Introduction to Formal Epistemology

May 12 - 17, ESSLLI 2007

Eric Pacuit   Rohit Parikh

July 3, 2007

Welcome to *Introduction to Formal Epistemology.* The course will consist of five 90 minute lectures roughly organized as follows:

Day 1: Introduction, Motivation and Basic Models of Knowledge

Day 2: Knowledge in Groups and Group Knowledge

Day 3: Reasoning about Knowledge and ....

Day 4: Logical Omniscience and Other Problems

Day 5: Reasoning about Knowledge in the Context of Social Software

This document contains an extended outline of the course including a bibliography. The idea of this reader is to provide a bird's eye view of the literature and to list the main examples, definitions and theorems we will use throughout the course. As a consequence, expository text will be kept to a minimum. The website for the course is

`staff.science.uva.nl/∼epacuit/formep_esslli.html`

On this website you will find the lecture notes and slides (updated each day). Enjoy the course and please remember to ask questions during the lecture, point out any mistakes and/or omitted references in this text.

The goal of this course is to introduce students to the field of formal epistemology. Although formal methods will be used, the focus of the course is not technical but rather on intuitions and the main conceptual issues (such as the logical omniscience problem). As such, there are no prerequisites for this course except some mathematical maturity.

Remark: *The text here is preliminary and may be updated before the course. Please check the website for the most up-to-date version.*

July 3, 2007

1 Introduction and Motivation

'Formal epistemology' is an umbrella term used to describe a field focused on formal methods from logic, probability and computability theory to study traditional epistemological problems. Researchers such as J. Hintikka, D. Lewis, R. Stalnaker, T. Williamson and others have repeatedly demonstrated that formal tools can guide and develop important philosophical insights. Nonetheless, much of the research on formal models of knowledge and beliefs during the latter half of the 20th century was motivated by applications in Game Theory and Computer Science. Recently, the focus has shifted back to Philosophy with an interest in "bridging the gap between formal and mainstream epistemology" (cf. Hendricks (2006) for a collection of essays on this topic).

For this course we will set our sights on yet a different domain: the Social Sciences. The main idea is that the formal models developed to reason about the knowledge and beliefs of a group of agents can be used to deepen our understanding of social interaction and aid in the design of successful social institutions. Our goal for this course is to provide a critical introduction to these formal models of (multi-agent) knowledge and beliefs focusing on how such models can fit into a larger theory of *Social Software* (Parikh, 2002) as outlined in (Parikh, 2007b,a) and (Pacuit and Parikh, 2006).

1.1 Social Software

Social software is an emerging interdisciplinary field devoted to the design and analysis of social procedures. First discussed by Parikh (2002), social software has recently gained the attention of a number of different research communities, including computer scientists, game theorists, social choice theorists and philosophers. The key idea is to analyze social procedures as rigorously and systematically as computer software is pursued by computer scientists. The main objective is a new theory of social interaction that is informed by results from the disciplines listed above *and* providing new insights for the analysis of social procedures.

Social procedures, such as fair division algorithms (see Brams and Taylor, 1996) or voting procedures (see Brams and Fishburn, 1994; Saari, 2001), have been analyzed in detail by mathematicians and political scientists. The analysis typically focuses on comparisons between the *mathematical* properties of various procedures. This is certainly a crucial part of social software, but one of the main goals of social software is to place these issues in the context of a larger discussion[1]

[1] This is not to say that discussions on fair division algorithms and voting procedures do not pay attention to the "larger picture". The point is that techniques from logics of programs and logics of knowledge can be useful in this discussion.

on "designing a *society*" or more modestly, "designing a good social procedure". When it comes to people taking part in social algorithms, a number of factors enter. They are

- The Logical Structure of the Procedure
- Communication and Knowledge
- Preferences and Incentives
- Co-ordination and Conflict
- Culture and Tradition

See (Parikh, 2007a) for an extended discussion of this point.

1.2 Motivating Examples

The following examples represent the type of situations we would like our formal models of knowledge and belief to be able to handle.

Knowledge Based Obligations

Example 1: Uma is a physician whose neighbour is ill. Uma does not know and has not been informed. Uma has no obligation (as yet) to treat the neighbour.

Example 2: Uma is a physician whose neighbour Sam is ill. The neighbour's daughter Ann comes to Uma's house and tells her. Now Uma does have an obligation to treat Sam, or perhaps call in an ambulance or a specialist.

The difference between Uma's responsibilities in examples 1 and 2 is that in the second one she has knowledge of a situation which requires action on her part. In the first case, none of us would expect her to address a problem whose existence she does not know of. Thus any decent social algorithm must allow for the provision of requisite knowledge. However, in the example 3 below, it is the agent's own responsibility to *acquire* the proper knowledge.

Example 3: Mary is a patient in St. Gibson's hospital. Mary is having a heart attack. The caveat which applied in case 1) does not apply here. The hospital cannot plead ignorance, but rather it has an obligation to *be aware* of Mary's condition at all times and to provide emergency treatment as appropriate.

In all the cases we mentioned above, the issue of an obligation arises. This obligation is circumstantial in the sense that in other circumstances, the obligation might not apply. Moreover, the circumstances may not be fully known. In such a situation, there may still be enough information about the circumstances to decide on the proper course of action. If Sam is ill, Uma needs to know that he is ill, and the nature of the illness, but not where Sam went to school.

Such knowledge issues arise all the time in real life. Suppose a shy student is in your office and you wonder if it is time for your next appointment. If you look at your watch, then you will know the time, but the student will also realize that you *wanted* to know the time, and may, being shy, leave even though he need not.

Levels of Knowledge

Suppose that Ann would like Bob to attend her talk; however, she only wants Bob to attend if he is interested in the subject of her talk, not because he is just being polite. There is a very simple procedure to solve Ann's problem: Have a (trusted) friend tell Bob the time and subject of her talk.

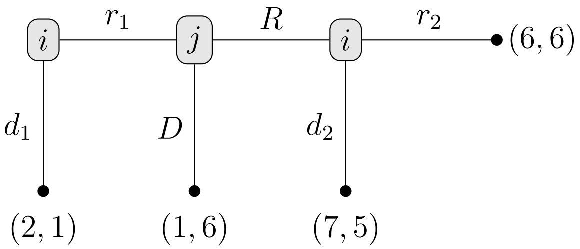
Just as we can show that Quicksort correctly sorts an array, perhaps we can show that this simple procedure correctly solves Ann's problem. While a correct solution for the Quicksort algorithm is easily defined, it is not so clear how to define a correct solution to Ann's problem. If Bob is actually present during Ann's talk, can we conclude that Ann's procedure succeeded? Not really. Bob may have figured out that Ann wanted him to attend, and so is there only out of politeness. Thus for Ann's procedure to succeed, she must achieve a certain level of knowledge between her and Bob. Besides both Ann and Bob knowing about the talk and Ann knowing that Bob knows about the talk, Ann must ensure that

Bob *does not know* that Ann knows about the talk.

This last point is important, since, if Bob knows that Ann knows that he knows about the talk, he may feel social pressure to attend. We now see that the procedure *to have a friend tell Bob about the talk, but not reveal that it is at Ann's suggestion*, will satisfy all the conditions. Telling Bob directly will satisfy the first three, but not the essential last condition.

Knowledge in Strategic Situations

The following example illustrates the role that the above reasoning can play in an analysis of a game-theoretic situation. Assume that there are two players i and j interacting according to the game tree and payoffs given below. Here agent i moves first, choosing between going down (action d_1) and moving right (action r_1). It is then player j's move who also chooses betwen going down (action D) and going right (action R). Finally, agent i has another chance to move choosing again between going down (action d_2) and moving right (action r_2). The payoffs are given as pairs of numbers (x_i, x_j) where the first component is i's payoff and the second j's payoff. So, if i chooses d_1 initially, agent i receives 2 "points" and agent j receives 1 "point". See (Osborne and Rubinstein, 1994) for more details about formal models of games.



Assuming that agents are rational if they choose an action that *guarantees* a higher payoff, then a *backwards induction*[2] argument implies that the only rational choice is for agent i to move down (select D at the first node) thus ending the game with i's receiving a payoff of 2 and j receiving a payoff of 1.

Exercise Give the details of this argument.

This creates a rather strange situation in which we are forced to claim that the only "rational" outcome is one in which *all* players are worse-off. Indeed, if i had a good reason to believe that j (if given the chance) would move R, then clearly i should move R ensuring herself 5. This suggests a refinement in what it means for an agent to be rational— a player is rational if *given its current information*, it acts so as to maximize its payoff. In other words, the rationality of an agent depends not only on its choice of action, but also on its state of knowledge.

Notice that the above game tree does not contain any information about the agents' beliefs or knowledge. Aumann (1999a), (Stalnaker, 1994), (Brandenburger,

[2] See any book on game theory for a discussion.

2005) and others have forcefully argued that adding such information to a description of a game-theoretic situation is important for analyzing the rationality of the agents.

Knowledge and Information

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own. After this is done, they will be asked repeatedly, beginning with Ann, if they know what their own number is.

Which agent can successfully answer (and when) that they know their own number?

2 Basic Frameworks

2.1 Epistemic Logic

The main idea of epistemic logic is to extend the language of propositional logic with symbols ($\Box$) that are used to formalize the statement "the agent knows φ" where φ is any formula. For example, the formula $K\varphi \rightarrow \varphi$ represents the widely accepted principle that agents can only know true propositions, i.e., if φ is known, then φ must be true.[3]

Definition 2.1 (Basic Modal Language) Let $\mathsf{At} = \{p_1, p_2, \ldots\}$ be a countable set of atomic propositions. The **basic modal language** is the smallest set $\mathcal{L}(\mathsf{At})$ generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi$$

where $p \in \mathsf{At}$. We write $\mathcal{L}(\mathsf{At})$ when At is clear from the context. We use the standard abbreviations for the other boolean connectives ($\leftrightarrow, \vee, \bot, \top$) and define $\Diamond\varphi$ as $\neg\Box\neg\varphi$. ◁

Exercise If the intended interpretation of $\Box\varphi$ is "the agent knows φ", what is the interpretation of $\Diamond\varphi$?

[3] Wittgenstein has pointed out in this context that we rarely argue from "Bob knows that φ" to "φ is true", but rather, from "φ is false" to "Bob cannot know φ."

Of course, $\Box\varphi$ is only a formal expression in some language as we have not yet provided a *formal* semantics. The goal is to give a precise interpretation of the expressions in this language that matches as much as possible our informal readings. Almost surely, making precise our intended interpretation will lead to unintended consequences[4]. One of the main goals of this course is to not only give the formal details of the interpretation of the knowledge statements, but also to highlight important idealizations.

Indeed, one may be interested in other motivational attitudes such as beliefs, desires, etc. In this course we focus primarily on the "knowledge" interpretation, but may on occasion prefer the "belief" interpretation. We use the following conventions:

- Write $K\varphi$ if the intended interpretation is "the agent knows φ"
- Write $B\varphi$ if the intended interpretation is "the agent believes φ"
- Write $\Box\varphi$ if the intended interpretation is not crucial

We now turn to the formal semantics for the language defined in Definition 2.1.

Definition 2.2 (Kripke Structures) A **Kripke Frame** is a pair $\langle W, R\rangle$ where W is non-empty set and $R \subseteq W \times W$. A **Kripke Model** based on a frame $\langle W, R\rangle$ is a triple $\langle W, R, V\rangle$ where $V : \mathsf{At} \rightarrow \wp(W)$. ◁

Elements $w \in W$ are called states, or worlds. We write wRv if $(w, v) \in R$. The relation R represents the uncertainty that the agent has about the "actual situation". In other words, if wRv and the actual situation is w, then for all agent i knows, the situation may be v.

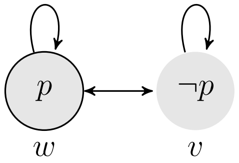
Definition 2.3 (Truth) Suppose $\varphi \in \mathcal{L}(\mathsf{At})$ and $\mathcal{M} = \langle W, R, V\rangle$ is a Kripke Model. We define φ is **true** at state $w \in W$ in model $\mathcal{M}$, written $\mathcal{M}, w \models \varphi$, inductively as follows:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \mathsf{At}$)
- $\mathcal{M}, w \models \varphi \vee \psi$ if $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \Box\varphi$ if for each $v \in W$, if wRv, then $\mathcal{M}, v \models \varphi$

[4] Of course, it is the "unintended consequences" of the formal semantics that is the subject of the most interesting debates.

If $\mathcal{M}, w \models \varphi$ for all states $w \in W$, then we say that φ is **valid in** $\mathcal{M}$ and write $\mathcal{M} \models \varphi$. If φ is valid in all models based on a frame $\mathcal{F}$, then we say φ is **valid in** $\mathcal{F}$ and write $\mathcal{F} \models \varphi$. $\triangleleft$

Example: We illustrate the above semantics with the following simple example. Consider the initial situation from Bob's point of view in the Levels of Knowledge example from Section 1. Before Ann has her friend send a message to Bob, Bob does not know the time of the talk. Suppose p means "the talk is at 2PM". We can represent this situation with the following Kripke structure $\mathcal{M} = \langle \{w, v\}, \{(w, v), (v, w), (w, w), (v, v)\}, V \rangle$ with $w \in V(p)$ and $v \notin V(p)$. This is pictured as follows:



In the above model, we have $\mathcal{M}, w \models \neg K p$, as desired (Bob does not know the time of the talk).

There are a number of principles about knowledge – listed below – expressible in the language of epistemic logic that have been widely discussed by many different communities.

K	$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$	*Kripke's axiom*
T	$\Box\varphi \rightarrow \varphi$	*Truth*
4	$\Box\varphi \rightarrow \Box\Box\varphi$	*Positive introspection*
5	$\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$	*Negative introspection*
D	$\neg\Box\bot$	*Consistency*

Exercise Discuss the plausibility of the epistemic interpretation of the above formulas.

The following technical result will aid in the discussion of the plausibility of the above formulas.

Definition 2.4 (Correspondence) A formula $\varphi \in \mathcal{L}(\mathsf{At})$ **corresponds** to a property P of a Kripke frame $\mathcal{F}$ provided $\mathcal{F} \models \varphi$ iff $\mathcal{F}$ has property P. $\triangleleft$

Lemma 2.5 *Each of the formulas in the left column correspond to properties in the right column.*

Formula valid in the frame	*Property of the relation*
$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	*(valid in all frames)*
$\Box\varphi \to \varphi$	*Reflexive*
$\Box\varphi \to \Box\Box\varphi$	*Transitive*
$\neg\Box\varphi \to \Box\neg\Box\varphi$	*Euclidean*
$\varphi \to \Box\Diamond\varphi$	*Symmetric*
$\neg\Box\bot$	*Serial*

Finally, we end this Section with a technical question. Given a particular class of Kripke frames $\mathbb{F}$, which formulas are valid on this class, i.e., valid in every frame $\mathcal{F} \in \mathbb{F}$?

