Belief and Rationality: Human and Animal

Rohit Parikh

City University of New York

IIT Kanpur, January 2008

Not long ago, if you wanted to seize political power in a country, you had merely to control the army and the police. Today it is only in the most backward countries that fascist generals, in carrying out a coup d'etat, still use tanks. If a country has reached a high level of industrialization, the whole scene changes. The day after the fall of Khruschev, the editors of Pravda, Izvestiia, *the heads of the radio and television were replaced; the army wasn't called out. Today, a country belongs to the person who controls communications.*

Umberto Eco
Towards a Semiological Guerrilla Warfare, 1967

The beliefs held by an agent are represented by a set of such belief-representing sentences. It is usually assumed that this set is closed under logical consequence, i.e. every sentence that follows logically from this set is already in the set.

Sven Hansson in the
Stanford Encyclopedia of philosophy

Socrates: Do you see, Meno, what advances he has made in his power of recollection? He did not know at first, and he does not know now, what is the side of a figure of eight feet: but then he thought that he knew, and answered confidently as if he knew, and had no difficulty; now he has a difficulty, and neither knows nor fancies that he knows.

Meno: True.

Socrates: Is he not better off in knowing his ignorance?

Meno: I think that he is.

Socrates: And that is the line which the learned call the diagonal. And if this is the proper name, then you, Meno's slave, are prepared to affirm that the double space is the square of the diagonal?

Boy: Certainly, Socrates.

Socrates: What do you say of him, Meno? Were not all these answers given out of his own head?

Meno: Yes, they were all his own.

A common semantics for the logic of knowledge uses Kripke structures with an accessibility relation R, typically assumed to be reflexive, symmetric, and transitive. If we are talking about belief rather than knowledge, then R would be serial, transitive, and euclidean.

Then some formula ϕ is said to be believed (known) at state s iff ϕ is true at all states R-accessible from s. Formally,

$$s \models B(\phi) \quad \text{iff} \quad (\forall t)(sRt \rightarrow t \models \phi)$$

- If a formula is logically valid then it is true at all states and hence it is both known and believed.
- If ϕ and $\phi \rightarrow \psi$ are believed then ψ is also believed at s.
- A logically inconsistent formula can be neither known nor believed.

logically omniscient humans?

Suppose that

- p stands for *Pandas live in Washington DC*,
- q stands for *Quine was born in Ohio*,
- r stands for *Rabbits are called* gavagai *at Harvard.*

Suppose that Jill believes that p is true and that q and r have the same truth values. Then she is allowing two truth valuations, $v = (t, t, t)$, and $v' = (t, f, f)$.

In particular she should believe ϕ, i.e., $(r \leftrightarrow (p \leftrightarrow q)$

Perhaps she *does*. But note that she will actually have to **make the calculations** rather than just sit back and say, "Now **do** I believe ϕ?"

The following example from Daniel Kahneman's Nobel lecture, 2002.

> *A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost? Almost everyone reports an initial tendency to answer 10 cents because the sum $1.10 separates naturally into $1 and 10 cents, and 10 cents is about the right magnitude. Frederick found that many intelligent people yield to this immediate impulse: 50% (47/93) of Princeton students, and 56% (164/293) of students at the University of Michigan gave the wrong answer. Clearly, these respondents offered a response without checking it.*

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations.

#6 Linda is a bank teller

#8 Linda is a bank teller and active in the feminist movement

89% of respondents rated item #8 higher in probability than item #6.

But the set of bank tellers who are active in the feminist movement is a proper subset (perhaps even a rather small subset) of the set of all bank tellers, so #8 *cannot* have higher probability than #6.

An agent **endorses** (agrees with) a sentence ϕ iff she asserts ϕ or chooses *Yes* when asked, *Do you think* ϕ*?*, and **denies** (or disagrees with) ϕ iff she chooses *No*. She may also choose *Not sure*, in which case of course she neither endorses nor denies.

An agent may endorse $X = \{\phi_1, \phi_2, ..., \phi_n\}$, X implies ψ, and either deny ψ or at least fail to endorse ψ. We will say in the first case that the agent is *logically incoherent*, and in the second that the agent is an incomplete reasoner – or simply *incomplete*.

If X is an inconsistent set of sentences, then X implies ψ for arbitrary ψ. Thus it is obvious that an agent who is logically incoherent, but *not* incomplete, and endorses X, will end up endorsing everything. Fortunately, most of us, though we *are* logically incoherent, tend also to be incomplete.

Imagine that Carol assigns probabilities of .3, .3, and .8 respectively to events $X, Y, X \cup Y$. One could say that these probabilities are **inconsistent**. But in fact nothing prevents Carol from *accepting* bets based on these probabilities. What makes them incoherent is that we can make Dutch book against Carol – i.e., place bets in such a way that no matter what happens, she will end up losing money.

