ERIC PACUIT, ROHIT PARIKH and EVA COGAN

THE LOGIC OF KNOWLEDGE BASED OBLIGATION*

ABSTRACT. *Deontic Logic* goes back to Ernst Mally's 1926 work, *Grundgesetze des Sollens: Elemente der Logik des Willens* [Mally. E.: 1926, Grundgesetze des Sollens: Elemente der Logik des Willens, Leuschner & Lubensky, Graz], where he presented axioms for the notion 'p ought to be the case'. Some difficulties were found in Mally's axioms, and the field has much developed. *Logic of Knowledge* goes back to Hintikka's work *Knowledge and Belief* [Hintikka, J.: 1962, Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell University Press] in which he proposed formal logics of knowledge and belief. This field has also developed quite a great deal and is now the subject of the *TARK* conferences. However, there has been relatively little work combining the two notions of knowledge (belief) with the notion of obligation. (See, however, [Lomuscio, A. and Sergot, M.: 2003, Studia Logica **75** 63–92; Moore, R. C.: 1990, In J. F. Allen, J. Hendler and A. Tate (eds.), *Readings in Planning*, Morgan Kaufmann Publishers, San Mateo, CA]) In *this paper* we point out that an agent's obligations are often dependent on what the agent knows, and indeed one cannot reasonably be expected to respond to a problem if one is not aware of its existence. For instance, a doctor cannot be expected to treat a patient unless she is aware of the fact that he is sick, and this creates a secondary obligation on the patient or someone else to inform the doctor of his situation. In other words, many obligations are situation dependent, and only apply in the presence of the relevant information. Thus a case for *combining* Deontic Logic with the Logic of Knowledge is clear. We introduce the notion of *knowledge based obligation* and offer an S5, history based Kripke semantics to express this notion, as this semantics enables us to represent how information is transmitted among agents and how knowledge changes over time as a result of communications. We consider both the case of an absolute obligation (although dependent on information) as well as the (defeasible) notion of an obligation which may be over-ridden by more relevant information. For instance a physician who is about to inject a patient with drug d may find out that the patient is allergic to d and that she should use d' instead. Dealing with the second kind of case requires a resort to non-monotonic reasoning and the notion of *justified belief* which is stronger than plain belief, but weaker than absolute knowledge in that it can be over-ridden. This notion of justified belief also creates a derived notion of default obligation where an agent has, as far as the agent knows, an obligation to do some action a. A dramatic application of this notion is our analysis of the Kitty Genovese case where, in 1964, a young woman was stabbed to death while 38 neighbours watched from their windows but did nothing. The reason was not indifference, but none of the neighbours had even a default obligation to act, even though, as a group, they did have an obligation to take some action to protect Kitty.

Synthese (2006) 149: 311–341
Knowledge, Rationality & Action 57–87
DOI 10.1007/s11229-005-3877-6

1. INTRODUCTION

Suppose we are given two functions α and β over some domain D. Let $\alpha \leqslant \beta$ iff $\forall x \in D, \alpha(x) \leqslant \beta(x)$, and moreover $\alpha < \beta$ iff $\alpha \leqslant \beta$ and $\beta \not\leqslant \alpha$. Suppose now that $\alpha(d)$ $(\beta(d))$ is the utility value of strategy α (β) in some circumstances d. If $\alpha < \beta$, then we will say that strategy β *dominates* strategy α. Hence, if some element d of D is chosen, and we are offered a choice between $\alpha(d)$ and $\beta(d)$ in dollars, we will choose $\beta(d)$ even if d is unknown to us. This paradigm comes in useful in two contexts: the decision theoretic context, where D is the set of possible states of nature and α, β represent payoff functions; and the game theoretic context, where D represents the (already chosen but unknown to us) choices of the other players, and α, β are possible strategies for us.

Now this comparison between α and β will not be possible for us if all we are given are the *ranges* of α and β. For instance if $\alpha(x) = x^2$ and $\beta(x) = x$ over the unit interval [0,1], then it is indeed the case that $\alpha < \beta$ even though the ranges of the two functions are the same. Moreover, the function $\gamma(x) = 1 - x$ has the same range as β, but while we do have $\alpha < \beta$ we do not have $\alpha < \gamma$. So the ranges by themselves give us too little information to be able to tell whether $\alpha < \beta$.