A Primer on Logics: We assume familiarity with axiomatizations of the propositional calculus (in particular, let MP denote the modus ponens rule). The **necessitation** (N) rule is: *from φ derive $\Box\varphi$. Note that this means that if φ is* **derivable** *then so is $\Box\varphi$.* Given formulas $\varphi_1, \ldots, \varphi_n$, a **logic** $\Lambda(PC, \varphi_1, \ldots, \varphi_n, MP, N)$ denotes the smallest set of formulas closed under the rules MP and N and containing all propositional tautologies and all instances of $\varphi_1, \ldots, \varphi_n$. We write $\vdash_\Lambda \varphi$ iff $\varphi \in \Lambda$. In this case there is a finite list of formulas each of which is an instance of a axiom or follows from previous elements of the list by a rule of the logic. The following are some well-known epistemic logics:

$$\begin{aligned}
\mathbf{S5} &= \Lambda(PC, K, T, 4, 5, MP, N)\\
\mathbf{KD45} &= \Lambda(PC, K, D, 4, 5, MP, N)\\
\mathbf{S4} &= \Lambda(PC, K, T, 4, MP, N)\\
\mathbf{T} &= \Lambda(PC, K, T, MP, N)\\
\mathbf{K} &= \Lambda(PC, K, MP, N)
\end{aligned}$$

Let Γ be a set of formulas and $\mathbb{F}$ a class of frames. For a frame $\mathcal{F}$, we write $\mathcal{F} \models \Gamma$ iff $\mathcal{F} \models \varphi$ for each $\varphi \in \Gamma$. We write $\Gamma \models_{\mathbb{F}} \varphi$ if for each $\mathcal{F} \in \mathbb{F}$, $\mathcal{F} \models \Gamma$ implies $\mathcal{F} \models \varphi$. We write $\Gamma \vdash_\Lambda \varphi$ if there is a derivation of φ in logic Λ using formulas from Γ.

Definition 2.6 (Soundness and Completeness)

- A logic Λ is **sound with respect to a class** $\mathbb{F}$ of Kripke frames if for all formulas φ, $\vdash_\Lambda \varphi$ implies $\models_{\mathbb{F}} \varphi$.

- A logic Λ is **strongly complete with respect to a class of frames** $\mathbb{F}$ provided for each $\Gamma \subseteq \mathcal{L}(\mathsf{At})$, $\Gamma \models_{\mathbb{F}} \varphi$ implies $\Gamma \vdash_{\Lambda} \varphi$.
- A logic Λ is **weakly complete with respect to a class of frames** $\mathbb{F}$ provided $\models_{\mathbb{F}} \varphi$ implies $\vdash_{\Lambda} \varphi$. $\lhd$

Theorem 2.7

- **S5** *is sound and strongly complete with respect to the class of all frames where the relation is an equivalence relation.*
- **S4** *is sound and strongly complete with respect to the class of all frames where the relation is reflexive and transitive.*
- **KD45** *is sound and strongly complete with respect to the class of all frames where the relation is serial, transitive and euclidean.*

Lemma 2.8 *The following formulas and rules are valid on the class of* all *Kripke frames.*

- *From* $\varphi \to \psi$ *infer* $\Box\varphi \to \Box\psi$
- *From* $\varphi \leftrightarrow \psi$ *infer* $\Box\varphi \leftrightarrow \Box\psi$
- $\Box\varphi \wedge \Box\psi \to \Box(\varphi \wedge \psi)$
- $\Box\top$

Exercise The above Lemma has been used to argue that there is an underlying assumption of **logical omniscience** in epistemic logic. Explain.

2.1.1 Pointers to Literature

Modern Epistemic Logic began with Jaakko Hintikka's seminal book *Knowledge and Beliefs: An Introduction of the Logic of the Two Notions* (Hintikka, 1962) (recently extended and republished by Vincent Hendricks and John Symons). A complete history of Epistemic Logic can be found in the article *Epistemic Logic* by P. Gochet and P. Gribomont (Gochet and Gribomont, 2006). The main textbook presentation of Epistemic Logic (focusing on applications in computer science) are (Fagin et al., 1995) and (Meyers and van der Hoek, 1995). See also the chapter on Epistemic Logic by J.-J. Meyer (Meyer, 2001) in the *Blackwell Guide to Philosophical Logic* (Goble, 2001) and (Zanaboni, 1991) for a somewhat different perspective.

Modal Logic: Epistemic Logic is a subarea of *Modal Logic*. The literature on modal logic is much too vast to survey here (see Goldblatt, 2006, for an account of the modern history of modal logic). Consult the *Handbook of Modal Logic* (Blackburn et al., 2006) for the current state of affairs on modal logic (Chapter 18, *Intelligent Agents and Common-sense Reasoning*, (Meyer and Veltman, 2006) and Chapter 20, *Modal Logic for Games and Interaction*, (van der Hoek and Pauly, 2006) contain discussions relevant to this course). There are a quite a few textbooks on modal logic — the most comprehensive is *Modal Logic* (Blackburn et al., 2002).

2.2 Aumann Structures

The approach sketched above for modeling knowledge of agents is *syntactic*. That is, a formal language is developed in which statements about an agent's knowledge about propositional facts about the world can be expressed. These syntactic expressions are then given meaning by *interpreting* them in a Kripke model. This approach was largely developed by Philosophers, Logicians and Computer Scientists. In the Economics and Game Theory literature, a set-theoretic approach was used to model knowledge. One of the first attempts to formalize knowledge in economic situations is by Robert Aumann (Aumann, 1976, 1999a).

As in the previous section, let W be a set of worlds, or states. Let S be the set of all states of nature. A state of nature is a complete description of the *exogenous* parameters (i.e. facts about the physical world) that do not depend on the agents' uncertainties. As noted above, the previous section started with an object language which could express knowledge-theoretic statements about the agents. However, in this section, reasoning about agents is done purely semantically. Thus we are making essential use of the fact that we can identify a proposition with the set of worlds in which it is true. Intuitively, we say that a set $E \subseteq W$, called an **event**, is true at state w if $w \in E$.

Aumann represents the uncertainty of each agent about the actual state of affairs by a partition over the set of states (Aumann, 1999a). Formally, there is a partition Π over the set W. (A partition of W is a pairwise disjoint collection of subsets of W whose union is all of W.). Elements of Π are called **cells**, and for $w \in W$, let $\Pi(w)$ denote the cell of Π containing w. Putting everything together,

Definition 2.9 (Aumann Model) An **Aumann model based on S** is a triple $\langle W, \Pi, \sigma\rangle$, where W is a nonempty set, Π is a partition over W and $\sigma : W \to S$. $\lhd$

So, σ is analogous to a valuation function, it assigns to each world a state of nature in which every ground fact (any fact *not about the uncertainty of the* *agents*) is either true or false. If $\sigma(w) = \sigma(w')$ then the two worlds w, w' will agree on all the facts, but the agents may have different knowledge in them. Elements of W are *richer* in information than the elements of S.

Given a state $w \in W$, the cell $\Pi(w)$ is called the agent's *information set* and an agent *knows event E at state w*, denote $w \in K(E)$, if $\Pi(w) \subseteq E$.

Definition 2.10 (Knowledge Function) Let $\mathcal{M} = \langle W, \Pi, \sigma \rangle$ be an Aumann model. The **knowledge function**, $\mathsf{K} : \wp(W) \to \wp(W)$, based on $\mathcal{M}$ is defined as follows:

$$\mathsf{K}(E) = \{w \mid \Pi(w) \subseteq E\}$$

$\triangleleft$

The obvious question is what is the precise connection between the Aumann models presented in this Section and Kripke models presented in the previous section? The next Lemma is a first step to answering this question.

Lemma 2.11 *Let $\mathcal{M} = \langle W, \Pi, \sigma \rangle$ be a Aumann model and K the knowledge function based on $\mathcal{M}$. For each $E, F \subseteq W$*

$$\begin{array}{ll}
E \subseteq F \Rightarrow \mathsf{K}(E) \subseteq \mathsf{K}(F) & \text{Monotonicity} \\
\mathsf{K}(E \cap F) = \mathsf{K}(E) \cap \mathsf{K}(F) & \text{Closure Under Intersection} \\
\mathsf{K}(E) \subseteq E & \text{Truth} \\
\mathsf{K}(E) \subseteq \mathsf{K}(\mathsf{K}(E)) & \text{Positive introspection} \\
\overline{\mathsf{K}(E)} \subseteq \mathsf{K}(\overline{\mathsf{K}(E)}) & \text{Negative introspection} \\
\mathsf{K}(\emptyset) = \emptyset & \text{Consistency}
\end{array}$$

where $\overline{E}$ means the set-theoretic complement of E (relative to W).

These are the analogues of the $K, T, 4, 5$ and D axiom schemes from the previous section. In fact, there is an obvious translation between Aumann structures and Kripke structures. Halpern (1999) formally compares the two frameworks pointing out similarities and important differences. We end this Section with some of the formal details of this comparison.

Definition 2.12 (Possibility Structures) An **possibility frame** is a pair $\langle W, \mathcal{P} \rangle$ where W is any set, and for each $\mathcal{P} : W \to \wp(W)$ is a function. Given a frame $\mathcal{F} = \langle W, \mathcal{P} \rangle$ and a set of **states** S, an **model based on S** is a triple $\langle W, \mathcal{P}, \sigma \rangle$, where $\sigma : W \to S$. $\triangleleft$

Definition 2.13 (Possibility Operator) Given any possibility function $\mathcal{P} : W \to \wp(W)$, we can associate a **possibility operator** $\mathsf{P} : \wp(W) \to \wp(W)$ defined by

$$\mathsf{P}(E) = \{w \mid \mathcal{P}(w) \subseteq E\}$$

for any subset $E \subseteq W$. $\triangleleft$

Definition 2.14 (Set Operator Properties) Let $\mathsf{P} : \wp(W) \to \wp(W)$ be a set operator and $E, F \subseteq W$. We define the following properties of P:

P1 $\mathsf{P}(E) \cap \mathsf{P}(F) = \mathsf{P}(E \cap F)$

P2 $\cap_{j\in J}\mathsf{P}(E_j) = \mathsf{P}(\cap_{j\in J}E_j)$, for any[5] index set J

P3 $\mathsf{P}(E) \subseteq E$

P4 $\mathsf{P}(E) \subseteq \mathsf{P}(\mathsf{P}(E))$

P5 $\overline{\mathsf{P}(E)} \subseteq \mathsf{P}(\overline{\mathsf{P}(E)})$

P6 $\mathsf{P}(E) \subseteq \overline{\mathsf{P}(\overline{E})}$ $\triangleleft$

Exercise Prove that **P2** follows from **P1**, **P3**, **P4** and **P5**. See Halpern (1999) for a discussion.

Definition 2.15 (Possibility Function Properties) Let $\mathcal{P} : W \to \wp(W)$ be any function. We define the following properties of P:

Reflexive $\forall w \in W$, $w \in \mathcal{P}(w)$

Transitive $\forall w, v \in W$, $v \in \mathcal{P}(w) \Rightarrow \mathcal{P}(v) \subseteq \mathcal{P}(w)$

Euclidean $\forall w, v \in W$, $v \in \mathcal{P}(w) \Rightarrow \mathcal{P}(w) \subseteq \mathcal{P}(v)$

Serial $\forall w \in W$, $\mathcal{P}(w) \neq \varnothing$ $\triangleleft$

Theorem 2.16 (Correspondence, (Halpern, 1999)) *Let $\mathcal{F} = \langle W, P\rangle$ be a frame. Let $\mathsf{P} : \wp(W) \to \wp(W)$ be defined from P as above. Then P satisfies $P2$ (and hence $P1$). Also we have the following correspondance: if P is reflexive, the P satisfies $P3$, if P is transitive, then P satisfies $P4$, if P is Euclidean, then P satisfies $P5$ and if P is serial, then P satisfies $P6$.*

Theorem 2.17 (Set-Theoretic Completeness, (Halpern, 1999)) *Suppose that P is any operator satisfying $P2$, then there is a frame $\langle W, P\rangle$ such that the operator defined from P is exactly P. Moreovere, if P satisfies $P3$, then P is reflexive, if P satisfies $P4$, then P is transitive, if P satisifies $P5$, then P is Euclidean, and if P satisfies $P6$, then P is serial.*

[5] Including, possibly, infinite sets. When $J = \varnothing$, we get $\mathsf{K}(\Omega) = \Omega$

2.2.1 Bayesian Structures

The above models assume a “crisp” notion of uncertainty of the agent, i.e., if w is a state of the world, then any other state $v \in W$ is either in or not in the same information set as w. In some cases it may be desirable to work in a probabilistic setting. Indeed, the “standard” game-theoretic models of knowledge and beliefs, *Harsanyi Type Spaces* (Harsanyi, 1967), are probabilistic[6].

Let W be a set of worlds and $\Delta(W)$ be the set of probability distributions[7] over W. We are interested in functions $p : W \to \Delta(W)$. The basic intuition is that for each state $w \in W$, $p(w) \in \Delta(W)$ is a probability function over W. So, $p(w)(v)$ is the probability the agent assigns to state v in state w. To ease notation we write p_w for $p(w)$.

Definition 2.18 (Bayesian Structure) The pair $\langle W, p\rangle$ is called a **Bayesian frame**, where $W \neq \emptyset$ is any set, and $p : W \to \Delta(W)$ is a function such that

$$\text{if } p_w(v) > 0 \text{ then } p_w = p_v$$

Given a Bayesian frame $\mathcal{F} = \langle W, p\rangle$ and a set of states S, an **Bayesian model based on S** is a triple $\langle W, p, \sigma\rangle$, where $\sigma : W \to S$. ◁

The above condition states that agents never consider the possibility (i.e., assign positive probability) that they might be wrong about their own probability functions. That is, if an agent in state w assigns nonzero probability to a state v, then her probability functions p_w, p_v must be the same. This suggests the following definition:

Definition 2.19 (Type Partition) Given any function $p : W \to \Delta(W)$, we define

$$\Pi(w) \quad := \quad \{v \mid p_w = p_v\}$$

It is easy to see that $\{\Pi(w) \mid w \in W\}$ forms a partition of W. The set $\{\Pi_w \mid w \in W\}$ is called the agent’s **type partition**. ◁

Intuitively, if you knew that the agent was the “type” of person to use proabability p_w at state w and the $\Pi(w)$ is the set of states that the agent is using that particular probability function. The set $\Pi(w)$ can also be thought of as the agent’s information partition at state w. This structure gives a more fine-grained definition of beliefs.

