursively in this manner, we find that if any one of us had the red cap on his head he would know it, and this would be true in particular for the boy t, whoever he might be." Having thus completed his argument, no. 17 is asked to please take off his cap. "What color is your cap?" he is asked. "Red." "Did you know you were wearing the red cap?" "No."

Let us examine this story carefully. Checking the condition (*) against the definition of the information partitions Π_i, we find that (*) is true if, and only if, $t<n$. Repeating, for the sake of emphasis, we have that the assertions $\{t\} \not\in \Pi_t$ and $t<n$ are **equivalent** to each other. Now suppose that at the outset we were to assume that the boy scouts somehow came to know $t<n$ instead of

assuming, as we had, that what they came to know was (*). Surely, since $t<n$ and (*) are equivalent, there can be no difference between the two settings. Yet think of that smart kid, no. 17. If $t<n$ were the assertion that was known to be true, he would no doubt have held his peace thus sparing himself an embarrassment. There simply is no hope for a recursive argument based on $t<n$, an inequality which says merely that the red cap is not on the last boy's head. Something is amiss.

Note that in recounting no. 17's argument we did not use (*) directly. Instead we used the English sentence "he who wears the red cap doesn't know that he is wearing it". Is it indeed the case that this sentence expresses the same reality that the formula (*) does? Not quite. A formula is only meaningful insofar as the symbols comprising it are well-defined. In particular, both the meaningfulness and the meaning of the formula

$$(*) \qquad \{t\} \not\in \Pi_t$$

depend on the definitions of the various symbols appearing in it, and primarily on the definitions of the information partitions $\Pi_1, \Pi_2, \ldots \Pi_n$. These, we recall, are defined as follows:

$$\Pi_i = \{\{1\}, \ldots, \{i-1\}, \{i, \ldots, n\}\}$$

for $i=1,\ldots,n$. The Π_i, in turn, are only well defined given the definition of the state-space Ω, i.e. given $\Omega=\{1,2,\ldots,n\}$. Thus, (*) is only meaningful – and only applicable – for this particular state-space and for those particular information structures. The English-language phrase "he who wears the red cap doesn't know that he is wearing it" is not so tightly tied down. In fact, what no. 17 did was to use this phrase while constantly **changing** the state-space and the corresponding information structures. Here is exactly what he did:

(1) We all know (*), and (*) is equivalent to $t<n$.

(2) $t<n$ means that n cannot be the true state of the world, so $\Omega=\{1,2,...,n\}$ is **not** the correct state-space. Rather, the correct state-space is $\Omega^{n-1}=\{1,2,...,n-1\}$.

(3) If Ω^{n-1} is the correct state-space, then the correct information structures are not $\Pi_1,...,\Pi_n$ but $\Pi'_1,...,\Pi'_n$ defined by

$$\Pi'_i = \{\{1\},...,\{i-1\},\{i,...,n-1\}\} \quad \text{for} \quad i=1,...,n-1$$
$$\Pi'_n = \Pi'_{n-1}$$

(4) Letting t be the true state and noting that t can only take on the values 1,2,...,n-1, we must now re-write (*) for the new setting, as follows:

$(*)^{n-1}$ $\qquad \{t\} \notin \Pi'_t$.

Translated into English, what $(*)^{n-1}$ says is "he who wears the red cap doesn't know that he is wearing it".

(5) The condition $(*)^{n-1}$ is equivalent to $t<n-1$. If $(*)^{n-1}$ is known to be true, then n-1 can never be the state of the world, so Ω^{n-1} is not the correct state space but rather $\Omega^{n-2}=\{1,2,...,n-2\}$.

and so forth.

What this shows is that in order to obtain a proper version of the Hangman's Paradox (or the Surprise Test Paradox) one needs a condition which is stronger than (*). Specifically, the condition that will give us the paradox, and provide the proper foundation for no. 17's argument, is this:

(**) Whatever the appropriate state-space and corresponding information structure Π_i, the true state t shall satisfy

$$\{t\} \notin \Pi_t \ .$$

Given (**), no. 17's argument seems flawless: The appropriate state-space keeps shrinking until the empty state-space (no one's got a red cap) is reached, where the condition $\{t\}\notin\Pi_t$ can be said to hold vacuously. Yet, when no. 17 takes a look at his own cap he finds that it is red, admitting that he did not know this when the cap was still on his head. Something must have gone wrong in his reasoning after all.