For instance we can bet \$3 on X, \$3 on Y, and \$2 against $X \cup Y$. If either X or Y happens, we earn \$7 (at least), and lose (at most) \$5, thus gaining \$2. If neither happens, we gain \$8 and lose \$6, so that we again make a profit – and Carol makes a loss.

Thus incoherent beliefs, on this account, are *unwise*, but **possible**.

Interpreting Mental States through Plans and Actions

Hayek [12] considers an isolated person acting over a period according to a preconceived plan. The plan

> *may, of course, be based on wrong assumptions concerning external facts and on this account may have to be changed. But there will always be a conceivable set of external events which would make it possible to execute the plan as originally conceived.*

The belief states which are implicit in plans are more general than the belief states which correspond to Kripke structures,

Having a plan does not require that the plan be formulated explicitly in language, or even that the planner *has* a language. It is perfectly possible for an animal to have a plan and to a smaller extent, it is also possible for a pre-lingual child to engage in deliberate behaviour which is plan-like.

Animals Can also have Plans!

..as has long been thought, is the attribution of mental states confined solely to humans?

In humans, the evidence for attribution almost always comes back to language. Therefore in order to answer the question of whether animals attribute mental states, a method for testing must be found that does not involve language.

Premack and Premack, [36] p. 139

It wasn't until an ape saved a member of our own species that there was public awakening to the possibility of nonhuman kindness. This happened on August 16, 1996 when an eight-year old female gorilla named Binti Jua helped a three-year-old boy who had fallen eighteen feet into the primate exhibit at Chicago's Brookfield Zoo. Reacting immediately, Binti scooped up the boy and carried him to safety.

Frans de Waal

Next to the ridicule of denying an evident truth is that of taking much pains to defend it; and no truth appears to me more evident, than that beasts are endow'd with thought and reason as well as men.

David Hume

It is for instance possible to say that a chicken believes a caterpillar of a certain sort to be poisonous,and mean by that merely that it abstains from eating such caterpillars on account of unpleasant experiences connected with them.

Frank Ramsey

Two kinds of beliefs

- **Non-linguistic beliefs** to be called **e-beliefs** which may also be possessed by animals
- **Linguistic beliefs** which we will call **i-beliefs** and which can only be possessed by humans; adults and older children.

Of course the last two groups will *also* have non-linguistic beliefs which must be somehow correlated with their linguistic beliefs.

Let $\mathcal{B}$ be the space (so far unspecified) of belief states of some agent. Then the elements of $\mathcal{B}$ will be identified with the choices which the agent makes.

Roughly speaking, if I believe that it is raining, I will take my umbrella, and if I believe that it is not raining, then I won't.

But clearly the choice of whether to take my umbrella or not is correlated with my belief only if I don't want to get wet. So my preferences enter **in addition** to my beliefs.

We assume that the agent has *some* space $\mathcal{P}$ of preferences, and that the choices are governed by the beliefs as well as the preferences.

We use $\mathcal{S}$ for the set of choice situations,
and C for the set of choices.

Thus the set $\{U, \neg U\}$ could be a choice situation (with U standing for *take the umbrella*) and both U and $\neg U$ are elements of C.

An agent who does have language can also be subjected to a purely linguistic choice. If asked **Do you think it is raining?** the agent may choose from the set **{Yes, No, Not sure}**.

Elements of $\mathcal{B}$ cannot be identified with propositions, for an agent may agree to one sentence expressing a proposition and disagree (or not agree with) another sentence expressing the same proposition.

An agent in some state $b \in \mathcal{B}$ may agree with **"Superman is strong"** while disagreeing with **"Clark Kent is strong"**.

On hearing, **"But Superman is the same person as Clark Kent!"** she will go into state b' in which she will presumably drop the disagreement with **"Clark Kent is strong"**.

But it is important that $b' \neq b$. The question, **What did she really believe in state b?** makes no sense.

An entirely different sort of incoherence arises when an agent's linguistic behaviour does not comport with his choices.

If an agent prefers not to get wet (which we knew somehow), says that it is raining, and does not take her umbrella, she may well be quite coherent in her linguistic behaviour, but her linguistic behaviour and her non-linguistic choices have failed to match.

But it is going to be **usually** the case that the agent will choose U in the situation, $\{U, \neg U\}$ iff she chooses *Yes* in the situation where she hears *Do you think that it is raining?*

Some technical details

We assume given a space $\mathcal{B}$ for some agent whose beliefs we are considering. The elements of $\mathcal{B}$ are the belief states of that agent.

There are three important update operations on $\mathcal{B}$ coming about as a result of (i) events observed, (ii) sentences heard, and (iii) deductions made.

Elements of $\mathcal{B}$ are also used to make **choices**.

Thus our three update operations are:

$\mathcal{B} \times \mathcal{E} \rightarrow_e \mathcal{B}$

A belief state gets revised by witnessing an event.