[6] That is not to say that logicians and philosophers ignore probabilistic reasoning. See, for example, Halpern (2003).

[7] We usually think of W as any finite or countable set. When W is infinite, since we are working with probabilities, we need to make some additional measure-theoretic assumptions.

Definition 2.20 (Probabilistic Beliefs) For each $r \in [0,1]$ define $B^r : 2^W \to 2^W$ as follows

$$B^r(E) = \{w \mid p_w(E) \geq r\}$$

$\triangleleft$

Intuitively, $B^r(E)$ is the set of states in which the agent assigns probability at least r to the event E.

Observation 2.21 *We can define a possibility model from a Bayesian model as follows. Let $\langle W, p, \sigma \rangle$ be a Bayesian model on a state space S. We define a possibility model $\langle W, P, \sigma \rangle$ base on S as follows: define $\mathcal{P} : W \to 2^W$ by*

$$\mathcal{P}(w) = \{v \mid \pi_w(v) > 0\}$$

It is easy to see that $\mathcal{P}$ is serial, transitive and Euclidean.

2.2.2 Pointers to Literature

This approach discussed in this section was put forward by Robert Aumann in his classic paper *Aggreing to Disagree* (Aumann, 1976). Aumann then extended this approach in a series of lectures given at the Cowles Foundation for Research in Economics at Yale University in 1989. This culminated with the publication of the article *Interactive Epistemology I: Knowledge* (Aumann, 1999a) (which includes a discussion of the syntactic approach to modeling knowledge). See also Aumann and Heifetz (2001), Halpern (1999), and Bonanno and Battigalli (1999) for general discussions on the set-theoretic models of knowledge.

▶ Add reference for Bayesian Frames: see Aumann (1999b); Bonanno and Battigalli (1999); Ely and Peski (2006) for general discussions.

3 Knowledge in Groups and Group Knowledge

The previous section presented formal models of knowledge for a *single* agent. However, all of the examples we presented in Section 1 involved more than one agent. We now extend the previous models to include more than one agent. Formally, this is completely straightforward. Let $\mathcal{A}$ be a set of agents. We give the details on how to extend Epistemic Logic to the multi-agent setting. The case for Aumann Structures and Bayesian Structures is analogous.

Definition 3.1 (Multi-agent Modal Language) Let $\mathsf{At} = \{p_1, p_2, \ldots\}$ be a countable set of atomic propositions and $\mathcal{A}$ a finite set of agents. The **multi-agent modal language** is the smallest set $\mathcal{L}_n(\mathsf{At})$ generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi$$

where $p \in \mathsf{At}$. We use the standard abbreviations for the other boolean connectives $(\leftrightarrow, \vee, \bot, \top)$ and define $L_i\varphi$ as $\neg K_i \neg\varphi$. $\lhd$

Definition 3.2 (Multi-agent Kripke Structures) A **multi-agent Kripke frame** is a pair $\langle W, \{R_i\}_{i\in\mathcal{A}}\rangle$ where W is non- empty set and $R_i \subseteq W \times W$. A **multi-agent Kripke model** based on a frame $\langle W, \{R_i\}_{i\in\mathcal{A}}\rangle$ is a triple $\langle W, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where $V : \mathsf{At} \to \wp(W)$. $\lhd$

Definition 3.3 (Truth) Suppose $\varphi \in \mathcal{L}(\mathsf{At})$ and $\mathcal{M} = \langle W, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ is a multi-agent Kripke model. We define φ is **true** at state $w \in W$ in model $\mathcal{M}$, written $\mathcal{M}, w \models \varphi$, inductively as follows:

- Boolean connectives and propositional variables are as in Definition 2.3
- $\mathcal{M}, w \models K_i\varphi$ if for each $v \in W$, if wR_iv, then $\mathcal{M}, v \models \varphi$ $\lhd$

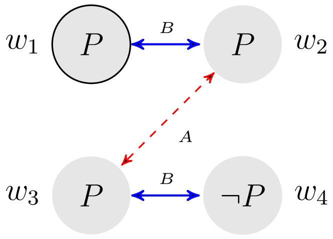
Note that we use K_i for the modal operators instead of $\Box_i$ as we are primarily interested in the knowledge interpretation in this Section.

In this new setting we can express not only what agents know about the world, but also what agents know the other agents know about the world, as so on.

Example: Recall the Levels of Knowledge example from Section 1. It is argued that the state of knowledge that Ann wants to achieve is the following: let p be the proposition 'the talks is at 2PM'.

1. K_Ap: *Ann knows that the talks is at 2PM*
2. K_Bp: *Bob knows that the talks is at 2PM*
3. K_AK_Bp: *Ann knows that Bob knows the talks is at 2PM*
4. $\neg K_BK_AK_Bp$: *Bob doesn't know that Ann knows that Bob knows that the talks is at 2PM*

The following multi-agent Kripke structure represents this situation:



Note that the reflexive arrows are not included in the above picture (so, $w_j R_i w_j$ for each $i \in \{A, B\}$ and $j = 1, 2, 3, 4$). One can check that each of the formulas given above is true at state w_1.

Group and Common Knowledge: You and I approach an intersection. You have the right of way. You should go, I should stop; that is the law and we both know it. However, that is not enough. You should know that *I* know. You do not want to risk your life merely to assert your right of way. Also, I should know that you know I am going to stop. If not, you will stop (even though you have the right of way) and neither of us will go; we do not want a deadlock.

Lewis (1969) and Clark and Marshall (1981) argue that the condition of *common knowledge* is necessary for such co-ordinated actions. Intuitively, a fact p is *common knowledge* if everyone knows p, everyone knows that everyone knows that p, everyone knows that everyone knows that p, and so on. For another example of the relevance of common knowledge, Halpern and Moses (1983) prove that clock synchronisation is impossible without common knowledge. Finally, Chwe (2001) provides many examples suggesting the everyday importance of the notion of common knowledge for co-ordination problems.

3.1 Formalizing Common Knowledge

We start by adding a common knowledge operator to the language of epistemic logic.

Definition 3.4 (Everyone Knows) The operator "everyone knows φ", denoted $E\varphi$, is defined as follows

$$E\varphi \quad := \quad \bigwedge_{i \in \mathcal{A}} K_i \varphi$$

◁

Intuitively, common knowledge of a formula φ (denoted $C\varphi$) is the following infinite conjunction:

$$\varphi \wedge E\varphi \wedge EE\varphi \wedge EEE\varphi \wedge \cdots$$

However, this involves an *infinite* conjunction, so cannot be a formula in the language of epistemic logic. This suggests that common knowledge is not definable in the language of multi-agent epistemic logic[8]. Thus we need to add a new symbol to the language $C\varphi$ whose intended interpretation is "common knowledge of φ".

Definition 3.5 (Multi-agent Epistemic Logic with Common Knowledge) Let $\mathsf{At} = \{p_1, p_2, \ldots\}$ be a countable set of atomic propositions and $\mathcal{A}$ a finite set of agents. The **multi-agent modal language with common knowledge** is the smallest set $\mathcal{L}_n^C(\mathsf{At})$ generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid C\varphi$$

where $p \in \mathsf{At}$. $\triangleleft$

Before giving semantics to $C\varphi$, we consider $EEE\varphi$. This formula says that "everyone knows that everyone knows that everyone knows that φ". When will this be true at a state w in a multi-agent Kripke model? First some notation. A **path on length** n in multi-agent Kripke model is a sequence of states $(w_0, w_2, \ldots, w_n)$ where for each $l = 0, \ldots, n-1$, we have $w_l R_i w_{l+1}$ where i is *any agent*. Thus, $EEE\varphi$ is true at state w iff every path of length 3 starting at w leads to a state where φ is true. The above intuitive definition of common knowledge suggests the following definition.

Definition 3.6 (Interpretation of C) Let $\mathcal{M} = \langle W, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ be a multi-agent Kripke model and $w \in W$. The truth of formulas of the form $C\varphi$ is as follows:

$$\mathcal{M}, w \models C\varphi \quad \text{iff} \quad \text{for all } v \in W, \text{ if } wR^*v \text{ then } \mathcal{M}, v \models \varphi$$

where $R^* := (\bigcup_{i\in\mathcal{A}} R_i)^*$, i.e., R^* is the reflexive transitive closure of the union of the R_i's. $\triangleleft$

Sometimes it is useful to work with the following equivalent characterization of common knowledge:

$$\mathcal{M}, w \models C\varphi \text{ iff every finite path starting at } w \text{ ends with a state satisfying } \varphi.$$

Remark 3.7 (Common Knowledge in a Subgroup) Note that $C\varphi$ means 'φ is common knowledge among the *entire group* $\mathcal{A}$ of agents.' It some cases, it may be useful to talk about common knowledge among a subgroup $G \subseteq \mathcal{A}$ of agents. In this case we write $C_G\varphi$. Formally, we restrict the union in Definition 3.6 to the agents in G.

[8] In fact, one can prove this using standard methods in modal logic.

We have extended the language of multi-agent epistemic logic with a new operator. An immediate technical question is how this effects the soundness and completeness results (Theorem 2.7). It turns out that in the presence of common knowledge, the situation is more complex (see Blackburn et al., 2002, for details). The following axiom and rule need to be added to **S5** to deal with the common knowledge operator:

- **Mix**: $C\varphi \rightarrow E(\varphi \wedge C\varphi)$
- **Induction**: from $\varphi \rightarrow E(\psi \wedge \varphi)$ infer $\varphi \rightarrow C\psi$

Exercise Verify that Mix and Induction are valid on the class of all **S5** frames.

Let $\mathbf{S5}^C$ be the logic including the axiom and rules from **S5** and Mix and Induction.

Theorem 3.8 $\mathbf{S5}^C$ *is sound and weakly complete with respect to the class of all Kripke frames where the relations are equivalence relations.*

Common knowledge in Aumann's setting (cf. Section 2.2) can be defined directly using the infinite conjunction above as there is no finitary object language to worry about.

Definition 3.9 (Common Knowledge Set Operators) Let $\mathsf{K}_i : \wp(W) \rightarrow \wp(W)$ be a knowledge operator for each $i \in \mathcal{A}$ (based on a multi-agent Aumann model). First define $\mathsf{K}^m : \wp(W) \rightarrow \wp(W)$ for each $m \geq 1$:

$$\mathsf{K}^1(E) = \bigcap_{i \in \mathcal{A}} \mathsf{K}_i(E) \qquad \text{and} \qquad \mathsf{K}^{m+1} = \mathsf{K}^1\mathsf{K}^m(E)$$

Then define $\mathsf{K}^\infty : \wp(W) \rightarrow \wp(W)$ by

$$\mathsf{K}^\infty(E) = \mathsf{K}^1(E) \cap \mathsf{K}^2(E) \cap \cdots$$

◁

Exercise Prove that for all $i \in \mathcal{A}$ and $E \subseteq W$, $\mathsf{K}_i\mathsf{K}^\infty(E) = \mathsf{K}^\infty(E)$. (cf. Aumann, 1999a, Lemma 2.3).

This exercise suggests an alternative characterization of common knowledge used by Aumann in (Aumann, 1999a). The main idea is nicely illustrated by the following quote:

> *Suppose you are told "Ann and Bob are going together,"' and respond "sure, that's common knowledge." What you mean is not only that everyone knows this, but also that the announcement is pointless, occasions no surprise, reveals nothing new; in effect, that the situation after the announcement does not differ from that before. ...the event "Ann and Bob are going together" — call it E — is common knowledge if and only if some event — call it F — happened that entails E and also entails all players' knowing F (like all players met Ann and Bob at an intimate party).* (Aumann, 1999a, pg. 271, footnote 8)

The following Definitions and Lemma make this informal statement precise.

Definition 3.10 (Self-Evident Event) An event F is **self-evident** if $\mathsf{K}_i(F) = F$ for all $i \in \mathcal{A}$. $\triangleleft$

Definition 3.11 (Knowledge Field) Let $\langle W, \{\Pi_i\}_{i\in\mathcal{A}}, \sigma\rangle$ be a multi-agent Aumann model. For each $i \in \mathcal{A}$, the **knowledge field of** i, denoted $\mathbb{K}_i$, is the family of all unions of cells in Π_i. $\triangleleft$

Lemma 3.12 *An event E is commonly known iff some self-evident event that entails E obtains. Formally, $\mathsf{K}^\infty(E)$ is the largest event in $\cap_{i\in\mathcal{A}}\mathbb{K}_i$ that is included in E.*

Agreeing to Disagree: In 1976, Aumann proved a fascinating result (Aumann, 1976). Suppose that two agents have the same prior probability and update their probability of an event E with some private information using Bayes' rule. Then Aumann showed that if the posterior probability of E is common knowledge, then they must assign the *same* posterior probability to the event E. In other words, if agents have the same prior probability and update using Bayes' rule, then the agents cannot "agree to disagree" about their posterior probabilities.

Definition 3.13 (Posterior Probability) Let $\langle W, \{\Pi_i\}_{i\in\mathcal{A}}, \sigma\rangle$ be a Aumann model and $p \in \Delta(W)$ a **prior probability** (common to all the agents). The **posterior** probability for agent i of an event A is defined as follows: for all $w \in W$

$$q_i(w) \quad := \quad p(A \cap \Pi_i(w))/p(\Pi_i(w))$$

$\triangleleft$

Theorem 3.14 ((Aumann, 1976)) *Given an Aumann model $\langle W, \{\Pi_i\}_{i\in\{1,2\}}, \sigma\rangle$ with a prior proability function $p \in \Delta(W)$. Let r' and r'' be two numbers in $[0,1]$. If it is common knowledge at w that $q_1(w) = r'$ and $q_2(w) = r''$ then $r' = r''$.*

The key idea is that common knowledge arises through communication. Suppose there are two agents who agree on a prior probability function. Suppose that each agent receives some private information concerning an event E and updates their probability function accordingly. Geanakoplos and Polemarchakis (1982) show that if the agents each announce their posterior probabilities and update with this new information, then the probabilities will eventually become common knowledge and the probabilities will be equal.