For consider the decision whether to exercise. Suppose some people are rich and some are poor, but all would be better off exercising. However, assume for a moment that it is better to be rich and lazy than to be poor and to exercise. Then the consequences of exercising are {rich∧exercised, poor∧exercised} whereas the consequences of being lazy are {rich∧lazy, poor∧lazy}. Not *all* consequences of exercising are better than every consequence of being lazy, even though *each* individual person, whether rich or poor, is better off exercising. To ask that *all* consequences of exercising be better than every consequence of being lazy, is too much. So we need to compare situations pairwise, a particular situation with exercising and the "same" situation with laziness. In other words, if choosing between an α and a β, we should choose β if for the *specific circumstance* we are concerned with, β yields a higher value than α. Choosing intelligently, or responsibly, may require some *knowledge* about the circumstances.

These considerations have relevance to the situation where the values represent some societal good and we ought to do what is best for society. For knowing what is good may involve knowing some facts.

The following examples illustrate the type of situations we have in mind.

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.

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.

EXAMPLE 4. Uma has a patient with a certain condition C who is in the St. Gibson's hospital mentioned above. There are two drugs d and d' which can be used for C, but d has a better track record. Uma is about to inject the patient with d, but unknown to Uma, the patient is allergic to d, and d' should be used instead for this particular patient. Nurse Rebecca is aware of the patient's allergy and also that Uma is about to administer d. It is then Rebecca's obligation to inform Uma and to suggest that drug d' be used in this case.

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.

Our purpose in this paper is to set forth a framework which can be used to study situations similar to those in the four examples above and to point out certain logical properties which will hold. We take as our starting point the history based models of Parikh and Ramanujam (1985, 2003), see also (Fagin et al. 1995; Halpern

et al. 2004). These models allow a representation of how agents acquire information, both from observations and from other agents. Our goal is a semantics and an axiomatic system in which we can formalize the agents' reasoning in the above examples. In particular, we should be able to *formally argue* that Ann is obliged to send a message to Uma in Example 2 (given the appropriate assumptions). In fact, this has been one of the goals of standard *deontic logic*. See (Hilpinen 2001; Horty 2001) and references therein for an up to date discussion of deontic logic.

In much of the deontic logic literature, an agent's knowledge is only informally represented or the discussion is focused on representing epistemic obligations, i.e., what an agent 'ought to know' (see (Lomuscio and Sergot 2003) for a recent discussion). The logic in this paper is intended to capture the *dependency* of individual obligation on knowledge.

The above discussion and examples point to four issues that are relevant to the task at hand.

1. The formal language and semantics must have machinery in which we can express statements of the form "after agent i performs action a".
2. The formal language and semantics must have machinery in which we can express statements of the form "agent i is obliged to perform action a" or constructs of the form "after performing the obligatory action $a \ldots$".
3. Certain actions are obligatory *only* in the presence of relevant information.
4. Certain obligations may disappear in the presence of relevant information (in Example 4, Uma's obligation to administer drug d disappears in the presence of relevant information).

Each of the above issues (except perhaps the third) have been the focus of much discussion in a variety of contexts. Certainly the notion of obligation has been widely discussed among philosophers, logicians, and more recently, computer scientists. See (Hilpinen 2001) for a survey of the literature. A complete survey of the literature relevant to each of the four issues above would require a book length treatment and would distract us from the task at hand. Instead, we point to a few references which are relevant to our formal treatment. For a treatment of obligatory actions in social situations, the reader is referred to Horty (2001). In Horty (2001), using

so-called '*see to it that*' modal operators, Horty shows how to represent obligatory actions.

The next sections will discuss each of the four issues in detail. Specifically, Section 2 discusses the history based models of Parikh and Ramanajum (1985, 2003) models in detail. Sections 3, 4 and 5 discuss actions, obligations, and default obligations respectively. We then return to our examples by showing how each example can be modelled in our framework. Finally, we conclude and discuss some directions for further research.