$\mathcal{B} \times \mathcal{L} \rightarrow_s \mathcal{B}$

A belief state gets revised through hearing a sentence.

$\mathcal{B} \rightarrow_d \mathcal{B}$

A deduction causes a change in the belief state (which we may sometimes represent as an **addition**).

Finally, we also have a space $\mathcal{S}$ of **choice sets** where an agent makes a particular choice among various alternatives. This gives us the map

$$\mathcal{B} \times \mathcal{S} \rightarrow_{ch} \mathcal{B} \times C$$

An agent with a certain belief makes a choice among various alternatives.

If we want to explicitly include preferences, we could write,

$$\mathcal{B} \times \mathcal{P} \times \mathcal{S} \rightarrow_{ch} \mathcal{B} \times C$$

While $\mathcal{S}$ is the family of choice sets, C is the set of possible choices and $\mathcal{P}$ is *some* representation of the agent's preferences. Thus **{take umbrella, don't take umbrella}** is a choice set and an element of $\mathcal{S}$, but *take umbrella* is a choice, and an element of C.

Thus our theory of an agent presupposes such a belief set $\mathcal{B}$, and appropriate functions $\rightarrow_e, \rightarrow_s, \rightarrow_d, \rightarrow_{ch}$.
We can identify two different spaces $\mathcal{B}, \mathcal{B}'$ if they are bisimilar – they need not be isomorphic.

We can understand an agent (with some caveats) if what we *see* as the effects of these maps conforms to some theory of what an agent wants and what the agent thinks. And we succeed pretty well. *Contra* Wittgenstein, we not only have a theory of what a lion wants, and what it means when it growls, we even have theories for bees and bats.

Many beliefs are expressed (or so we think) by sentences.

The Setting

In our setting we imagine an **observer** o who is pondering on what some **agent** i believes. We assume (for convenience) that o thinks of a proposition expressed by a sentence as a set of possible worlds where that sentence is true, but that the observee i need not even have a language or a notion of truth.

However, it is assumed that i does have some plans. Even if i is just a dog digging for a bone, o understands that i has a plan and roughly what that plan is. And we shall use this plan to make it possible for o to attribute beliefs to i.

We also assume that there is a **context** C which is the set of **relevant** possible worlds, and that worlds outside C, even though they are there, are not considered in deliberating about i's belief or beliefs.

So let P be i's plan at the moment, and let $\pi(P)$ be the set of worlds w in C such that the plan is *possible* at w.

Formally, $\pi(P) = \{w|\ w \in C \wedge w\ \mathit{enables}\, P\}$.

Let ϕ be a sentence. Then $||\phi|| = \{w|w \models \phi\}$, the set of worlds where ϕ is true, is the *proposition* corresponding to the sentence ϕ. If ϕ and ψ are logically equivalent, then $||\phi|| = ||\psi||$.

Definition 0.1 *We will say that i* **e-believes** *ϕ, $B_e^i(\phi)$ if $\pi(P) \subseteq ||\phi||$. We will suppress the superscript i when it is clear from context.*

It is obvious in terms of the semantics which we just gave that the statement "The dog e-believes that there is a bone where he is digging" is true.

Also, if an agent e-believes ϕ and ψ then the agent also e-believes $\phi \wedge \psi$ and that if the agent e-believes ϕ and $\phi \rightarrow \psi$ then the agent e-believes ψ.

Oddly enough, creatures which do not use language do not suffer from a lack of logical omniscience!

Suppose someone has a plan P consisting of, "If ϕ then do α, else do β" and another plan P' consisting of "If ϕ then do γ, else do δ". Now we find him doing α and also doing δ (we are assuming that the truth value of ϕ has not changed). We could accuse him of being illogical, but there is no need to appeal to logic. For he is doing Dutch book against himself.

Presumably he assumed that $u(\alpha|\phi) > u(\beta|\phi)$ but $u(\alpha|\neg\phi) < u(\beta|\neg\phi)$. Thus given ϕ, α was better than β but with $\neg\phi$ it was the other way around. Similarly, $u(\gamma|\phi) > u(\delta|\phi)$, but $u(\gamma|\neg\phi) < u(\delta|\neg\phi)$. And that is why he had these plans. But then his choice of α, δ results in a loss of utility whether ϕ is true or not. If ϕ is true then he lost out doing δ and if ϕ is false, then he lost out doing α.

For a concrete example of this, suppose that on going out I advise you to take your umbrella, but fail to take mine. If it is raining, there will be a loss of utility for I will get wet. If it is not raining, there will be a loss of utility because you will be annoyed at having to carry an umbrella for no good reason. My choice that I advise you to take your umbrella, but fail to take mine, is not *logically* impossible. It just makes no pragmatic sense.

A similar argument will apply if someone endorses ϕ, endorses $\phi \rightarrow \psi$ and denies ψ. If such a person makes plans comporting with these three conditions, then he will make choices which do not maximise his utility.