Parikh and Krasucki (1990) look at the general situation where there may be more than two agents[9] and communication is restricted by a communication graph. They show that under certain assumptions about the communication graph, consensus can be reached even though the posterior probabilities of the agents may not be common knowledge[10]. Before stating their result, some clarification is needed. Note that a communication graph tells us which agent *can* communicate with which agent, but not when two agents *do* communicate. To represent this information, Parikh and Krasucki introduce the notion of a **protocol**. A protocol is a pair of functions (r, s) where $r : \mathbb{N} \rightarrow \mathcal{A}$ and $s : \mathbb{N} \rightarrow \mathcal{A}$. Intuitively, $(r(t), s(t))$ means that $r(t)$ receives a message from $s(t)$ at time t. Say a protocol (r, s) respects a communication graph $\mathcal{G} = (\mathcal{A}, E)$ if for each $t \in \mathbb{N}$, $(r(t), s(t)) \in E$. A protocol is said to be **fair** provided every agent can send a message to any other agent, either directly *or indirectly*, infinitely often[11].

Parikh and Krasucki show that if the agents are assumed to have finitely many information sets, then for any protocol, if the agents send the current probability[12] (conditioned on the agent's current information set) of proposition A, then after a finite amount of time t for each agent i, the messages received after time t will not change i's information set. Furthermore, if the protocol is assumed to be fair (i.e., the communication graph is strongly connected) then all the agents will eventually assign the same probability to A. Krasucki takes the analysis further in Krasucki (1996) and provides conditions on the protocol (and implicitly on the underlying communication graph) which will guarantee consensus regardless of the agents' initial information.

[9] Cave (1983) also considers more than two agents, but assumes all communications are public announcements.

[10] This point was formally clarified by Heifetz in Heifetz (1996). He demonstrates how to enrich the underlying partition space with time stamps in order to formalize precisely when events become common knowledge.

[11] Consult Parikh and Krasucki (1990) for a formal definition of "fairness".

[12] Actually, Parikh and Krasucki consider a more general setting. They assume agents communicate the value of some function f that maps events to real numbers. Intuitively, f need *not* be the probability of an event given some information set. The only condition imposed on f is a convexity condition: for any two disjoint close subsets X and Y, $f(X \cup Y)$ lies in the open interval between $f(X)$ and $f(Y)$. Here closed is defined with respect to the agents information sets. This generalizes a condition imposed by Cave Cave (1983) and Bacharach Bacharach (1985).

3.2 Levels of Knowledge

While it is true that co-ordinated actions and, supposedly, common knowledge do happen, it may also be relevant to consider other levels of knowledge, short of the infinite, common-knowledge, level.[13] Such levels also arise in certain pragmatic situations, e.g. with e-mail or snailmail or messages left on telephones as voice mail. Thus one purpose of these notes (and this course) is to study levels *other* than common knowledge and how they affect the actions of groups.

In typical co-operative situations, even if a certain level of knowledge is needed, a higher level would also do. If Bob wants Ann to pick up the children at 4 PM, it is enough for him to know that she knows. Thus if he sends her e-mail at 2 PM and knows that she always reads hers at 3 PM, he can be satisfied. In such a situation Bob knows that Ann will know about the children in time, or symbolically $K_b(K_a(C))$ and he may feel this is enough. However, if he telephones her at 3 PM instead, this will create common knowledge of C, much *more* than is needed. But no harm done, since in this context, Ann and Bob have the same goals. The following example from (Parikh, 2003) illustrates this point.

Example: Suppose a pedestrian is crossing the street and sees a car approaching him. It happens in many cities, e.g., Boston, Naples, etc., that the pedestrian will pretend not to notice the car, thereby preventing $K_dK_p(C)$ with C representing the car, d being the driver and p the pedestrian. If the driver knew that the pedestrian knew, he might drive aggressively and try to bully the pedestrian into running or withdrawing. But if he does not know that the pedestrian knows, he will be more cautious.

Let S be the situation where a pedestrian is crossing the street and a car is coming. Let S' be the same situation without the car. In S the pedstrian has two options, g, i.e., to go, and n, i.e., to not go. The motorist also has two similar options, G and N. Here are the payoffs for the two in state S.

[13] The following, possibly apocryphal story about the mathematician Norbert Wiener, well known for his absent mindedness, illustrates something even more subtle. At one time the Wieners were moving and in the morning as he was going to work, Mrs. Wiener said to him, "Now don't come home to this address in the evening." And she gave him a piece of paper with the new address. However, in the evening Wiener found himself standing in front of the old address and not knowing what to do – he had already lost the slip of paper with the new address. He went to a little girl standing by and said, "Little girl, do you know where the Wieners have moved to?" The little girl replied, "Daddy, Mom knew what would happen so she sent me to fetch you." The moral of the story, for *us*, is that common knowledge works only if the memory of all parties involved is reliable.

Motorist choices

		G	N
Pedestrian choices	g	(-100,-10)	(1,0)
	n	(0,1)	(0,0)

Note that there are two Nash equilibria: at (g, N) and at (n, G). However, the penalty for the pedestrian (injury or loss of life) to depart from (n, G) is much greater than the penalty for the motorist (fine or loss of license) to depart from (g, N). Thus the equilibrium (g, N) is less stable than (n, G), and this fact creates the possibility for the motorist to 'bully' the pedestrian.

However, if the pedestrian is unaware of the existence of the car, then the picture is much simpler and his payoffs are 1 for g and 0 for n. g dominates n, and once this choice is made by the pedestrian, it is dominant for the motorist to choose N. This is why the pedestrian tries to achieve the state of knowledge represented by the formulas $K_p(C), \neg K_m(K_p(C))$ indicating that the pedestrian knows the car is there but the motorist does not know that the pedestrian knows. The pedestrian chooses the action g, and knowing that the pedestrian will do this, the motorist must choose N. However, if the motorist has a horn, he can change the knowledge situation. The existence of the car becomes common knowledge and thus the possibility for the motorist to bully the pedestrian arises again.

A number of technical questions suggest themselves when one takes the perspective outlined in this section. A number of these are explored in (Parikh and Krasucki, 1992; Parikh, 2003). We give a sample of one of the questions here: *How many levels of knowledge are there of a given fact P?*

Definition 3.15 (Level of Knowledge) Let $\mathcal{A} = \{1, 2, \ldots, n\}$ be a set of agents. The **modal alphabet** based on $\mathcal{A}$ is the set $\Sigma_{\mathcal{A}} = \{\Box_1, \ldots, \Box_n\}$. We will write Σ for $\Sigma_{\mathcal{A}}$ when the set of agents is understood. Let $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathcal{A}}, V \rangle$ be a multi-agent Kripke model, $w \in W$ and P a propositional formula. The **level of knowledge of P at state** w is the set:

$$L(w, P) = \{x \mid x \in \Sigma^*_{\mathcal{A}}, \mathcal{M}, w \models x\varphi\}$$

where $\Sigma^*_{\mathcal{A}}$ is the set of finite strings over $\Sigma_{\mathcal{A}}$. $\triangleleft$

A natural question is what types of sets can arise as levels of some formula in a Kripke model? The answer to this questions depends on the underlying logic.

Fact 3.16 *Let $\Sigma = \{\Box_1, \ldots, \Box_n\}$ and $\mathbb{F}$ a class of multi-agent Kripke frames. If for all formulas φ and all strings $x, y \in \Sigma^*$, $a \in \Sigma$ we have*

$$\models_{\mathbb{F}} xay\varphi \leftrightarrow xaay\varphi$$

Then for all Kripke models based on frames from $\mathbb{F}$ and states w, $xay \in L(w, \varphi)$ iff for all $j \geq 1$, $xa^jy \in L(w, \varphi)$.

Theorem 3.17 *There are countably many levels of knowledge, but uncountably many levels of beliefs.*

Thus there are levels of belief which *cannot* be levels of knowledge, no matter how things are!

3.2.1 Beyond Finite Levels of Knowledge

States in Kripke structures are supposed to be *complete descriptions of the world* including descriptions of the knowledge of the other agents. In other words, they describe the ground facts, the agents' knowledge of these facts, the agents' knowledge of the other agents' knowledge of these facts, the other agents' knowledge of the other agents' knowledge of the other agents' knowledge of these facts, and so on *ad infinitum*. One may wonder whether this description is adequate. Strangely, it turns out that such descriptions are, in general, *not* adequate descriptions of the world, since they do not completely describe an agent's unceratainty. This fact was pointed about by Heifetz and Samet (1998, 1999), Fagin et al. (1999) and Parikh (1991).

This point should be contrasted with the probabilistic models. Under certain technical assumptions, descriptions of all finite levels of probabilistic beliefs is, in general, sufficient to describe all possible states of the worlds. The existence of such a **universal probabilistic belief space** was first shown to exist by Mertens and Zamir (1985). The technical details are beyond the scope of this course (see Bonanno and Battigalli, 1999; Moss and Viglizzo, 2005, and references therein for details).

3.2.2 Critiques on Common Knowledge

▶ **Critiques on the that common knowledge is required for co-ordinated action. Cf., the work of Bacharach (Bacharach, 2006) and others.**

4 Reasoning about Knowledge and ....

4.1 ....and Time

Suppose we fix a social interactive situation involving a (finite) set of agents $\mathcal{A}$. What aspects are relevant for the analysis of social procedures? First of all, since the intended application of our models is to study agents *executing a procedure*, it is natural to assume the existence of a global discrete clock (whether the agents have access to this clock is another issue that will be discussed shortly). The natural numbers $\mathbb{N}$ will be used to denote clock ticks. Note that this implies that we are assuming a finite past with a possibly infinite future. The basic idea is that at each clock tick, or moment, some *event* takes place.

This leads us to our second basic assumption. Typically, no agent will have *all* the information about a situation. For one thing agents are computationally limited and can only process a bounded amount of information. Thus if a social situation can only be described using more bits of information than an agent can process, then that agent can only maintain a portion of the total information describing the situation. Also, the observational power of an agent is limited. For example, suppose that the exact size of a piece of wood is the only relevant piece of information about some situation. While an agent may have enough memory to remember this single piece of information, measuring devices are subject to error. Furthermore, some agents may not *see*, or be aware of, many of the events that take place. Therefore it is fair to assume that two different agents may have different views, or interpretations, of the same situation. We now turn to the formal details of the model. A variant of these models were first defined in (Parikh and Ramanujam, 1985, 2003).

Let Σ be any set of **events**. Given any set X, X^* is the set of finite strings over X and X^ω is the set of infinite strings over X. Elements of $\Sigma^* \cup \Sigma^\omega$ will be called **histories**. Given $H \in \Sigma^* \cup \Sigma^\omega$, $\mathsf{len}(H)$ is the **length** of H, i.e. the number of characters (possibly infinite) in H. Given $H, H' \in \Sigma^* \cup \Sigma^\omega$, we write $H \preceq H'$ if H is a *finite* prefix of H'. If $H \preceq H'$ we call H an **initial segment** of H' and H' an **extension** of H. Given an event $e \in \Sigma$, we write $H \prec_e H'$ if $H' = He$. Finally, let ϵ be the empty string and $\mathsf{FinPre}(\mathcal{H}) = \{H \mid \exists H' \in \mathcal{H} \text{ such that } H \preceq H'\}$ be the set of finite prefixes of the elements of $\mathcal{H}$ and $\mathsf{FinPre}_{-\epsilon}(\mathcal{H}) = \mathsf{FinPre}(\mathcal{H}) - \{\epsilon\}$.

Definition 4.1 (Protocol) Let Σ be any set of events. A set $\mathcal{H} \subseteq \Sigma^* \cup \Sigma^\omega$ is called a **protocol** provided $\mathsf{FinPre}_{-\epsilon}(\mathcal{H}) \subseteq \mathcal{H}$. A **rooted protocol** is any set $\mathcal{H} \subseteq \Sigma^* \cup \Sigma^\omega$ where $\mathsf{FinPre}(\mathcal{H}) \subseteq \mathcal{H}$. $\lhd$

Intuitively, a protocol is the set of all possible ways an interactive situation may evolve. Once the underlying temporal structure is in place, we can add the uncertainty of the agents. The most general models we have in mind are 'forests' with epistemic relations between finite branches.

Definition 4.2 (ETL Structure) An **ETL frame** is a tuple $\langle \Sigma, \mathcal{H}, \{\sim_i\}_{i\in\mathcal{A}}\rangle$ where Σ is a (finite or infinite) set of events, $\mathcal{H}$ is a protocol, and for each $i \in \mathcal{A}$, $\sim_i$ is an equivalence relation on the set of finite strings in $\mathcal{H}$. An **ETL model** based on an ETL frame $\langle \Sigma, \mathcal{H}, \{\sim_i\}_{i\in\mathcal{A}}\rangle$ is a tuple $\langle \Sigma, \mathcal{H}, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle$ where V is a valuation function $V : \mathsf{At} \to 2^{\mathsf{FinPre}(\mathcal{H})}$. $\lhd$

Making assumptions about the underlying event structure corresponds to "fixing the playground" where the agents will interact. The assumptions of interest are as follows: Let $\mathcal{F} = \langle \Sigma, \mathcal{H}, \{\sim_i\}_{i\in\mathcal{A}}\rangle$ be an *ETL* frame. If Σ is assumed to be finite, then we say that $\mathcal{F}$ is **finitely branching**. If $\mathcal{H}$ is a rooted protocol, $\mathcal{F}$ is a **tree frame**. We will be interested in **protocol frames** which satisfy both of these conditions. These are finitely branching trees with epistemic relations between the finite branches.

Remark 4.3 (*Three Equivalent Approaches*) : There are at least two further approaches to uncertainty in the literature. The first, discussed in Parikh and Ramanujam (1985), represents agents' "observational" power. That is, each agent i has a set E_i of events it *can* observe[14]. For simplicity, we can assume $E_i \subseteq \Sigma$ but this is not necessary. A **local view** function is a map $\lambda_i : \mathsf{FinPre}(\mathcal{H}) \to E_i^*$. Given a finite history $H \in \mathcal{H}$, the intended interpretation of $\lambda_i(H)$ is "the sequence of events observed by agent i at H". The second approach comes from Fagin et al. Fagin et al. (1995). Each agent has a set L_i of **local states** (if necessary, one can also assume a set L_e of environment states). Events e are tuples of local states (one for each agent) $\langle l_1, \ldots, l_n\rangle$ where for each $i = 1, \ldots, n$, $l_i \in L_i$. Then two finite histories H and H' are i-equivalent provided the local state of the last of event on H and H' is the same for agent i. From a technical point of view, the three approaches to modeling uncertainty are equivalent (Pacuit (2007) provides the relevant intertranslations). However, they may still be different for modeling purposes.