2. AN ABSTRACT MODEL

Our main tool will be the distinction between global histories and local histories as in (Parikh and Ramanujam 1985, 2003; Fagin et al. 1995). The *global histories* include all (relevant) events which have taken place. An agent i's *local history* is those events which i has actually seen. Here we make the assumption that if we knew every event that has taken place we would know all facts, but our ignorance of some facts is due to the fact that some events have not been observed by us. Thus for instance if Uma does not know that Sam is ill, it is because she has not seen him throwing up. The events which she *has* seen, including perhaps the sight of Sam mowing his lawn are quite compatible with another state of affairs where he is in fact quite fine.

We shall use letters H, H' etc. to range over global histories and h, h' etc. to range over local ones. To express the notion of a *moment*, we will assume a global clock. This will allow us to translate sentences like, "At 10 AM, Uma is unaware that Sam is ill, but at 11 AM she knows." Being able to mention the time t (e.g. 10 AM) allows us to talk simultaneously about a moment for Uma and the *corresponding* moment for Sam. Letters t, t' will range over time, and given a moment t of time the global history H restricts to H_t, the global history *upto* (and including) time t.

Following (Parikh and Ramanujam 1985, 2003) we now present an abstract extensional representation of a communication system in which the system is described as a set of *global histories*, each of which represents one possible system evolution given by a sequence of global events. For each system, the set of *agents* that participate in its events is assumed to be a fixed finite set. Similarly, for each system, the set of possible global events is fixed.

For convenience, we fix $n > 0$, and consider only systems with agents from $\mathcal{A} = \{1, 2, \ldots, n\}$, and events from a fixed set E. E will be finite in the examples we give but the infinite case can be handled too. E^* is the set of all finite sequences over E and E^ω is the set of all infinite sequences over E; we will let $H, H', \ldots$ range over the set $E^* \cup E^\omega$. Let $H \preceq H'$ denote that H is a finite prefix of H'. We write $H_1; H_2$ or just $H_1 H_2$ to denote the concatenation of the finite history H_1 with the possibly infinite history H_2. When H is infinite or of length $\geqslant t$, we let H_t denote the finite prefix of H consisting of the first t elements. For a set of histories $\mathcal{H}$, let $\texttt{FinPre}(\mathcal{H})$ denote the set $\{H' \mid H' \preceq H \text{ for some } H \in \mathcal{H}\}$ containing all finite prefixes of sequences in $\mathcal{H}$. The set $\mathcal{H}$ is called a *protocol*. Of course, we assume that protocols are closed under finite prefixes.

The set of events E typically consists of actions by agents in the system (including the sending and receipt of messages), but may also include other events (perhaps due to actions of the environment) that affect the knowledge of agents. We do not have a specific syntax of messages here, but choose to identify the message with the event that denotes its sending or receipt; in this sense, when we talk of the meaning of a message, we are referring to what the sending (receiving) of that message (at a specific time, in a context) signifies to the sender (receiver). Thus we are really discussing the semantics of event occurrences as perceived by agents in the system.

DEFINITION 2.1. A *history based frame* is a tuple $\langle \mathcal{H}, E_1, \ldots, E_n \rangle$, where $\mathcal{H} \subseteq E^\omega \cup E^*$ (our *protocol*) is the set of all possible global histories, and for $i \in \mathcal{A}$, $E_i \subseteq E$ is the set of local events of agent i (not necessarily disjoint from E_j for $j \neq i$).

The role of the protocol $\mathcal{H}$ is to limit the possible global histories which any agent may consider. It is this limitation on what can happen globally that permits an agent to make inferences from locally observed events to non-observed events. Thus for instance, when Sam throws up or vomits, that event v is not witnessed by Uma, but the event m, which Uma *does* observe, of Ann saying "My dad is throwing up," creates in Uma the knowledge of the event v which she did not observe, for every global history H which includes an event like m also includes a previous event like v. If the protocol 'allowed' Ann to lie, i.e., if it contained histories where an event m was not preceded by an event v, then clearly Uma could not infer v from m.