A Second Notion of Belief – language enters

We now define a second notion of belief which does *not* imply logical omniscience. This is a more self-conscious, language-dependent notion of belief.

For agents i who do have a language (assumed to be English from now on), their plan may contain linguistic elements. At any moment of time they have a finite stock of currently believed sentences. This stock may be revised as time passes. These agents may perform atomic actions from time to time, and also make observations which may result in a revision in their stock of believed sentences.

Thus Lois seeing Superman in front of her will add the sentence "Superman is in front of me", to her stock, but, since she does not know that Clark Kent is Superman, she will *not* add the sentence "Clark Kent is in front of me". Someone else may add the sentence "I see the Evening Star", but not the sentence "I see the Morning Star" at 8 PM on a summer night. A person who knows that $ES = MS$, may add the sentence, "Venus is particularly bright tonight." In any case, this stock consists of sentences and not of propositions.

The basic objects in the agents' **plans** are atomic actions and observations which may be active (one looks for something) or passive (one happens to see something). These are supplemented by the operations of concatenation (sequencing), *if then else*, and *while do*, where the tests in the *if then else* and *while do* are on *sentences*. There may also be recursive calls to the procedure: *find out if the sentence ϕ or its negation is derivable within the limits of my current resources, from my current stock of beliefs*. Thus if i's plan has currently a test on ϕ, then, to be sure, the stock of sentences will be consulted to see if ϕ or its negation is in the stock. But there may also be a recursive call to a procedure for deciding ϕ. If someone asks "Do you know the time?", we do not usually say, "I don't", but look at our watches. Thus consulting our stock of sentences is typically only the first step in deciding if some sentence or its negation can be derived with the resources we have.

Sentences and Propositions

Suppose for instance that Lois Lane has invited Clark Kent to dinner but he has not said yes or no. So she forms the plan,

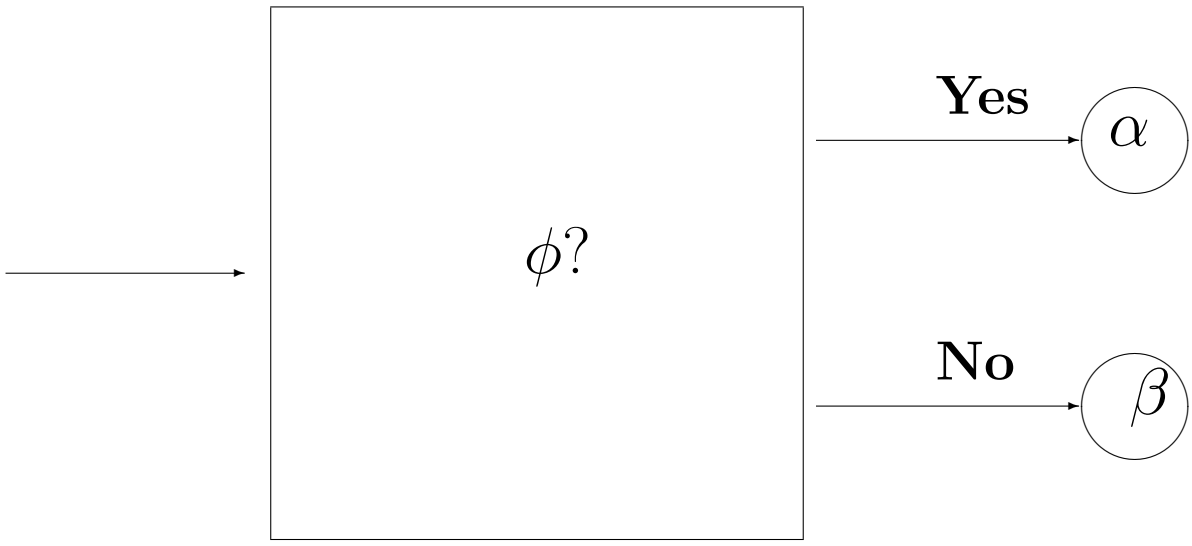
While I do not have a definite answer one way or another, *if* I see Clark Kent, I will ask him if he is coming to dinner.

Here *seeing Clark Kent* is understood to consist of an observation followed by the addition of the sentence "I am seeing Clark Kent" to her stock.

Suppose now that she sees Superman standing on her balcony. She will *not* ask him if he is coming to dinner as the sentence "I am seeing Clark Kent" will not be in her stock of sentences. And this is the sense in which she does *not* know that when she is seeing Superman, she is also seeing Clark Kent. If she *suspects* that Clark Kent is Superman, then it may happen that her recursive call to the procedure "decide if I am seeing Clark Kent" will take the form of the question, "Are you by any chance Clark Kent, and if so, are you coming to dinner?" addressed to Superman.