Agent Oriented Conditions: Various types of agents place constraints on the interplay between the epistemic and temporal relations. We survey some conditions from the literature.

Definition 4.4 (No Miracles) Fix an epistemic temporal frame $\langle \Sigma, \mathcal{H}, \{\sim_i\}_{i\in\mathcal{A}}\rangle$. An agent $i \in \mathcal{A}$ satisfies the property **No Miracles** (sometimes called, somewhat

[14] This may be different from what the agent *does* observe in a given situation.

misleadingly, **No Learning**) if for all finite histories $H, H' \in \mathcal{H}$ and events $e \in \Sigma$ with $He \in \mathcal{H}$ and $H'e \in \mathcal{H}$, if $H \sim_i H'$ then $He \sim_i H'e$. $\lhd$

Thus, unless a 'miracle' happens, uncertainty of agents cannot be erased by the same event. The next condition is the dual property.

Definition 4.5 (Perfect Recall) An agent $i \in \mathcal{A}$ satisfies the property **Perfect Recall** provided for all finite histories $H, H' \in \mathcal{H}$ and events $e \in \Sigma$with $He \in \mathcal{H}$ and $H'e \in \mathcal{H}$, if $He \sim_i H'e$ then $H \sim_i H'$. $\lhd$

Perfect Recall means that the histories an agent considers possible can only decrease or remain the same, unless new indistinguishable events occur.

Definition 4.6 (Synchronized Communication) An agent $i \in \mathcal{A}$ is **synchronized** provided for all finite histories $H, H' \in \mathcal{H}$, if $H \sim_i H'$ then $\mathsf{len}(H) = \mathsf{len}(H')$. $\lhd$

Intuitively, if an agent is synchronized, then that agent knows the value of the global clock (this may or may not be expressible in the formal language). For other assumptions that can be made about the interaction between the epistemic relation and time, the reader is referred to Fagin et al. (1995); van Benthem and Liu (2004). Finally, note that in general we do not assume that all agents have the same reasoning capabilities. When they do, we say, for example, that a frame $\mathcal{F}$ is synchronous if all agents are synchronized.

Modal Languages: Different modal languages can reason about the above structures (see the Handbook chapter Hodkinson and Reynolds (ming)), with 'branching' or 'linear' variants. Here we give just the bare necessities.

Let At be a countable set of atomic propositions. We are interested in languages with various combinations of the following modalities: $P\varphi$ (φ is true *sometime* in the past), $F\varphi$ (φ is true *sometime* in the future), $Y\varphi$ (φ is true at *the* previous moment), $N\varphi$ (φ is true at *the* next moment), $K_i\varphi$ (agent i knows φ) and $C_B\varphi$ (the group $B \subseteq \mathcal{A}$ commonly knows φ). Dual operators are written as usual (eg., $L_i\varphi = \neg K_i \neg \varphi$). If X is a sequence of modalities from $\{P, F, Y, N\}$ let $\mathcal{L}_n^X$ be the language with n knowledge modalities $K_1, \ldots, K_n$ together with the modalities from X. For a sequence of modalities X, $\mathcal{L}_C^X$ is the language $\mathcal{L}_n^X$ closed under the common knowledge modality C. Let $\mathcal{L}_{ETL}$ be the full epistemic temporal language, i.e., it contains all of the above temporal and knowledge operators.

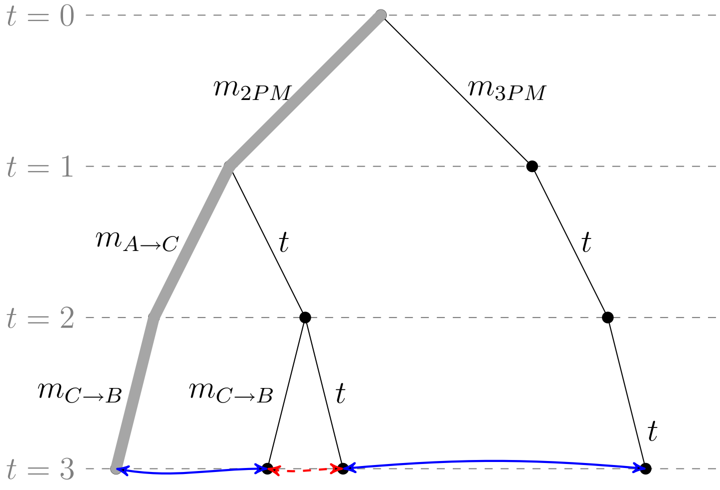
Regardless of whether the language has *branching time* or *linear time* temporal operators, formulas express properties about finite histories. The difference lies in the format of the satisfaction relation. In a linear temporal setting, formulas are interpreted at pairs H, t where H is a 'maximal' (possibly infinite) history and t

an element of $\mathbb{N}$. The intended interpretation of $H, t \models \varphi$ is that *on the branch* H at time t, φ is true. In the branching time setting, we only need the moment, and formulas can be interpreted at finite histories H. In the interest of a unified approach we will interpret formulas at branch-time pairs.

Formulas are interpreted at pairs H, t where $t \in \mathbb{N}$ and $H \in \mathcal{H}$ has length longer than t (finite or infinite). Truth for the languages $\mathcal{L}_n^X$ is defined as usual: see Fagin et al. (1995) and Hodkinson and Reynolds (ming) for details. We only remind the reader of the definition of the knowledge and some temporal operators:

- $H, t \models P\varphi$ iff there exists $t' \leq t$ such that $H, t' \models \varphi$
- $H, t \models F\varphi$ iff there exists $t' \geq t$ such that $H, t' \models \varphi$
- $H, t \models K_i\varphi$ iff for each $H' \in \mathcal{H}$ and $m \geq 0$ if $H_t \sim_i H'_m$ then $H', m \models \varphi$

Example: We again return to the Levels of Knowledge example from Section. In Section 3, we gave an example of a multi-agent Kripke model in which all of the necessary knowledge statements are true. Given the framework in this Section, we can be more precise about *where the Kripke structure comes from.* In the model below, the event t stands for a clock tick, $m_{A\to C}$ is the event that "Ann tells Charles that the talk is at 2PM", $m_{C\to B}$ is the event that "Charles tells Bob the talks is at 2PM", m_{2PM} is the event that the "Ann receives the message that the talk is at 2PM" and m_{3PM} is the event that "Ann receives the message that the talk is at 3PM".



The uncertainty lines are derived according to the following rules: Ann is only aware of the events m_{2PM}, m_{3PM} and $m_{A\to C}$ while Bob (only) is aware of the event $m_{C \to B}$. Let H be the highlighted branch. Then we can check that all of the knowledge statements from Section 3 are true at $H, 3$ (eg., $H, 3 \models \neg K_B K_A K_B P_{2PM}$

▶ Add references

4.1.1 ....and Obligation

▶ Formal details of the Knowledge based obligation examples from (Pacuit et al., 2006)

4.2 ....and Effort

Moss and Parikh (1992) introduce a bimodal logic intended to formalize reasoning about points and sets. This new logic called *Topologic* can also be understood as an epistemic logic with an effort modality. Formally, the two modalities are: K and $\Diamond$. The intended interpretation of $K\varphi$ is that φ is known; and the intended interpretation of $\Diamond\varphi$ is that after some amount of effort φ can become true. For example, the formula

$$\varphi \to \Diamond K\varphi$$

means that if φ is true, then after some "work", $K\varphi$ can become true, i.e., φ is known. In other words, the formula says that if φ is true, then φ can be known with some effort. What exactly is meant by "effort" depends on the application. For example, we may think of effort as meaning taking a measurement, performing a calculation or observing a computation.

There is a temptation to think that the effort modality can be understood as (only) a temporal operator, reading $\Diamond\varphi$ as "φ is true some time in the future". While there is a connection between the logics of knowledge and time and logics of knowledge and effort, following (Moss and Parikh, 1992) we will assume that such effort leaves the base facts about the world unchanged. In particular, in any topologic model, if φ does not contain any modalities, then $\varphi \leftrightarrow \Box\varphi$ is valid. Thus, effort will not change the base facts about the world – it can only change knowledge of these facts.

Given a set W, a subset space is a pair $\langle W, \mathcal{O}\rangle$, where $\mathcal{O}$ is a collection of subsets of W. A point $x \in W$ represents a complete description of the world in which all ground facts are settled, whereas a set $U \in \mathcal{O}$ represents an *observation*. The pair (x, U), called a *neighborhood situation*, can be thought of as an actual situation together with an observation made about the situation. Formulas are interpreted at neighborhood situations. Thus the knowledge modality K represents movement within (consistent with) the current observation, while the effort modality $\Diamond$ represents a refining of the current observation.

Formally,

1. $x, U \models K\varphi$ iff $(\forall y \in U)(y, U \models \varphi)$
2. $x, U \models \Diamond\varphi$ iff $(\exists V \in \mathcal{O})((x \in V \subseteq U)$ and $(x, V \models \varphi))$

Moss and Parikh (1992) provide a sound and complete axiomatization for all subset spaces. Georgatos (1994, 1997) provides a sound and complete axiomatization for subset spaces that are topological spaces and complete lattices. For a complete discussion of topologic and the resulting literature consult (Parikh et al., 2006).

4.3 ....and Communication

The study of *Dynamic Epistemic Logic* attempts to combine ideas from dynamic logics of actions and epistemic logic. The main idea is to start with a formal model that represents the uncertainty of an agent in a social situation, i.e., a Kripke model. Then define an 'epistemic update' operation that represents the effect of a communicatory action, such as a public announcement, on the original model. For example, publicly announcing a true formula φ, shifts from the current model to a submodel in which φ is true at each state. Starting with (Plaza, 1989) and more recently (Baltag and Moss, 2004; Kooi, 2003; van Ditmarsch, 2000; Gerbrandy, 1999; van Benthem, 2002), logical systems have been developed with the intent to capture the dynamics of information in a social situation. See (van Ditmarsch et al., 2007) and (van Benthem, 2002) for a thorough discussion of the current state of affairs.

► Add details of Public Announcement Logic

4.4 ....and Games

In a game-theoretic situation[15] two types of uncertainty can be distinguished: *incomplete* information and *imperfect* information. The former concerns uncertainty about *structure* of the game (eg., available moves, players payoffs, etc.) and the latter concerns uncertainty about the current *stage of the game* (i.e., precisely which moves are currently available to a player).

[15] See Osborne and Rubinstein (1994) for an introduction to game theory. Technical details about game theory that are important for this course will be introduced as needed.

▶ **Add a discussion about Common Knowledge of Rationality and the Backwards Induction Solution. (Aumann, 1995), (Stalnaker, 1996), (Halpern, 2001)**

5 Logical Omniscience and Other Problems

The Single Agent Case: Suppose that Jack knows φ and $\varphi \rightarrow \psi$. Must Jack know ψ in that case? Of course a logically omniscient Jack *would* know. In Plato (Plato) Socrates suggests that Jack *must* know ψ in such a case, but when we follow Socrates' conversation with the slave boy, it is clear that a person who knows φ and $\varphi \rightarrow \psi$ can be *brought* to see ψ, but not necessarily that he already knows ψ. Similarly, the Kripke structure account going back to Hintikka and others suggests that Jack must know ψ, but this account is seen to be defective on this point. For we are well aware that most Jacks, and Jills, are not logically omniscient. It is quite possible that they know φ and $\varphi \rightarrow \psi$, but not ψ. So let us accept this and then proceed.

It is important to note now that the fact that Jack knows φ and $\varphi \rightarrow \psi$, but not ψ is a *contingent* truth. Surely no logic is going to *prevent* Jack from knowing ψ, and hence to discover this fact about Jack, like any other contingent fact, we must resort to observation and experiment. Similarly, the situation that Jack knows φ and $\varphi \rightarrow \psi$ but not ψ cannot be represented by a Kripke structure, as a Kripke structure which represents Jack as knowing φ and $\varphi \rightarrow \psi$, would necessarily represent him as knowing ψ as well. So we need another representation of knowledge besides Kripke structures and also a method for directly *measuring* Jack's knowledge.

Such a method was suggested (for belief) long ago by Ramsey (1931). How do we know that a chicken thinks that a particular caterpillar is poisonous? *It refuses to eat it.* Of course, we might need to know a bit more about the chicken's past before we can attribute such a belief to it, but perhaps we already know that the chicken ate such a caterpillar yesterday and was sick. If the current caterpillar is of the *same kind* as yesterday's, we will say that the chicken knows that it is poisonous, and if it is of a different, benign species, then we will say that the chicken has a false belief that it is poisonous. Such false beliefs can be deliberately created by a supposedly lower form of life.

> *Mimicry is the term applied to the phenomenon presented by certain species which, being themselves eatable, and belonging to groups which are attacked and devoured by numerous enemies, obtain protection by their close resemblance to some of the brightly coloured species which*

> *are free from attack on account of their nauseous odour or general inedibility. In most cases it is not a general but a special resemblance which serves this purpose, sometimes carried so far that the mode of flight and general habits are imitated, as well as colour and marking. The most numerous examples of mimicry occur among butterflies, but there are almost equally remarkable cases among beetles and other orders of insects, as well as a few among reptiles and birds.*
>
> From "Protective Mimicry in Animals"
> by Alfred Russel Wallace, *Science for All,* 1881.

This is a case of what Robin Dunbar (2004) would describe as a second level of mental representation occurring in the animal world. An animal A of species X is creating a false belief in an animal B of species Y (the likely predator). Thus *A knows that B believes "A is dangerous."*

In any case, our knowledge about the chicken is going to be independent of any Kripke structure and will be based on its *behaviour.*

With Jack, we have two methods. If Jack picks up an umbrella as he is going out, and it is indeed raining, we will say that he knows it is raining. We might even ask him and be rewarded with a piece of verbal behaviour (he says, "You know, it is raining!") which corresponds with his action (though it might not). Also, if he picks up his umbrella, but neglects to remind you to take yours, he might be guilty of a lack of consideration, or perhaps a lack of logical omniscience is the culprit here.

It is important to note that non-verbal behaviour will be a response to a proposition, or a state of affairs. A chicken responds to how things are, and if φ and φ' are logically equivalent, then the chicken will act the same way in worlds which satisfy φ as it does in worlds which satisfy φ' – for they are in fact the same worlds.. Verbal behaviour on the other hand is going to be a response to *sentences* and it is quite easy for us to imagine that someone may assent to "Is φ true?" and dissent from or express doubt about a logically (or necessarily) equivalent φ'. Lois Lane may say "Yes" to "Does Superman fly?" and respond with "Are you kidding?" to "Does Clark Kent fly?" Until a few years ago, many of us who willingly assented to $0 = 0$, expressed doubts about Fermat's theorem, even though the two are logically equivalent.