Local histories are got by 'projecting' global histories to local components. For $i \in \mathcal{A}$, let $\lambda_i : \texttt{FinPre}(\mathcal{H}) \to E_i^*$ be the *local view function* for i. In this paper, we assume that the local view functions are defined as follows. Let $H \in \mathcal{H}$ be a finite history, then $\lambda_i(H)$ is obtained by mapping each event in E_i into itself, and each event from $E - E_i$ into a non-informative clock tick c. That is, the local history of agent i corresponding to global history H at time t consists simply of those events from H_t which are *seen* by agent i. Thus if $H_1 \preceq H_2 \preceq H \in \mathcal{H}$, then $\lambda_i(H_1) \preceq \lambda_i(H_2)$ as well. In particular, if h is the local history of agent i at some stage, and event e visible to i takes place next (that is, $e \in E_i$), then $h; e$ will be the resulting local history. If e is not visible, then the new local history would be hc where c is a clock tick. Note that we are also assuming that both $H_t, \lambda_i(H_t)$ will have the same length t.

In general we can define λ_i to be *any* function from finite strings of events to the set of i's local histories. The above conditions amount to assuming that the agents all have *perfect recall* and the system is synchronous (the agents all have access to the global clock). These assumptions are not necessary for our analysis, but are made to ease exposition. Note also that the domain of the local view functions are the *finite* strings of $\mathcal{H}$. This is in line with the assumption that at any moment only a finite number of events have already taken place. This assumption can be dropped and the definitions can be modified to allow agents the ability to remember an infinite number of events, but since our intended application is the analysis of social interactive situations and these situations typically have a starting point, we will stay with this more realistic assumption.

DEFINITION 2.2. Let $H, H' \in \mathcal{H}$ be global histories (of length $\geqslant t$). For $i \in \mathcal{A}$, define $H_t \sim_i H'_t$ iff $\lambda_i(H_t) = \lambda_i(H'_t)$.

It is easy to see that $\sim_i$ is an equivalence relation. We can consider this relation as giving the information partition for i in a history based frame; that is, given the information available to i, the histories H and H' cannot be distinguished. Agent i can only know properties *common* to H, H'.

Since the basic elements of the model are sequences, a linear time temporal logic suggests itself. Let $At = \{p_0, p_1, \ldots, p_m\}$ be a finite set of atomic propositions. Formally, the syntax of $\mathcal{L}_n^{KT}$ is given by

$$\phi, \psi \in \mathcal{L} ::= p \in At \mid \neg\phi \mid \phi \vee \psi \mid \phi U \psi \mid \bigcirc \phi \mid K_i \phi$$

Here $\bigcirc$ stands for "in the next moment", U for "until", and K_i for "i knows that". Given a history based frame $\mathcal{F} = \langle \mathcal{H}, E_1, \ldots, E_n \rangle$, a *model* is a pair $\mathcal{M} = \langle \mathcal{F}, V \rangle$, where $V : \texttt{FinPre}(\mathcal{H}) \to 2^{At}$ is a valuation map on finite prefixes of global histories which assigns truth values to the atomic predicates. We can now inductively define the notion $H, t \models \phi$, for infinite histories $H \in \mathcal{H}$:

1. $H, t \models p$ iff $p \in V(H_t)$, for $p \in At$.
2. $H, t \models \neg\phi$ iff $H, t \not\models \phi$.
3. $H, t \models \phi \vee \psi$ iff $H, t \models \phi$ or $H, t \models \psi$.
4. $H, t \models \bigcirc\phi$ iff $H, t+1 \models \phi$.
5. $H, t \models \phi U \psi$ iff for some $m > t$, $H, m \models \psi$ and for all k, $t < k < m$, $H, k \models \phi$.
6. $H, t \models K_i\phi$ iff for all $H' \in \mathcal{H}$ such that $H_t \sim_i H'_t$, $H', t \models \phi$.

The operators F (sometimes in the future), G (always in the future), can all be defined from U (as can $\bigcirc$). Their semantics is given by

1. $H, t \models F\phi$ iff for some $m > t$, $H, m \models \phi$.
2. $H, t \models G\phi$ iff for all $m > t$, $H, m \models \phi$.

Indeed, $F\phi$ is $true\, U \phi$, $\bigcirc\phi$ is $false\, U \phi$ and $G\phi$ is $\neg F \neg \phi$.

Notice that we are only interpreting formulas at *infinite* global histories. This is because the definition of truth of $\bigcirc\phi$ may not make sense if the global history is finite. That is if $\mathsf{len}(H) = k$, then how should we interpret $H, k \models \bigcirc\phi$? It is easy to see that specifying that $\bigcirc\phi$ is always true (or always false) conflicts with the valid principle $\bigcirc\neg\phi \leftrightarrow \neg\bigcirc\phi$.