Definition 0.2 *If an agent a comes to a point in her plan where her appropriate action is* If ϕ then do α else do β, *and she does* α, *then we will say that she* i-believes ϕ. *If, moreover,* ϕ *is true, and we believe that in a similar context she would judge it to be true only if it* is *true, then (within the context of this plan) we will say that she* i-knows ϕ.

A common example of such a plan is the plan to answer a question correctly. Thus if an agent is asked "Is ϕ true?", the agent will typically call the procedure "decide if ϕ is true", and then answer "yes", "no", or "I don't know" in the appropriate cases.



If ϕ then α else β

We no longer have the law that if the agent i-knows ϕ and ϕ implies ψ then the agent of necessity i-knows ψ. But if the agent has the resources to decide ϕ and the proof of ψ from ϕ is easy, then she might well also know ψ. But her notion of "easy" may be different from ours, and how much effort she devotes to this task will depend on her mood, how much energy she has, etc.

Many Agent States of Knowledge

In a study of the genesis and development of grooming, Plooij found that infant chimpanzees of two to four months of age "request" grooming from the mother without first "checking to see" whether she is "looking" at them. At that age they simply extend and arm or leg toward the mother. At about ten and a half months, however, the infant looks into the mother's eyes, establishes that she is looking at it, and then extend an arm or leg toward her. In other words, Plooij established that making eye contact is an early pre-condition for social interaction among chimpanzees.

Premack and Premack, [36] p. 139

Ann is sitting on a chair in front of which there is a vase with a dozen roses in it. Bob can see both Ann and the roses. Charlie can see Ann and Bob and the roses.

We could now ask:

- *Does Ann know p?* where $p =$ There are roses in front of her. I.e., $K_a(p)$?
- Does Bob know that she knows? $(K_b K_a(p)\,?)$
- Does Charlie know that Bob knows that Ann knows $(K_c K_b K_a(p))\,?$

Both common sense and the corresponding Kripke structure tell us that the answer to all three questions is *yes*. Indeed if they can see each other then p is *common knowledge* among them.

Let us now change the meaning of p. In this new example, Ann, Bob and Charlie are all as before, but what is in front of Ann is not a vase of roses, but a blackboard with the number 1243 written on it. Let p now denote the fact that the number on the blackboard is composite.

Logically the situation is not changed. Since 1243 is composite (113 times 11), this is a necessary truth, Ann knows it, Bob knows that Ann knows it, and Charlie knows that Bob knows that Ann knows it.

But are *we* sure that this is the case? It could be that Ann finds numbers greater than 100 to be a mystery. Or perhaps she *is* actually a number theorist but sexist Bob thinks that she is number-challenged. Or perhaps Bob knows her quite well, but Charlie thinks that Bob is a chauvinist who has a poor opinion of the mathematical abilities of women.

So we are no longer sure that $K_a(p)$, $K_bK_a(p)$ and $K_cK_bK_a(p)$ are all true.

Consider now the following game. Ann is sitting (again) in a chair in front of a blackboard on which the number n is written. In front of her are three buttons, 1, 2, 3. Bob can see her and the blackboard, and Charlie can see both Ann and Bob and the blackboard. Bob and Charlie also have buttons. No one can see the buttons of the other people.

The Game:

Ann should push button 1 if she thinks n is prime, button 2 if she thinks it is composite, and button 3 if she does not know.

If she presses the right button she gets one Euro. If she guesses wrong, she *pays* \$20 310. And if she presses 3, there is no gain or loss.

Bob has four buttons, and he should press a button corresponding to Ann's if he knows which button it is, and he presses button 4 if he does not know. If he guesses right, he gets \$1, if he guesses wrong, he pays \$10, and if he presses 4, no gain or loss.

Charlie has 5 buttons, buttons 1-4 to indicate what he thinks Bob pressed, and button 5 if he does not know. His payments are similar to Bob's.

If p denotes the fact that n is composite, then we ought to have $K_a(p)$, $K_bK_a(p)$ and $K_cK_bK_a(p)$. Thus all three should press button 2, all of them getting \$1.

Will this happen? Not necessarily! As we saw, Ann may not realize that the number is composite, or if she does, Bob might think the number is too big for her to factorize etc. Thus in fact we do not have a definite map from physical situations to Kripke structures. The physical set up leaves out the mental facts, and there are many interpretations (not all of which are Kripke structures) for the *same* physical situation.

So how will the game be played? It depends, even if *some* of the three payers are logically omniscient. But note that Ann's best strategy is to press button 2 regardless of what Bob and Charlie press. *Given that she presses button 2*, Bob's best strategy is to press 2 also, and *given* that they are both pressing 2, Charlie should also play 2.