Parikh (2005, 2007c) investigates this phenomenon in detail. An agent's belief state is represented as an element b of a space $\mathcal{B}$ with the property that in a decision situation where an agent has to make a choice, either between two (or more) actions, or between two statements, b gives us a decision. That decision depends also on the agent's space $\mathcal{P}$ of preferences. If Jack takes an umbrella when going out, this action depends not only on his belief that it is raining, but also on his preference for not getting wet. Thus what we have is a map $\rightarrow_{ch}$:

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

Here $\mathcal{S}$ is the space of choice situations, and C is the space of actual choices. Thus {*take umbrella,don't take umbrella*} is an element of $\mathcal{S}$ whereas *take umbrella* is an element of C.

The belief state $b \in \mathcal{B}$ can be revised by several means. It can be revised by witnessing an external event, e.g., raindrops falling on your head; it can be revised by hearing a sentence from someone, "Hey, it is raining!"; or it can be revised by a deduction like, "Didn't I just see Jill coming in with a wet umbrella? It must be raining."

Clearly a theory based on more observations and less theoretical deduction is going to be messier to deal with than the current one based on Kripke structures. But it will be more realistic and more useful.

We have talked so far about belief and not knowledge. The gap between belief and knowledge is important of course. Knowledge requires truth, justification, and that other magic factor which we still do not quite have (Gettier, 1963). However, we canot understand knowledge unless we understand belief as well, and it is important to address both kinds of lack of logical omniscience.

The Many Agent Case: More complex problems arise with many person knowledge. How does Jack know that Jill knows it is raining? Perhaps he saw her take *her* umbrella. But common knowledge of the fact that is raining is going to be harder to measure and we suspect that it does not exist.

Suppose A travelling north sees a green light and also sees a car, driven by B heading east on a cross road. No doubt B will see a red light. At this moment, perhaps A and B have common knowledge that A should go and B should stop. But perhaps it is much simpler, that A is conditioned to go when the light is green and B is conditioned to stop when it is red. Such cultural habits are likely to be the right explanation when there is co-ordinated action which we explain by appealing to common knowledge.

We will now carry out a detailed mathematical investigation of a puzzle from Section 1 which goes back to Littlewood (1953).

Imagine the following situation[16]. Two players Ann and Bob are told that the

[16] This sort of problem has been discussed elsewhere, e.g. Littlewood (1953); van Emde Boas et al. (1984), etc. See Parikh (1991) for a discussion of such dialogues in both the finite and infinite casse.

following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own. After this is done, they will be asked repeatedly, beginning with Ann, if they know what their own number is.

Let us denote the situation where Ann has a and Bob has b as (a, b), and of course $|a - b| = 1$. Consider now the situation (1,2). When Ann is asked if she knows her number, she sees that Bob has a 2 and so her own number must be either 1 or 3. Not knowing which one, she will say, "I don't know". However, if Bob is asked next, he will realise that n must be 1, written on Ann's forehead, with 2 on his, since 0 is not a positive integer.

This argument also leads to a solution for the situation (3,2). Ann will respond as before, since her evidence in the beginning is the same as before. However, this time, Bob will also have to say, "I don't know", and so, when Ann is asked a *second* time, she will realise that the situation is not the same as the one just above, and hence that her number cannot be 1. Since 3 is the only other possibility, she will now say, "My number is 3"

Can we continue this argument beyond 3? If the situation is, say, (4,5), then not only must each party say, "I don't know" at the first stage, the other party must already *expect* this response. Thus Bob, seeing a 4, knows that his own number must be either 3 or 5, and in either case, Ann must say, "I don't know". Similarly, Bob must also say "I don't know" when first asked, and Ann must expect this answer. If the answers were already expected, then how can there be any learning, and if there is no learning, how can there be any progress?

Nonetheless, there is a "proof" by induction on n, that the dialogue will always terminate with one or the other player guessing his/her number. In the following, a *stage* will be a single question. A *round* will therefore consist of two stages.

Theorem 5.1 *In those cases where Ann has the even number, the reponse at the nth stage will be, "my number is $n+1$", and in the other cases, the response at the $(n+1)$st stage will be "my number is $n+1$". In either case, it will be the person who sees the smaller number, who will respond first.*

Proof. By induction on n. We divide the cases into four categories.

$(A)_n$: n is even, Ann has n.In this case, Bob sees n and concludes that his own number is $n-1$ or $n+1$. In the first case, we are in case $(B)_{n-1}$ and by induction hypothesis, if Bob's number is $n-1$, then Ann should guess her own number at stage $n-1$. Since she said "I don't know my number", Bob realises that his number is not $n-1$ and hence must be $n+1$, which he will say at the next stage, i.e. $n+1$.

(B)$_n$: n is odd, Bob has n. If n is 1, then at the very first stage, Ann, seeeing a 1, will say, "my number is 2". If $n > 1$, then we reduce to the case (A)$_{n-1}$ as above.

(C)$_n$: n is even, Bob has n. Ann knows that her number is $n - 1$ or $n + 1$. If it were $n - 1$, Bob would say at stage n that his number is n. Hence, when Bob says "I don't know my number", she realises that she is in case (C)$_n$ rather than in (D)$_{n-1}$ and at the next stage she guesses her number.

(D)$_n$: n is odd, Ann has n. This case is like the case (B). Note that if n is 1, then the number will be guessed at stage 2, since that is Bob's first chance to speak.

QED

However, there is a gap in this argument in that both Ann and Bob's reasoning depends heavily on what the other one is thinking, including a consideration of what the other does not know. Ann's reasoning is justified if Bob thinks as she believes he does, and Bob's reasoning is justified if she thinks as he believes she does. But there is no guarantee that they do indeed think this way. How do we justify what each thinks and what each does and does not know?

In order to deal with this question we need some apparatus. We will use the Kripke structures as defined in Section **??**

In the example we are looking at, the set of states $W = \{(m, n) | m, n \epsilon N^+$ and $|m - n| = 1\}$. If $s, t \in W$ and $i \in \{1, 2\}$, then sR_it iff $(s)_j = (t)_j$, where $j = 3 - i$, and $(s)_j$ is the j-the component of s. Intuitively, sR_it means that when the dialogue begins, player i cannot distinguish between s and t, where Ann is player 1 and Bob is player 2.

Definition 5.2 (Closed Sets) A subset X of W is ***i*-closed** if $s \in X$ and sR_it imply that $t \in X$. X is **closed** if it is both 1-closed and 2-closed. ◁

Using this notion, we can give an alternative characterization of common knowledge (cf. Definition 3.6).

Definition 5.3 (Common Knowledge, again) Given multi-agent Kripke model $\mathcal{M} = \langle W, \{R_i\}_{i\in\mathcal{A}}, V\rangle$, $X \subseteq W$, and $s \in X$, then i **knows** X at s iff for all t, sR_it implies that $t \in X$. X is **common knowledge** at s iff there is a closed set Y such that $s \in Y \subseteq X$. ◁

What we have given above amounts to endorsing a Kripke structure account of knowledge, which justifies the theorem above, but which we have already found defective. Now we proceed to a somewhat more realistic and more behavioral account of how Ann and Bob may actually proceed.

Definition 5.4 (Interactive Discovery System) An *IDS* **(interactive discovery system)** for M is a map $f : W \times N^+ \rightarrow \{\text{“}no\text{”}\} \cup W$ such that for each odd n, $f(s,n)$ (Ann’s response at stage n) depends only on the R_1 equivalence class of s and on $f(s,m)$ for $m < n$. For each even n, $f(s,n)$ depends only on the R_2 equivalence class of s and on $f(s,m)$ for $m < n$. ◁

Thus, e.g. if n is odd and sR_1t and for all $m < n$, $f(s,m) = f(t,m)$ then $f(s,n) = f(t,n)$.

The answer “no” means “I don’t know my number”, whereas saying one’s own number is equivalent to giving the full state. We shall refer to “no” as the trivial response. Any other response will be non-trivial.

Definition 5.5 (Sound IDS) The IDS f is **sound** if for all s, if $f(s,n) \neq \text{“}no\text{”}$, then $f(s,n) = s$. We define $i_f(s) = \mu_n(f(s,n) \neq \text{“}no\text{”})$ and $p(s) = 1$ if $i_f(s)$ is odd and 2 if $i_f(s)$ is even. (Here μ stands for “least”. $i_f(s) = \infty$ if $f(s,n)$ is always “no”. We may drop the subscript f from i_f if it is clear from the context.) ◁

Note that we allow people to be ignorant even when they should not be, but a sound IDS requires that all nontrivial responses be correct. Thus the IDS which takes the constant value “no” is sound though it may not be very interesting. The IDS used by most non-mathematicians may correspond to the strategy, “if you see a 1, then say 2. If you see a 2 and the other player has already said ‘no’, then say 3. Otherwise say ‘I don’t know’ ”. This strategy is also sound, but not optimal. Without loss of generality we will confine ourselves to functions f where the dialogue after any non-trivial response is constant. I.e. if one person says the state s, then it is s thereafter.

Lemma 5.6 *Let f be a sound IDS. Let sR_it, $i(s) = k < \infty$ and $p(s) = i$. Then $i(t) < k$ and $p(t) \neq i$.*

Proof. At stage $i(s)$, i has evidence distinguishing between s and t. Since all previous utterances associated with s were “no”, some previous utterance associated with t must have been nontrivial. Formally, $f(s, i(s)) = s \neq f(t, i(s))$. But sR_it. Hence $(\exists m < i(s))(f(s,m) \neq f(t,m))$. Since $m < i(s)$, $f(s,m) = \text{“}no\text{”}$ and so $f(t,m) \neq \text{“}no\text{”}$. Thus $i(t) \leq m < i(s)$. Now, if $p(t) = i$, then, by a symmetric argument, we could prove also that $i(t) < i(s)$. But this is absurd. Hence $p(t) \neq i$. QED

Corollary 5.7 *Suppose that $p(s) = i$ and there is a chain $s = s_1R_1s_2R_2s_3R_1 \cdots s_m$. Then $i(s) \geq m$.*

Proof. $i(s_m) \geq 1$. Now we can show using induction on k and lemma 1, that $i(s_{m-k}) \geq k+1$. For we have $i(s_{m-k}) > i(s_{m-k+1}) = i(s_{m-(k-1)}) \geq k$ (by induction hypothesis). Taking $k = m-1$ we get $i(s_1) \geq m$. QED

Corollary 5.8 *Suppose that there is a chain $s_1 R_1 s_2 R_2 s_3 R_1 ... s_m R_2 s_1$, with $m > 1$. Then $i(s_i) = \infty$ for all i.*

Proof. If, say, $i(s_1) = k < \infty$, we would get $i(s_1) > i(s_2) > ... > i(s_m) > i(s_1)$, a contradiction. QED

Remark 5.9 (What does Theorem 5.1 say?) We now return to a discussion of the proof of theorem 5.1 above. The theorem is really a proof that the IDS f is sound where f is defined by:

Ann's strategy: If you see 2n+1, then say n "no"'s and then, if Bob has not said his number, say "2n+2". If you see 2n, then say n "no"'s and if Bob has not said his number, say "2n+1".

Bob's strategy: If you see 2n+1, then say n "no"'s and then, if Ann has not said her number, say "2n+2". If you see 2n, then say n "no"'s and if Ann has not said her number, say "2n+1".

These strategies yield: $i(2n+2, 2n+1) = 2n+1$, $i(2n, 2n+1) = 2n$, $i(2n+1, 2n+2) = 2n+2$ and $i(2n+1, 2n) = 2n+1$. In other words, the smaller number if Ann's number is even, and the bigger number if it is odd. These strategies are *optimal*. E.g. we have

$$(6,5) \quad R_1 \quad (4,5) \quad R_1 \quad (4,3) \quad R_1 \quad (2,3) \quad R_2 \quad (2,1)$$

and hence $i(6,5)$ has a minimum value of 5, the value achieved by the strategy above.

Theorem 5.10 *The strategies implicit in theorem 1 and described in remark 1 are optimal. I.e. if h is any other sound IDS, then $i_f(s) \leq i_h(s)$ for all s.*

Proof. By cases. Suppose, for example, that Ann has an even number and $s = (2n, 2n-1)$. $i_f(s) = 2n-1$. Suppose Bob is the one who first notices the state. Then we have $(2n, 2n-1) R_2 (2n, 2n+1) R_1 (2n+2, 2n+3)$..., and by lemma 1, $i_h(s)$ could not be finite. So Ann *does* first discover s. But then we have $(2n, 2n-1) R_1 (2n-2, 2n-1) R_2 (2n-2, 2n-3) \cdots R_2 (2,1)$ and so, by lemma 1, $i_h(s) \geq 2n-1$. QED

Remark 5.11 (Common Knowledge of the IDS) We have not said whether the IDS f itself is common knowledge between Ann and Bob. The reason is, it does not matter. If they do act as described in the strategy, then their utterances will be correct whenever they are non-trivial.

How can Ann and Bob learn such co-ordinated strategies? Perhaps they both start with a naive strategy where each says "I don't know" except when (s)he sees a 1, and says "My number is 2," if a 1 is seen. But once each is acting that way, it is possible for one of them to proceed to a more sophisticated situation. If the other person has a 2 on his forehead, but does not say, "My number is 2," then that signals that *my* number must be 3 since it cannot be 1. Thus group activities involving some sort of co-ordination may *evolve* without anyone planning that they evolve just this way.

6 Reasoning about Knowledge in the Context of Social Software

We end these notes by re-examining the original motivation for developing formal models of knowledge in multi-agent situations. Namely, how the formal analysis we developed in the previous sections fits into a larger theory of social software. We do this by highlighting three examples where knowledge-theoretic properties are important for the analysis of a social procedure.

6.1 Knowledge and Social Networks

The topic "who knew what and when" is not just of interest to epistemic logicians. Often it is the subject of political scandals (both real and imagined). For example, consider the much talked about Valerie Plame affair. A July 2003 column in the Washington Post reported that Plame was an undercover CIA operative. This column generated much controversy due to the fact that such information (the identity of CIA operatives) is restricted to the relevant government officials. Of course, in this situation, we know full well "Who knew what and when": in July of 2003, Robert Novak (the author of the article) knew that Plame was a CIA operative. What creates a scandal in this situation is *how* Novak came to know such information. Since the CIA goes to great lengths to ensure that communication about sensitive information is contained within its own organization, the only way Novak could have known that Plame was a CIA operative was if a communication channel had been created between Novak and someone inside the CIA organization.