Since the truth value of a formula of the form $K_i\phi$ at H, t depends only on $h = \lambda_i(H_t)$, we shall occasionally abuse language and write $h \models K_i(\phi)$ when we mean $H, t \models K_i(\phi)$. The operators $\bigcirc, F, G$ are more likely to arise in our own examples. However, U *could itself* arise in other examples which we intend to discuss in future.

We shall *extend our language* in later sections in order to express notions, like *values* and *good actions*, which are relevant to our examples.

The formula ϕ is said to be *satisfiable* if there exists a model $\mathcal{M}$, a global history $H \in \mathcal{H}$ in $\mathcal{M}$ and $t \geqslant 0$ such that $H, t \models \phi$. The formula ϕ is said to be *valid* iff $\neg\phi$ is not satisfiable. The following laws of the logic **S5** are easily seen to be valid:

– $K_i(\phi \to \psi) \to (K_i\phi \to K_i\psi)$.
– $K_i\phi \to \phi$.

– $K_i\phi \rightarrow K_i K_i \phi$.
– $\neg K_i\phi \rightarrow K_i \neg K_i \phi$.

A sound and complete axiomatization for knowledge and time under various assumptions can be found in (Halpern et al. 2004), using a slightly different framework. The precise connection between the two frameworks will be discussed below. The reader is referred to (Halpern et al. 2004) for the relevant results. See also (Pacuit and Parikh 2004) for a logic of learning from other agents.

There is a related approach to defining semantics for an epistemic temporal logic called *interpreted systems* (see (Fagin et al. 1995) Chapter 5 for an explanation). It turns out that these approaches are modally equivalent, i.e., both semantics validate the same epistemic temporal formulas (see (Pacuit 2005) for a discussion). In particular, this implies that the soundness and completeness proofs from (Halpern et al. 2004) can be applied here. So, is the difference between the two semantics only linguistic? Technically, perhaps the answer is yes. However, there is a difference from the modeler's point of view. The intuition guiding interpreted systems is that there is a computational procedure that each agent is following and the local states describe the internal states of the agents at different moments in time. So the difference lies in the intended application in the models. For interpreted systems, the intended application is an analysis of distributed computational procedures whereas for history based structures the intended application is social interactive situations. For example, in (Parikh and Ramanujam 2003), Parikh and Ramanajam argue that this framework very naturally formalizes many social situations by providing a semantics of messages in which notions such as Gricean implicature can be represented.

3. ACTIONS

We think of an action as something which is performed at a *finite* global history H and which yields a set $a(H)$ of global extensions of H (provided that the action a *can* be performed at H). In general there will be *other* extensions of H in which a has not been performed. Formally, we assume a finite set, Act, of actions that is a subset of E (the set of possible events). We assume that each action is tagged to a particular agent who is the only one performing that action. Thus if l stands for *turning on the light*, then l_u will be *Uma turning on the light*, and l_s will be *Sam turning on the light*. Clearly

Uma cannot perform the action l_s. For the *sake of simplicity* we assume that at any moment of time *only* one agent can perform any action, although if that agent does nothing, then nature is free to perform a clock tick.

Formally, we assume that the set of actions $\mathsf{Act} \subseteq E$ is partitioned into sets Act_i for each agent $i \in \mathcal{A}$. That is, $\mathsf{Act} = \cup_{i \in \mathsf{A}} \mathsf{Act}_i$ where $\mathsf{Act}_i \cap \mathsf{Act}_j = \emptyset$ for $i \neq j$. Elements of Act_i will be denoted by a_i, b_i, etc. If it is clear from the context which agent can perform which action, then we will leave out the indices.

We understand an action $a \in \mathsf{Act}$ as a partial function from the set of finite histories to sets of global histories. If Ann turns on the light at H_t then the corresponding set is the set of all histories H' such that H' extends $H_t l_a$. Formally, given an infinite global history H and a time $t \in \mathsf{N}$:

$$a(H_t) = \{H' \mid H_t a \preceq H' \text{ and } H' \in \mathcal{H}\}$$

This implies that when an action is performed, it is performed at the next moment of time. We could weaken this assumption and assume that performing an action means performing that action eventually. In this case, $a(H_t)$ will be the set of global histories H' such that there is an $H_1 \in E^*$ and $H_t H_1 a \preceq H'$. Note that in this case for two different actions a and b which both can be performed at finite global history H_t, $a(H_t)$ and $b(H_t)$ need not be disjoint. We will use either definition depending on the application – it should be clear from the context which is intended.