The *standard* Kripke structure that we get out of the physical situation does not necessarily represent the mental situation, but *it does represent the unique Nash equilibrium.*

A Formalism:

Suppose we have a finite n-agent Kripke structure $\mathcal{M}$. The set of states is W with cardinality m. We use this structure to construct a game $\mathcal{G}$. Each agent is told what $\mathcal{M}$ loooks like. Moreover, each agent has a set of symbols corresponding to the (finitely many) equivalence classes of that agent. I.e. the space W splits into finitely many pieces which are the equivalence classes of the agent's accessibility relation and the agent has a symbol for each such class. Thus each agent has his own alphabet. Let $[s]_i$ be i's symbol (equivalence class) for $s \in W$. Thus $s \sim_i t$ iff $[s]_i = [t]_i$. When the agent sees the symbol, he knows which equivalence class he is in, but not *where* he is in that class.

At any moment of time, some state $s \in W$ is picked with probability $1/m$. Then each agent i is given the symbol $[s]_i$. i is also given a finite set X_i of formulas with the following properties. Only atoms are negated in any formula – there are no other negations in any formula. The only connectives are $\wedge$, $\vee$, K_j, $L_j = \neg K_j \neg$. Every knowledge formula (without common knowledge) can be written in this way with all negations driven in using de Morgan's laws, etc.

If $A \wedge B$ $(A \vee B)$ is in X_i, then so are A, B.

If $K_j(A)$ or $L_j(A)$ is in X_i, then A is in X_j.

At time t, each agent i is asked to mark each formula in X_i of level $t-1$ with a *yes*, or a *no*, or a *don't know*. The process goes on until all formulas have been marked. (We could have made this a one shot game, but the extended form is a bit prettier.)

After this, each agent gets \$1 for each formula **correctly marked**, \$0 for each **don't know**, and is fined $\$(m \times k)$, for each **incorrectly marked** formula, where m is the cardinality of W, and k is the cardinality of the finite set $\bigcup X_i$. A formula marked with *don't know* is not considered marked.

A literal (atomic formula or its negation) is considered correctly marked by i iff it is true and marked *yes*, or false and marked *no*. ("true, false" are relative to $\mathcal{M}, s$.)

Formulas $A \wedge B$ and $A \vee B$ are considered correctly marked by i if the yes/no corresponds to the truth value (of $K_i(A \wedge B)$ and $K_i(A \vee B)$) at state s.

A formula $K_j(A)$ is considered to be correctly marked by i if
either $K_j(A)$ is true and A marked *yes* by j or
$K_j(A)$ is marked *no* and either A is false, or A is not marked *yes* by j.

A formula $L_j(A)$ is considered correctly marked by i if either $L_j(A)$ is marked *yes*, it is true, and A is not marked *no* by j or $L_j(A)$ is false, marked *no*, and A is marked *no* by j.

Each agent may have a strategy for playing this game given by the Kripke structure and the sets X_i. We will say that an n-tuple $S = (s_1, ..., s_n)$ of strategies is **safe** for i if i does not have a negative expected value. It is **safe** if no agent makes an expected loss.

Clearly the strategy where some agent says *don't know* for all formulas, is safe for him. Indeed the *don't know* strategy is safe regardless of how the other agents play.

On the other hand, a strategy where an agent says *yes* for a formula A when he does not know A (i.e., where $s \models \neg K_i(A)$), can never be safe, because if he does not know A,, then there is probability at least $1/m$ that he is wrong once, and his loss of $m \times k$ will make up for all possible gains from other cases where he is accidentally right.

Definition: A **knowledge state** for n-agents is a set of *safe* strategies for them. A **belief state** is a set of not necessarily safe strategies.

A state where **every** agent marks formulas according to their knowledge value is safe.

Bob could have a false belief that Ann does not know that 1243 is composite. That is not (on the face of it) a false belief about the world, *but it is a false belief nonetheless. And if Bob has such a false belief, he will make a bad move and pay for it in our game.*

Theorem: The only Nash equilibrium is where each agent marks each formula correctly according to its value at s, where A is considered to be correctly marked by A if it is marked *yes* and $s \models K_i(A)$ or it is marked *no* and $s \models K_i(\neg A)$. (Formulas A where the agent does not know whether A should be marked with a *don't know*).

Proof: Straightforward by induction on formula complexity.

Definition: Let the **knowledge depth** $d(A)$ be the maximum length of a chain of embedded knowledge operators (K or L) in formula A. We will say that a strategy s of some agent is l-complete if the agent correctly marks all formulas of knowledge depth at most l.

Lemma: Suppose all agents other than i are ℓ-complete. Then agent i can safely be $(\ell + 1)$-complete.

Thus for agent i to infer to level $\ell + 1$ it is sufficient that other agents do infer to level ℓ. In the Ann, Bob, Charlie example, if Ann correctly infers facts (that p is true) then Bob can safely infer one level higher, and if he does, then Charlie can safely infer two levels higher.

Thus there can be evolution towards the Nash equilbrium as follows. Each agent can safely start by marking true all knowledge-free formulas which the Kripke structure says they know, and marking false all knowledge-free formulas which the Kripke structure says they know to be false. They are not dependent on other players being intelligent.