To put this a bit more formally, given a set of agents $\mathcal{A}$, call any graph $\mathcal{G} = (\mathcal{A}, E)$ a **communication graph** where the intended interpretation of an edge between agent i and agent j is that i and j *can communicate*. In this setting, the CIA can be represented as a connected component of $\mathcal{G}$. Given that the CIA is the only group of agents that (initially) knows the identity of CIA operatives, and Novak is not an element of the CIA component of $\mathcal{G}$ then we can conclude that Novak did not originally know the identity of CIA operatives and no amount of communication *that respects the graph* $\mathcal{G}$ can create a situation in which Novak does know the identity of a CIA operative. Thus Novak's report in the Washington Post implied that our original communication graph was incorrect[17]. That is, there must be an edge (or a chain) between Novak and *some* agent inside the CIA component. Since in principle, Novak could be connected to any member of the CIA component, much resources and time has been spent discussing the possible edges.

A multi-agent epistemic logic with a communication modality where agents are assumed to communicate according to some fixed communication graph is developed in (Pacuit and Parikh, 2005, 2007). Agents are assumed to have some private information at the outset, but may refine their information by acquiring information possessed by other agents, possibly via yet other agents. That is, each agent is initially informed about the truth values of a finite set of propositional variables. Agents are assumed to be connected by a *communication graph*. In the communication graph, an edge from agent i to agent j means that agent i can directly receive information from agent j. Agent i can then refine its information by learning information that j has, including information acquired by j from another agent, k.

In keeping with the CIA-theme, we give an example from Pacuit and Parikh (2005) of the type of situations that we have in mind. Let $K_i\varphi$ mean that according to i's current information φ is true. Given a communication graph $\mathcal{G} = (\mathcal{A}, E)$, we say that a sequence of communications (i learns a fact from j who learns a fact from k, and so on) **respects the communication graph** if agents only communicate with their immediate neighbors in $\mathcal{G}$. Let $\Diamond\varphi$ mean that φ becomes true after a sequence of communications that respects the communication graph. Suppose now that φ is a formula representing the exact whereabouts of Bin Laden, and that Bob, the CIA operative in charge of maintaining this information knows φ. In particular, $K_{\text{Bob}}\varphi$, but suppose that at the moment, Bush does not know the exact whereabouts of Bin Laden ($\neg K_{\text{Bush}}\varphi$). Presumably Bush *can* find out the exact whereabouts of Bin Laden ($\Diamond K_{\text{Bush}}\varphi$) by going through Hayden, but

[17] Of course, it could also mean that we were incorrect about the agents' initial information — Novak could have had previous knowledge about the identity of CIA agents. In this paper, we are interested in studying communication and so will not consider this case.

of course, *we* cannot find out such information ($\neg\Diamond K_{\mathrm{E}}\varphi \wedge \neg\Diamond K_{\mathrm{R}}\varphi$) since we do not have the appropriate security clearance. Clearly, then, as a *pre-requisite* for Bush learning φ, Hayden will also have to come to know φ. We can represent this situation by the following formula:

$$\neg K_{\mathrm{Bush}}\varphi \wedge \Box(K_{\mathrm{Bush}}\varphi \rightarrow K_{\mathrm{Hayden}}\varphi)$$

where $\Box$ is the dual of diamond ($\Box\varphi$ is true if φ is true after every sequence of communications that respect the communication graph).

6.2 Manipulating a Fair Division Procedure

Suppose there are two players, called Ann (A) and Bob (B), and n (divisible[18]) goods ($G_1, \ldots, G_n$) which must be distributed to Ann and Bob. Steven Brams and Alan Taylor invented an algorithm called *Adjusted Winner* (*AW*) to "fairly" divide these goods between Ann and Bob (Brams and Taylor, 1996). We begin by discussing an example which illustrates the Adjusted Winner algorithm.

Suppose Ann and Bob are dividing four goods: G_1, G_2, G_3 and G_4. *Adjusted Winner* begins by giving each of Ann and Bob 100 points to divide among the four goods as they see fit. Say Ann assigns points 15, 46, 14, and 25 respectively to G_1 through G_4, and Bob assigns 7,36,12, and 45. We get the following table:

Item	Ann	Bob
G_1	$\underline{15}$	7
G_2	$\underline{46}$	36
G_3	$\underline{14}$	12
G_4	25	$\underline{45}$
Total	100	100

The first step of the procedure is to give each agent the goods for which it assigns more points. So, Ann receives the goods G_1, G_2 and G_3 while Bob receives G_4. However this is not an equitable outcome since Ann has received 75 points while Bob only received 45 points (each according to their personal valuation). We must now transfer some of Ann's goods to Bob. In order to determine which goods should be transfered from Ann to Bob, we look at the ratios of Ann's valuations to Bob's valuations. For G_1 the ratio is $15/7 \approx 2.14$, for G_2 the ratio is $46/36 \approx 1.28$, and for G_3 the ratio is $14/12 \approx 1.17$. Since 1.17 is the smallest

[18] Actually all we need to use is that *one particular* good is divisible. However, since we do not know before the algorithm begins *which* good will be divided, we assume all goods are divisible. See (Brams and Taylor, 1996) for a discussion of this fact.

ratio (i.e., the ratio closest to 1), we transfer as much of G_3 as needed from Ann to Bob[19] to achieve equitability.

However, even giving all of item G_3 to Bob will not create an equitable division since Ann still has 61 points, while Bob has only 57 points. In order to create equitability, we must now transfer part of item G_2 from Ann to Bob. Let p be the proportion of item G_2 that Ann will keep. Then p should satisfy

$$15 + 46p = 45 + 12 + 36(1 - p)$$

yielding $p = 78/82 \approx 0.9512$, so Ann will keep 95.12% of item G_2 and Bob will get 5.12% of item G_2. Thus both Ann and Bob receive 58.76 points. It turns out that this allocation (Ann receives all of G_1 and 95.12% of G_3 and Bob receives all of items G_4 and G_3, plus 5.12% of item G_2) is *envy-free*, *equitable* and *efficient*, or *Pareto optimal*. In fact, Brams and Taylor show that Adjusted Winner *always* produces such an allocation Brams and Taylor (1996).

It turns out that agents may improve their total allocation by *misrepresenting* their preferences. The following example from Brams and Taylor (1996) illustrates how Ann can deceive Bob. Suppose that Ann and Bob are dividing two paintings: one by Matisse and one by Picasso. Suppose that Ann and Bob's actual valuations are given by the following table.

Item	Ann	Bob
Matisse	75	25
Picasso	25	75

Ann will get the Matisse and Bob will get the Picasso and each gets 75 of his or her points.

But now suppose Ann knows Bob's preferences, but Bob does not know Ann's. Can Ann benefit from being insincere? Suppose that Ann announces the following allocation:

Item	Ann	Bob
Matisse	26	25
Picasso	74	75

So Ann will get the Matisse, receiving 26 of her announced (and insincere) points and Bob gets 75 of his announced points. But Ann will *also* get a time share in the Picasso! Let x be the fraction of the Picasso that Ann will get, then we want

$$26 + 74p = 75 - 75p$$

[19] When the ratio is closer to 1, a unit gain for Bob costs a smaller loss for Ann.

Solving for p gives us $p = 0.33$ and each gets 50 of his or her announced preference. In terms of Ann's *true* preference, however, the situation is very different. She is getting from her true preference $75 + 0.33 * 25 = 83.33$ (cf. Theorem ?? from Brams and Taylor (1996) for a proof that this is, in fact, the *best* Ann can do).

However, while honesty may not always be the best policy, it is the only *safe* one; i.e., it is the only one which will guarantee 50%. For suppose that Ann's actual valuation is $(a_1, ..., a_n)$ but she reports $(c_1, ..., c_n)$. We show how she can end up with less than 50%. Suppose that Bob also reported $(c_1, ..., c_n)$. We know that in that case both Ann and Bob would get *exactly* 50% of their declared valuations. So Ann would receive 50% according to her *declared* valuation and this might be different from her actual valuation. To see how it might be *less* consider the eventuality that Bob reports slightly more than c_i when $c_i < a_i$ and slightly less than c_i when $c_i > a_i$. In the initial allocation then Bob will get all the pieces where Ann's declared valuation is less than her actual valuation, and Ann will get those where it is more. There will be adjustments of course, but Ann will still tend to get pieces where her declared valuation is *more* than her actual valuation. If she gets (approximately) 50% by her declared valuation, then it will be *less than* 50% by her actual valuation. Thus she can lose out by being dishonest (unless of course she *knows* something about Bob's declared values).

Suppose *both* players know each other's preferences but neither knows that the other knows their own. Their announced point allocations might then be as follows:

Item	Ann	Bob
Matisse	26	74
Picasso	74	26

Each will get 74 of his or her announced points, but each one is really getting only 25 of his or her *true* points. The following theorem of Brams and Taylor describes the situation when agents divide two goods.

Suppose *both* players know each other's preferences. Moreover, Ann knows that Bob knows her preference and Bob doesn't know that Ann knows, then the announced allocation will be as follows:

Item	Ann	Bob
Matisse	73	74
Picasso	27	26

Now suppose they both know each other's preference and each know that the other person knows his or her preference. Then the announced valuations will be:

Item	Ann	Bob
Matisse	73	27
Picasso	27	73

What happens as the level of knowledge increases?

6.3 Strategic Voting

The following example is from (Brams and Fishburn, 1994). Suppose that there are four candidates $\mathcal{O} = \{o_1, o_2, o_3\}$ and nine voters divided into three groups: A, B and C. Suppose that the sizes of the groups are given as follows: $|A| = 4$, $|B| = 3$, and $|C| = 2$. We assume that all the voters in each group have the same true preference and that they all vote the same way. Suppose that the voting procedure is plurality voting. Assume the voters' true preferences are as follows:

$$\begin{aligned} P_A^* &= o_1 >_{P_A^*} o_3 >_{P_A^*} o_2 \\ P_B^* &= o_2 >_{P_B^*} o_3 >_{P_B^*} o_1 \\ P_C^* &= o_3 >_{P_C^*} o_1 >_{P_C^*} o_2 \end{aligned}$$

Since we assume that in the absence of additional information, the voters will vote sincerely[20], candidate o_1 will win an initial election with a total of 4 votes. Now, Brams and Fishburn make the following assumption about the effect of poll information on a candidates choice of vote: "After the poll, voters will adjust their voting strategies to differentiate between the top two candidates, as indicated by the poll, if they prefer one of these candidates to the other one of these choices. Given that they are not indifferent between the top two candidates in the poll, they will vote after the poll for the one of these two they prefer" Brams and Fishburn (1994). Following this protocol, only the voters in group C will change their votes. Given that they prefer o_1 to o_2, group C will give their votes to candidate o_1, thus strengthening the lead of o_1. However, note that it is candidate o_3 who is the *Condorcet candidate*, i.e., a candidate who defeats every other candidate in a pairwise contest.

Brams and Fishburn go on to generalize this example and show that if the voters follow the protocol described above, then under plurality voting, if the Condorcet candidate is not one of the top two candidates identified by the poll, then that Condorcet candidate will always lose. In the above example, the protocol is set up so that the second round of votes is a fixed point, i.e., the voters will not change their votes a second time. The next example (from Chopra et al. (2004)) describes a situation in which a fixed point does not occur until round IV:

Example: Suppose that there are four candidates $\mathcal{O} = \{o_1, o_2, o_3, o_4\}$ and five groups of voters: A, B, C, D and E. Suppose that the sizes of the groups are given

[20] See Parikh and Pacuit (2005) for a *proof* of the fact that voting honestly is the only protocol which dominates not voting under plurality voting.

as follows: $|A| = 40$, $|B| = 30$, $|C| = 15$, $|D| = 8$ and $|E| = 7$. We assume that all the voters in each group have the same true preference and that they all vote the same way. Suppose that the voting procedure is plurality voting. The voters' true preferences are as follows:

$$\begin{aligned}
P^*_A &= o_1 >_{P^*_A} o_4 >_{P^*_A} o_2 >_{P^*_A} o_3 \\
P^*_B &= o_2 >_{P^*_B} o_1 >_{P^*_B} o_3 >_{P^*_B} o_4 \\
P^*_C &= o_3 >_{P^*_C} o_2 >_{P^*_C} o_4 >_{P^*_C} o_1 \\
P^*_D &= o_4 >_{P^*_D} o_1 >_{P^*_D} o_2 >_{P^*_D} o_3 \\
P^*_E &= o_3 >_{P^*_E} o_1 >_{P^*_E} o_2 >_{P^*_E} o_4
\end{aligned}$$

We assume that the voters all use the following protocol. If the current winner is o, then voter i will switch its vote to some candidate o' provided:

1. i prefers o' to o, formally $o' >_{P_i} o$, and
2. the current total for o' plus voter i's group's votes for o' is greater than the current total for o.

By this protocol a voter (thinking only one step ahead) will only switch its vote to a candidate which is currently not the winner. Initially, we assume that the voters all report their (unique) sincere vote. The following table describes what happens if the voters use this protocol. The candidates in bold are the winner of the current election round.

Size	Group	I	II	III	IV
40	A	$\mathbf{o_1}$	o_1	o_4	$\mathbf{o_1}$
30	B	o_2	$\mathbf{o_2}$	$\mathbf{o_2}$	o_2
15	C	o_3	$\mathbf{o_2}$	$\mathbf{o_2}$	o_2
8	D	o_4	o_4	o_1	o_4
7	E	o_3	o_3	o_1	$\mathbf{o_1}$

In round I, everyone reports their top choice and o_1 is the winner. C likes o_2 better than o_1 and its own total plus B's votes for o_2 exceed the current votes for o_1. Hence by the protocol, C will change its vote to o_2. A will not change its vote in round II since its top choice is the winner. D and E also remain fixed since they do not have an alternative like o' required by the protocol. In round III, group A changes its vote to o_4 since it is preferred to the current winner (o_2) and its own votes plus D's current votes for o_4 exceed the current votes for o_2. B and C do not change their votes. For B's top choice o_2 is the current winner and as for C, they have no o' better than o_2 which satisfies condition 2). Ironically, Group D and E change their votes to o_1 since it is preferred to the current winner is o_2 and group A is currently voting for o_1. Finally, in round IV, group A notices that E is voting for o_1 which A prefers to o_4 and so changes its votes back to o_1. The situation stabilizes with o_1 which, as it happens, is also the Condorcet winner.

Finally, an example in which the strategizing does not stabilize.