As we have seen, no. 17's backward induction argument works through the manipulation of the state-space. Well, it turns out that manipulating the state-space is a very tricky business. In order to find out where the crux of the matter lies, let us reconstruct no. 17's argument in such a way that his conclusion will in fact be correct.

We start with $\Omega=\{1,2,\ldots,n\}$, together with the condition $\{t\}\notin\Pi_t$ which, as we know, is equivalent to $t<n$. At this point, no. 17 makes the following suggestion: "Since $t<n$, it is certain that the n-th boy cannot be wearing the red cap. So, what's the point of keeping no. n stupidly standing in line like this?" This leads to no. n being told that he can put his cap in his pocket and go to the cafeteria for a cup of hot chocolate. We are now left with just n-1 boys standing in line, the n-th having left for the cafeteria. The new state-space is given (**physically**) by $\Omega^{n-1}=\{1,2,\ldots,n-1\}$ and $\{t\}\notin\Pi_t$ tells us that the red cap cannot be on (n-1)'s head. So, n-1, who no longer has any relevance for who's wearing the red cap, is excused and goes to the cafeteria for hot chocolate. The line has physically shrunk once again, the new state space is $\Omega^{n-2}=\{1,\ldots,n-2\}$ and the process can continue. No. 17, who will be completing his argument in the cafeteria, is now absolutely right: The state-space has been reduced to nought and the condition $\{t\}\notin\Pi_t$, if at all meaningful, holds vacuously.

In this version of the story, the shrinking of the state-space is **physical**. It becomes part of what we have called "the physics of the situation". In the previous version, with all the boys staying in line with their caps on their heads, the shrinking of the state-space was **virtual**. The very same argument can be sound when the revisions of the state-space are physical and faulty when these revisions are virtual. It thus becomes a matter of utmost importance for agents to know the precise grounds for admitting a candidate to membership in the state-space or denying it. This kind of knowledge can be neither purely factual nor purely analytic. We call it comprehension.

5. Conclusion

The most consistent and eloquent advocate of the view that knowledge must be either sensory or else analytic was, of course, David Hume. From the very start, however, one of the nagging questions which Hume's analysis did not appear to resolve was how to deal with knowledge of "other selves". Some critics went so far as to assert that Hume's position must necessarily lead to solipsism. It would not be incorrect, we think, to describe this essay as an echo of that criticism of David Hume: In the study of interaction among agents, the requirement that agents' knowledge be subject to sensory/analytic dichotomy cannot be sustained. Our suggestion for relaxing this dichotomy involves the introduction of what we have called "comprehension". Whether or not "comprehension" can be modelled explicitly remains to be seen.

References

[1] Aumann, R.J., "Agreeing to Disagree", **The Annals of Statistics**, 4, 1236-39, 1976.

[2] Aumann, R.J., **Notes on Interactive Epistemology**, version for 17 July, 1992, unpublished.

[3] Davidson, D., "Radical Interpretation", in **Inquiries into Truth and Interpretation**, 125-39, Oxford: Clarendon Press, 1984.

[4] Geanakoplos, J., **Common Knowledge**, Yale University, 1990, unpublished.

[5] Hintikka, J., "Information, Deduction, and the A-Priori", in **Logic, Language-Games and Information**, 222-41, Oxford: Clarendon Press, 1973.

[6] Lewis, C.I., and Langford, C.H., **Symbolic Logic**, New York: Dover, 1932 (2nd Ed., 1959).

[7] Lewis, D., "Radical Interpretation", in **Philosophical Papers**, Vol. 1, 108-22, New York: Oxford University Press, 1983.

[8] Sorenson, R.A., "Reculcitrant Variations of the Prediction Paradox", **Australian Journal of Philosophy**, 60, 355-62, 1982.