Suppose now that all the agents proceed from some level ℓ to $\ell + 1$. They are still safe since all agents were ℓ-complete. In a finite number of steps, they will arrive at a stage where all formulas A where agent i knows *whether* A according to the Kripke structure, have been marked. Now the agents have earned the maximum they possibly could and the Nash equilibrium has been reached.

We can make a stronger assertion. Starting with the strategy where all agents say *don't know* all the time, there is a sequence of changes where at each stage, *only one* agent changes his valuation of *one formula*, and which ends up with the Nash equilibrium. Moreover, no agent is unsafe at any stage of these transformations.

What about **common knowledge**? We could extend the game by saying that Ann and Bob can mark the formula $C_{a,b}(p)$ *yes*, provided it is true in the conventional sense and they *both* mark it *yes*. But now there is no individually safe way to proceed to this situation! They must do it together.

However, if the Kripke structure $\mathcal{M}$ does satisfy $C_{a,b}(p)$, then for each formula A of the form $K_aK_bK_a....K_b(p)$ (for example) there is a way for the two agents to proceed to a stage where both agents mark A with *yes*.

We can now consider the case where some agents are – or are believed to be, logically deficient by other agents. Thus suppose that of agents 1,2,3, agents 1 and 2 are logically adequate, but they know that agent 3 has no notion that other people even have minds (perhaps he is autistic). All three are looking at a vase of flowers. Let p stand for *There is a vase of flowers.* Then p will be common knowledge among 1 and 2, and in fact, *that 3 knows* p will be common knowledge among 1 and 2. But p cannot be common knowledge among $\{1,2,3\}$, for 3 has no notion of what 1 and 2 are thinking! For example, 1 cannot mark $K_3K_1(p))$ *yes*, because he cannot count on 3 marking $K_1(p)$ *yes*. Thus the formula $K_1K_3K_1(p)$ fails to be true not because of a deficiency in 1, but because of a deficiency in 3.

Thus there will be a sort of Nash equlibrium where agents 1, 2 are doing their best *given* 3's deficiency!

Conclusion: We have defined a more general set of knowledge states than those provided by Kripke structures. Hopefully, this more flexible notion will allow us to address various puzzles like that of the **No Trade theorem**, or the issue of mathematical knowledge.

Acknowledgements: We thank Sergei Artemov, Samir Chopra, Jill Cirasella, Horacio Arlo Costa, Juliet Floyd, Haim Gaifman, Isaac Levi, Mike Levin, Ruth Marcus, Larry Moss, Eric Pacuit and Catherine Wilson for comments. The information about chess came from Danny Kopec. This research was supported by a grant from the PSC-CUNY faculty research assistance program.

Earlier versions of the second part of this talk were given at *TARK-05*, *ESSLLI-2006*, at the *Jean Nicod Institute*, and at a seminar in the philosophy department at Bristol University. Some of the research for this paper was done when the author was visiting Boston University and the Netherlands Institute for Advanced Study.

References

[1] Carlos Alchourron, Peter Gärdenfors and David Makinson, "On the logic of theory change: partial meet contraction and revision functions", *J. Symbolic Logic* 50 (1985) 510–530

[2] Artemov, S., and E. Nogina, "On epistemic logics with justifications", *Theoretical Aspecits of Rationality and Knowledge*, ed. Ron Meyden, University of Singapore press, (2005), 279–294.

[3] Aumann, R., "Agreeing to disagree", *Annals of Statistics*, **4** (1976) 1236-1239.

[4] Cristina Bicchieri, "Learning to co-operate," in *The Dynamics of Norms*, ed. Biccheri, Jeffrey, Skyrms, Cambridge1997, pp. 17-46.

[5] Brandom, Robert, "Unsuccessful seminatics", *Analysis*, **54** (1994) 175-178.

[6] Chopra, S., and L. White, "Attribution of Knowledge to Artificial Agents and their Principals", *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, Edinburgh, August 2005, forthcoming.

[7] Michael A. E Dummett *The justification of deduction*, (Henriette Hertz Trust annual philosophical lecture ; 1973).

[8] Fagin, R., Halpern, J., Moses, Y. and Vardi, M., *Reasoning about knowledge*, M.I.T. Press, 1995.

[9] de Finetti, Bruno, "Foresight: its logical laws, its subjective sources", *Annales de l'Institut Henri Poincare*, **7** (1937).

Translation by Henry Kyburg in *Studies in Subjective Probability*, ed. Kyburg and Smokler, Krieger publishing company, (1980) 53-118.

[10] Fitting, M., "A logic for explicit knowledge", to appear in proceedings of *Logica 2004*.

[11] Gaifman, H., "Reasoning with limited resources and assigning probabilities to arithmetical statements", *Synthese*, **140** (2004) 97-119.