Example: Consider three candidates $\{o_1, o_2, o_3\}$, and 100 voters. Suppose that there are three groups of voters A, B, and C. The sizes of the groups are $|A| = 40$, $|B| = 30$ and $|C| = 30$. The actual preferences are given as follows:

$$\begin{aligned} P^*_A &= o_1 >_{P^*_A} o_2 >_{P^*_A} o_3 \\ P^*_B &= o_2 >_{P^*_B} o_3 >_{P^*_B} o_1 \\ P^*_C &= o_3 >_{P^*_C} o_1 >_{P^*_C} o_2 \end{aligned}$$

Assume that the voters use the following protocol. A voter i will switch its vote for o to o' provided (assume w is the current winner)

1. o' is i's second choice and the current winner is i's last choice, or

2. o' is i's top choice and the current winner is i's top choice.

Assuming that the voting protocol is plurality voting and that all voters follow the above protocol generates the following table.

Size	Group	I	II	III	IV	V	VI	VII	VIII	IX	$\cdots$
40	A	$\mathbf{o_1}$	o_1	o_2	$\mathbf{o_2}$	$\mathbf{o_2}$	$\mathbf{o_1}$	$\mathbf{o_1}$	o_2	o_1	$\cdots$
30	B	o_2	$\mathbf{o_3}$	$\mathbf{o_3}$	$\mathbf{o_2}$	$\mathbf{o_2}$	o_2	o_3	$\mathbf{o_3}$	$\mathbf{o_3}$	$\cdots$
30	C	o_3	$\mathbf{o_3}$	$\mathbf{o_3}$	o_3	o_1	$\mathbf{o_1}$	$\mathbf{o_1}$	$\mathbf{o_3}$	$\mathbf{o_3}$	$\cdots$

After reporting their initial preferences, candidate o_1 will be the winner with 40 votes. The members of group B dislike o_1 the most, and will strategize in the next election by reporting o_3 as their preference. So, in the second round, o_3 will win. But now, members of group A will report o_2 as their preference, in an attempt to draw support away from their lowest ranked candidate. o_3 will still win the third election, but by changing their preferences (and making them public) group A sends a signal to group B that it should report its true preference - this will enable group A to have its second preferred candidate o_2 come out winner. This cycling will continue indefinitely; o_2 will win for two rounds, then o_1 for two rounds, then o_3 for two, etc.

7 Notes

Relevant Conferences: A number of regular conferences address issues surrounding formal epistemology:

- FEW: *Formal Epistemology Workshop* is a yearly conference aimed at general issues in formal epistemology:

  `ist-socrates.berkeley.edu/∼fitelson/few/`

- TARK: *Theoretical Aspects of Rationality and Knowledge* is a bi- annual conference on the interdisciplinary issues involving reasoning about rationality and knowledge:

  `www.tark.org`

- LOFT: *Logic and the Foundations of Game and Decision Theory* is a bi-annual conference which focuses, in part, on applications of formal epistemology in game and decision theory:

  `www.econ.ucdavis.edu/faculty/bonanno/loft.html`

- KR: *Conference on the Principles of Knowledge Representation and Reasoning* is a bi-annual conference geared towards computer scientists that emphasizes both theoretical and practical applications (although the focus in on knowledge representation as opposed to reasoning about knowledge).

  `www.kr.org`

- See also the website

  `www.cs.gc.cuny.edu/∼kgb`

References

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics 4*, 1236 — 1239.

Aumann, R. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior 8*, 6 – 19.

Aumann, R. (1999a). Interactive epistemology I: Knowledge. *International Journal of Game Theory 28*, 263–300.

Aumann, R. (1999b). Interactive epistemology II: Probability. *International Journal of Game Theory 28*, 301 – 314.

Aumann, R. and A. Heifetz (2001, June). Incomplete information. Social Science Working Paper 1124, California Institute of Technology.

Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory 37*, 167 – 190.

Bacharach, M. O. L. (2006). *Beyond Individual Choice.* Princeton University Press.

Baltag, A. and L. Moss (2004). Logics for epistemic programs. *Synthese: Knowledge, Rationality, and Action 2*, 165 – 224.

Blackburn, P., M. de Rijke, and Y. Venema (2002). *Modal Logic.* Campbridge University Press.

Blackburn, P., J. van Benthem, and F. Wolter (Eds.) (2006). *Handbook of Modal Logic*, Volume 3. Elsevier.

Bonanno, G. and P. Battigalli (1999, June). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics 53*(2), 149–225.

Brams, S. J. and P. C. Fishburn (1994). Voting Procedures. In *Handbook of Social Choice and Welfare*. North-Holland.

Brams, S. J. and A. D. Taylor (1996). *Fair Division: From cake-cutting to dispute resolution.* Cambridge University Press, Cambridge.

Brandenburger, A. (2005). The power of paradox: Some recent developments in interactive epistemology. Unpublished Manuscript.

Cave, L. (1983). Learning to agree. *Economics Letters 12*, 147 – 152.

Chopra, S., E. Pacuit, and R. Parikh (2004). Knowledge-theoretic properties of strategic voting. In J. J. Alferes and J. Leite (Eds.), *Proceedings of Logics in Artificial Intelligence: 9th Euorpean Conference (JELIA).*

Chwe, M. S.-Y. (2001). *Rational Ritual.* Princeton University Press.

Clark, H. and C. R. Marshall (1981). Definite reference and mutual knowledge. In Joshi, Webber, and Sag (Eds.), *Elements of Discourse Understanding.* Cambridge University Press.

Dunbar, R. (2004). *The Human Story*, Chapter 3, “Mental Magic”. Faber and Faber.

Ely, J. and M. Peski (2006). Hierarchies of belief and interim rationalizability. *Theoretical Economics 1.*

Fagin, R., J. Geanakoplos, J. Halpern, and M. Vardi (1999). The hierarchical apporach to modeling knowledge and common knowledge. *International Journal of Game Theory 28*, 331–365.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi (1995). *Reasoning about Knowledge.* The MIT Press.

Geanakoplos, J. and H. Polemarchakis (1982). We can’t disagree forever. *Journal of Economic Theory 28*(1), 192 – 200.

Georgatos, K. (1994). Knowledge theoretic properties of topological spaces. In M. Masuch and L. Polos (Eds.), *Lecture Notes in Artificial Intelligence*, Volume 808, pp. 147–159. Springer-Verlag.

Georgatos, K. (1997). Knowledge on treelike spaces. *Studia Logica.*

Gerbrandy, J. (1999). *Bisimulations on Planet Kripke.* Ph. D. thesis, University of Amsterdam.

Gettier, E. (1963). Is justified true belief knowledge? *Synthese*, 121 – 123.

Goble, L. (Ed.) (2001). *The Blackwell Guide to Philosophical Logic.* Blackwell Publishers.

Gochet, P. and P. Gribomont (2006). Epistemic logic. In D. Gabbay and J. Woods (Eds.), *Logic and the Modalities in the Twentieth Century*, Volume 7 of *The Handbook of History and Philosophy of Logic.* Elsevier.

Goldblatt, R. (2006). Mathematical modal logic: A view of its evolution. In D. Gabbay and J. Woods (Eds.), *Logic and the Modalities in the Twentieth Century*, Volume 7 of *Handbook of the History and Philosophy of Logic.* Elsevier.

Halpern, J. (1999). Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence 26*, 1–27.

Halpern, J. (2001). Substantive rationality and bacward induction. *Games and Economic Behavior* (37), 425 – 435.

Halpern, J. (2003). *Reasoning about Uncertainty.* The MIT Press.

Halpern, J. and Y. Moses (1983). Knowledge and common knowledge in a distributed environment. *ACM-PODC*, 50 – 61.

Harsanyi, J. C. (1967). Games with incompletet information played by bayesian players parts I-III. *Management Sciences 14*.

Heifetz, A. (1996). Comment on consensus without common knowledge. *Journal of Economic Theory 70*, 273 — 277.

Heifetz, A. and D. Samet (1998). Knowledge spaces with arbitrary high rank. *Games and Economic Behavior 22*, 260–273.

Heifetz, A. and D. Samet (1999). Hierarchies of knowledge: an unbounded stairway. *Mathematical Social Sciences 38*, 157–170.

Hendricks, V. (Ed.) (2006). *8 Bridges between Formal and Mainstream Epistemology*, Volume 128. Spinger.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Ithaca: Cornell University Press.

Hodkinson, I. and M. Reynolds (forthcoming). Temporal logic. In *Handbook of Modal Logic.*

Kooi, B. (2003). *Knowledge, Chance and Change.* Ph. D. thesis, University of Groningen.

Krasucki, P. (1996). Protocols forcing consensus. *Journal of Economic Theory 70*, 266 – 272.

Lewis, D. (1969). *Convention.* Harvard University Press.

Littlewood, J. E. (1953). *A Mathematicians Miscellany.* Methuen and Company.

Mertens, J.-F. and S. Zamir (1985). Formulation of bayesian analysis for games with incomplete infomration. *International Journal of Game Theory 14*, 1–29.

Meyer, J.-J. (2001). *The Blackwell Guide to Philosophical Logic*, Chapter Epistemic Logic. Blackwell Publishers.

Meyer, J.-J. and F. Veltman (2006). Intelligent agents and common-sense reasoning. In P. Blackburn, J. van Benthem, and F. Wolter (Eds.), *Handbook of Modal Logic.* Elsevier.

Meyers, J.-J. and W. van der Hoek (1995). *Epistemic Logic for AI and Computer Science*. Number 41 in Cambridge Tracks in Theoretical Computer Science. Cambridge University Press.

Moss, L. and R. Parikh (1992). Topological reasoning and the logic of knowledge. In Y. Moses (Ed.), *Proceedings of TARK IV*. Morgan Kaufmann.

Moss, L. and I. Viglizzo (2005). Harsanyi type spaces and final coalgebras constructed from satisfied theories. In *Electronic Notes in Theoretical Computer Science*, Volume 106.

Osborne, M. and A. Rubinstein (1994). *A Course in Game Theory*. The MIT Press.

Pacuit, E. (2007). Some comments on history based structures. *Journal of Applied Logic*.

Pacuit, E. and R. Parikh (2005). The logic of communication graphs. In J. Leita, A. Omicini, P. Torroni, and P. Yolum (Eds.), *Proceedings of DALT 2004*, Number 3476 in Lecture Notes in AI, pp. 256 – 269. Springer.

Pacuit, E. and R. Parikh (2006). Social interaction, knowledge and social software. In D. Goldin, S. Smolka, and P. Wegner (Eds.), *Interactive Computation: The New Paradigm*. Springer.

Pacuit, E. and R. Parikh (Forthcoming, 2007). Reasoning about communication graphs. In J. van Benthem, D. Gabbay, and B. Löwe (Eds.), *Interactive Logic, Proceedings of the 7th Augustus de Morgan Workshop*. King's College Press.

Pacuit, E., R. Parikh, and E. Cogan (2006, March). The logic of knowledge based obligation. *Knowledge, Rationality and Action: A Subjournal of Synthese 149*(2).

Parikh, R. (1991). Finite and infinite dialogues. In Y. Moschovakis (Ed.), *Proceedings of a Workshop on Logic and Computer Science*, pp. 481–498. Springer.

Parikh, R. (2002, September). Social software. *Synthese 132*, 187–211.

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

Parikh, R. (2005). WHAT do we know, and what do WE know? In R. van der Meyden (Ed.), *Proceedings of TARK 2005*. National University of Singapore.

Parikh, R. (2007a). Is there a logic of society? Forthcoming in *Proceedings of First Indian Conference on Logic and its Applications*.

Parikh, R. (2007b). Knowledge and structure in social algorithms (extended abstract). Presented at the Stony Brook Conference on Game Theory.

Parikh, R. (2007c). Sentences, propositions and logical omniscience. *The Review of Symbolic Logic (forthcoming)*.

Parikh, R. and P. Krasucki (1990). Communication, consensus, and knowledge. *Journal of Economic Theory 52*(1), 178 – 189.

Parikh, R. and P. Krasucki (1992, March). Levels of knowledge in distributed systems. *Sadhana — Proc. Ind. Acad. Sci. 17*, 167–191.

Parikh, R., L. Moss, and C. Steinsvold (2006). Topology and epistemic logic. In M. Aiello, I. Pratt-Hartmann, and J. van Benthem (Eds.), *Handbook of Spatial Reasoning*. Springer.

Parikh, R. and E. Pacuit (2005). Safe votes, sincere votes, and strategizing. Working Paper.

Parikh, R. and R. Ramanujam (1985). Distributed processes and the logic of knowledge. In *Logic of Programs*, Volume 193 of *Lecture Notes in Computer Science*, pp. 256 – 268. Springer.

Parikh, R. and R. Ramanujam (2003). A knowledge based semantics of messages. *Journal of Logic, Language and Information 12*, 453 – 467.

Plato. Meno. Available online at http://classics.mit.edu/Plato/meno.html, , translation by B. Jovett.

Plaza, J. (1989). Logics of public communications. In *Proceedings, 4th International Symposium on Methodologies for Intelligent Systems*.

Ramsey, F. (1931). Truth and probability. In *The Foundations of Mathematics*. Routledge and Kegan Paul.

Saari, D. (2001). *Decisions and Elections*. Cambridge University Press.

Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision 37*(42).

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy 12*, 133 – 163.

van Benthem, J. (2002). 'One is a lonely number': on the logic of communication. Technical Report Technical Report PP-2002-27, ILLC, University of Amsterdam. Available at http://staff.science.uva.nl/∼johan/Muenster.pdf.

van Benthem, J. and F. Liu (2004). Diversity of logical agents in games. *Philosophia Scientiae 8*(2), 163 – 178.

van der Hoek, W. and M. Pauly (2006). Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter (Eds.), *Handbook of Modal Logic*. Elsevier.

van Ditmarsch, H. (2000). *Knowledge Games*. Ph. D. thesis, University of Groningen.

van Ditmarsch, H., W. van der Hoek, and B. Kooi (2007). *Dynamic Epistemic Logic*. Springer.

van Emde Boas, P., J. Groendijk, and M. Stockhof (1984). The conway paradox: its solution in an epistemic framework. In J. Groenendijk, T. Janssen, and M. Stokhof (Eds.), *Truth, Interpretation and Information, selected papers from the third Amsterdam Colloquium*. Foris Publications.

Zanaboni, A. M. (1991, June). Reasoning about knowledge: Notes of rohit parikh's lectures. Published in Italy: Cassa di Risparmio di Padova e Rovigo. Based on lectures given at the 3rd International School for Computer Science Researchers, Acireale, June 1991.