In order to reason about actions in our formal language, we introduce a **PDL** style modal operator. If $a \in Act$, then $[a]\phi$ is intended to mean that in all histories in which a is performed (by the appropriate agent), ϕ is true. I.e., all executions of a make ϕ true. Its dual $\langle a \rangle \phi$ will mean that after some execution of a, ϕ is true. Given a global history H and time t, we define truth of $[a]\phi$ as follows

$$H, t \models [a]\phi \text{ iff for all } H' \in a(H_t),\ H', t+1 \models \phi$$

whereas the $\bigcirc$, F and U modal operators are linear time operators, i.e., they range over moments on a single global history, the dynamic modalities just defined are branching time operators.

Note that we are assuming that actions are primitive, i.e., an action is just an element of the set of events E. One could develop a calculus of actions, where complicated actions are built up from

primitive actions using standard **PDL** style operators. We refer the reader to van der Meyden (1996) for more on this topic. However, a large class of examples, including all the ones we consider here can be handled without adding the complications of a calculus of actions; and so we leave this line of reasoning for future research.

One last assumption is that each agent knows *when* it can perform an action. Thus if $H_t \sim_i H'_t$ and i can perform a_i at H_t then it can also perform a_i at H'_t. We note that if the power has been off and an agent does not know whether the electric power is back on, then the agent still knows it can perform the action 'flip the light switch', but does not know whether it can perform the action 'turn on the light'. Since we stipulate that an agent knows when it can perform an action, our notion of action will correspond to flipping the switch but not to turning on the light (unless there is no doubt that the power *is* on.) It is not too hard to see that this assumption will force the following axiom scheme to be valid:

$$\langle a_i \rangle \top \rightarrow K_i \langle a_i \rangle \top$$

4. VALUES

We move to the second issue discussed in the introduction: formalizing an agent's obligation. The basic idea is to assign a real number to each infinite global history (called the *value* of the history) and assume that higher valued histories are "better" than lower valued histories. Notice that we are not making any attempt to explain *why* histories are assigned the values they are – that is a job for an ethicist (or perhaps a court). We are interested in formalizing the agents' reasoning about obligatory actions *given an assignment of values*. Furthermore, it is worth pointing out that the actual values assigned to histories do not matter – it is only the induced ordering among global histories which will be of interest for us. At this stage, the use of real numbers eases presentation and suggests parallels with a game theoretic analysis. The basic idea is that our models can be thought of as extensive games in which all agents are playing the *same* utility function, or at least each agent's utility function induces the same order over global histories.

Under natural assumptions, (e.g. that the set of values is finite or compact) there will be a set of extensions of finite histories H which have the highest possible value. For a finite history H, we will refer to this set as the *H-good* histories and denote it as $\mathcal{G}(H)$.

Now, since all global histories have a value, so will those global histories which extend some finite history H in which a has been performed. We will say that *a is good* to be performed at a finite history H, if $\mathcal{G}(H) \subseteq a(H)$, i.e., there are no H-good histories which do not involve the performing of a. And we say that *a may* be performed at H if $\mathcal{G}(H) \cap a(H)$ is non-empty. Note that this definition seems compatible with the inference that if a letter may be posted then it may be posted or burned. But we can avoid this apparent paradox by saying that the permission to post or burn a letter really amounts to a permission to post the letter plus the permission to burn it. This can be formally expressed as, $(\mathcal{G}(H) \cap a(H) \neq \emptyset)$ and $(\mathcal{G}(H) \cap b(H) \neq \emptyset)$ rather than the more obvious interpretation $(\mathcal{G}(H) \cap (a(H) \cup b(H)) \neq \emptyset)$ which does justify burning the letter as an option. Here, of course, a is the action of posting the letter and b is the action of burning it. The formula $(\mathcal{G}(H) \cap a(H) \neq \emptyset)$ expresses permission to post the letter. It does imply $(\mathcal{G}(H) \cap (a(H) \cup b(H)) \neq \emptyset)$ but, in our view, the latter formula does not express the intent of the English sentence "You may post the letter or burn it."