[12] Hayek, F.A., *Individualism and Economic Order*, University of Chicago Press (1936). See especially chapters II and IV.

[13] David Hume, *A Treatise of Human Nature*, ed. Selby-Brigge, Oxford U. Press 1978, 176-179.

[14] Daniel Kahneman, "Maps of Bounded Rationality," Nobel prize lecture, 2002.

[15] S. Kripke, "A Puzzle about belief," in *Meaning and Use*, ed. A. Margalit, Reidel 1979.

[16] I. Levi, *The Enterprise of Knowledge*, MIT press 1980.

[17] R. Marcus, "Some revisionary proposals about belief and believing," *Philosophy and Phenomenological Research*, **50** (1990) 133-153.

[18] R. Marcus, "The anti-naturalism of some language centere accounts of belief," *Dialectica*, **49** (1995) 112-129.

[19] M. MicKinsey, "The semantics of belief ascriptions," *Nous* **33:4**, (1999) 519-557.

[20] Ruth Millikan, "Styles of Rationality", in *Rationality in Animals*, ed. M. Nudds and S. Hurley, Oxford 2006, pp. 117-126.

[21] Y.Moses, "Resource bounded knowledge," in *Proc. Theorectical Aspects of Reasoning about Knowledge*, ed. M. Vardi, Morgan Kaufmann 1988, pp. 261-276.

[22] Nozick, R., *Philosophical Explanations*, Harvard University Press, 1981.

[23] R. Parikh, "Knowledge and the Problem of Logical Omniscience" *ISMIS- 87* (International Symp. on Methodology for IntelligentSystems), North Holland (1987) 432-439.

[24] R. Parikh, "Finite and Infinite Dialogues", in the *Proceedings of a Workshop on Logic from Computer Science*, Ed. Moschovakis, MSRI publications, Springer 1991 481-498.

[25] R. Parikh, "Propositions, propositional attitudes and belief revision" in K. Segerberg, M. Zakharyaschev, M. de Rijke, H. Wansing, editors, *Advances in Modal Logic, Volume 2*, CSLI Publications, 2001.

[26] R. Parikh, "Logical omniscience" in *Logic and Computational Complexity*, LNCS 960, 22-29.

[27] R. Parikh, Knowledge based computation (Extended abstract), in *Proceedings of AMAST-95*, Montreal, July 1995, Edited by Alagar and Nivat, Lecture Notes ib Computer Science no. 936, 127-42.

[28] R. Parikh, "Logical omniscience", in *Logic and Computational Complexity* Ed. Leivant, Springer Lecture Notes in Computer Science no. 960, (1995) 22-29.

[29] R. Parikh, "Social Software", *Synthese*, **132**, Sep 2002, 187-211.

[30] R. Parikh, Levels of knowledge, games, and group action, in *Research in Economics*, **57**, (2003) 267-281.

[31] Irene Pepperberg, "Talking with Alex: Logic and Speech in Parrots; Exploring Intelligence," *Scientific American Mind,* August 2004.

[32] R. Parikh, and P. Krasucki, "Levels of knowledge in distributed computing", *Sadhana - Proc. Ind. Acad. Sci.* **17** (1992) 167-191.

[33] Pacuit, E., and R. Parikh, A Logic for communication graphs, to appear in the proceedings of *DALT 2004*.

[34] R. Parikh, and R. Ramanujam, A Knowledge based Semantics of Messages, in *J. Logic, Language and Information*, **12**, (2003) 453-467.

[35] *Meno*, by Plato, translation by Benjamin Jovett. Available online at http://classics.mit.edu//Plato/meno.html

[36] David and Ann Premack, *Original Intelligence*, McGraw-Hill (2003).

[37] Ramsey, F. P., "Facts and propositions", in *Philosophical Papers*, edited by D.H. Mellor, Cambridge U. Press 1990, 34-51.

[38] Ramsey, F.P., 'Truth and probability', in *The Foundations of Mathematics*, Routledge and Kegan Paul (1931), 156-198.

[39] A. Rubinstein, *Lecture Notes in Microeconomic Theory*, Princeton 2006.

[40] L. J. Savage, *The Foundations of Statistics*, Wiley 1954.

[41] S. Schiffer, "Propositional content", in *The Oxford Handbook of Philosophy of Language*, ed. E. Lepore and B. Smith, OUP 2006, 267-294.

[42] E. Schwitzgebel, "A phenomenal, dispositional account of belief," *Nous*, **36:2** (2002) 249-275.

[43] Stalnaker, R., *Context and Content*, Oxford University Press, 1999.

[44] Frans de Waal, *Our Inner Ape*, Penguin 2005.

[45] Whyte, J.T., "Success semantics", *Analysis*, **50** (1990), 149-157.

[46] Wittgenstein, L., *Philosophical Investigations*, MacMillan, 1958.