We now make the above discussion more formal, but first some notation. Let $\mathcal{H}$ be a protocol and $H \in \mathcal{H}$ an infinite global history. Define for each $t \in \mathsf{N}$, $\mathcal{F}(H_t) = \{H' \in \mathcal{H} \mid H_t \preceq H'\}$. That is, $\mathcal{F}(H_t)$ is the "fan" of global histories (in $\mathcal{H}$) that contain H_t as an initial segment. Let $\mathcal{K}$ be any set of histories, $f : \mathcal{K} \to \mathsf{R}$ be any function, and define $f[\mathcal{K}] = \{f(H) \mid H \in \mathcal{K}\}$. Given a protocol $\mathcal{H}$, let $\mathsf{Inf}(\mathcal{H})$ be the set of infinite histories of $\mathcal{H}$.

DEFINITION 4.1. Let $\mathcal{H}$ be any protocol. A function $\mathsf{val} : \mathsf{Inf}(\mathcal{H}) \to \mathsf{R}$ is called a *value function* if for each infinite global history $H \in \mathcal{H}$,

1. For all $t \in \mathsf{N}$, $\mathsf{val}[\mathcal{F}(H_t)]$ is a closed and bounded subset of R.
2. $\cap_{t \in \mathsf{N}} \mathsf{val}[\mathcal{F}(H_t)] = \{\mathsf{val}(H)\}$

Condition 2 is a 'discounting' condition which ensures that values of histories depend only on what happens in a finite amount of time. If two histories agree for a long time then their values should be close. Formally, it is easy to see that condition 2 implies the following fact:

$$\forall \epsilon > 0,\ \exists t \geqslant 0,\ \forall H' \in \mathcal{H},\ (H'_t = H_t \Rightarrow |\mathsf{val}(H'_t) - \mathsf{val}(H_t)| < \epsilon)$$

Since $\mathsf{val}[\mathcal{F}(H_t)]$ is closed and bounded for all t, there are maximal and minimal elements. Thus we define,

$$\mathcal{G}(H_t) = \{H' \mid H' \in \text{argmax}(\mathsf{val}[\mathcal{F}(H_t)])\}$$

Thus $\mathcal{G}(H_t)$ is the set of maximally good, (or just maximal) extensions of H_t. Put another way, $\mathcal{G}(H_t)$ the set of extensions of H_t that maximize the val function.

In order to reason about good actions using our language, we must extend our formal language. For each action $a \in \mathsf{Act}$, introduce a formal symbol $G(a)$. The intended interpretation of $G(a)$ is "action a is 'good'". Truth is defined as follows: $H, t \models G(a)$ iff $\mathcal{G}(H_t) \subseteq a(H_t)$. We will return to this issue in Section 6.

We can now define knowledge based obligation.

DEFINITION 4.2. Agent i is *obliged* to perform action a at global history H and time t iff a is an action which i (only) can perform, and i *knows* that it is good to perform a. Formally, $(\forall H')(H_t \sim_i H'_t$ and $H' \in \mathcal{G}(H'_t) \Rightarrow H' \in a(H'_t))$. Putting this in terms of the agent's local history $h = \lambda_i(H_t)$, all maximal extensions of *any* H'_t with $\lambda_i(H'_t) = h$ belong to the range of the action a.

Note that in our semantics at any moment, only one action attached to a particular agent is good. In theory nothing prevents it from being good that Ann puts the tea-kettle on the stove while Uma is treating her father, but we prefer not to overburden an already heavy semantics.

4.1. *Comparison with Horty*

This above definition of a good action generalizes Horty's notion of dominance of actions (Horty 2001). In Horty (2001) actions are *defined to be* sets of global histories and at any moment m an agent i is faced with a set $Choice_i^m$ of possible actions. This set is a partition of the possible global histories that extend a global history at a particular moment m. Each history H is assumed to have a value $Value(H)$. Since actions are in fact sets of global histories, one is tempted to compare actions pointwise so that action a is 'better' than a' just in case $Value(H) \geqslant Value(H')$ for each $H \in a$ and $H' \in a'$. In such a case we will write $a \geqslant a'$ ($\leqslant$, $<$, $>$ can then be defined in similar ways). However, using the *sure-thing principle*[1] of Savage, Horty demonstrates some problems with this definition. In order to get around this complication, actions are given a functional flavor.

For each agent i and moment m let $State_i^m$ be the actions available to each agent other than i. Thus $State_i^m$ is a collection of actions available to agent i which are themselves sets of global histories. That is

$$State_i^m = Choice^m_{\mathsf{A}-\{i\}}$$

where $\mathcal{A}$ is the set of all agents.[2] Horty can now compare actions as follows (recall that actions are defined to be sets of global histories)

DEFINITION 4.3 (Horty 2001). Let i be an agent, m a moment and a and a' be two members of $Choice_i^m$. Then (a' *weakly dominates* a) $a \preceq a'$ if and only if $a \cap s \leqslant a' \cap s$ for each $s \in State_i^m$; and $a \prec a'$ if $a \preceq a'$ and not $a' \preceq a$.

Thus when comparing actions a and a', they are treated as functions over the domain of choices of the other agents (i.e., the domain is $State_i^m$). As functions, a and a' are then compared pointwise. Our approach is to make this idea explicit and define actions as partial functions on the set of all possible histories. We then can compare actions pointwise on their domains.

5. DEFAULT HISTORIES

As we have already seen from Example 4, the notion of a *default* history is important for our analysis. Since the notion of obligation in this chapter depends on the definition of knowledge, we must first weaken our definition of knowledge. We introduce a modal operator B_i which is intended to mean that "i is justified in believing ...". Our approach will be to define a system of Grove spheres on the set $\mathcal{H}$ (Grove 1988).

DEFINITION 5.1. Let $\mathcal{H}$ be a set of global histories. A *system of spheres* on $\mathcal{H}$ is a set $\mathsf{S} = \{\mathcal{H}_1, \mathcal{H}_2, \dots\}$ where for each $i \geqslant 1$, $\mathcal{H}_i \subseteq \mathcal{H}_{i+1} \subseteq \mathcal{H}$, and $\cup_{i=1}^{\infty} \mathcal{H}_i = \mathcal{H}$.

The intuition is that if $i < j$, then the histories in $\mathcal{H}_i$ are "more plausible" than those in $\mathcal{H}_j - \mathcal{H}_i$. There are two issues that will be important. The first is: *Given a finite global history H, which histories are the most plausible given that the situation has evolved according to H?* Denote this set of histories by $\mathcal{D}(H)$. Of course we want

$\mathcal{D}(H) \subseteq \mathcal{F}(H)$ (the set of global histories extending H). In order to formally define $\mathcal{D}$, we define an index function I for a system of spheres S. Given a finite global history H, $I(H) = \mu i.(\exists H', H'_t = H \text{ and } H' \in \mathcal{H}_i)$, i.e., $I(H)$ is the smallest index of a sphere containing an infinite extension of H. Then given a finite global history H,

$$\mathcal{D}(H) = \mathcal{H}_{I(H)} \cap \mathcal{F}(H)$$

That is, $\mathcal{D}(H)$ is the set of the most plausible histories extending H. The second notion is the set of global histories that a particular agent considers most plausible given the events he has seen. Formally, let $i \in \mathcal{A}$ be an agent and suppose that h is a local history for agent i. Then define the i-index function $I_i(h) = \mu j.(\exists H \in \mathcal{H}_j, \quad \lambda_i(H_t) = h)$, where $t = \mathsf{len}(h)$ (the length of h).

So, $I_i(h)$ is the least index of a sphere containing a history in $\lambda^{-1}(h)$. We can then define the set of histories that i considers plausible, given the events that i has seen. Denote this set $\mathcal{D}_i(h)$ and define it as follows

$$\mathcal{D}_i(h) = \{H' \mid \lambda_i(H'_t) = h\} \cap \mathcal{H}_{I_i(h)}$$

and here t is the length of all finite histories, local and global, mentioned.

We can now define the notion of *justified beliefs*. We say that agent i *justifiably believes* ϕ at H, t, denoted $B_i\phi$, if ϕ is true in all i-plausible histories. Formally,

$$H, t \models B_i\phi \text{ iff for all } H', H'_t \in \mathcal{D}_i(\lambda_i(H_t)), \quad H', t \models \phi$$

or putting it in terms of the local history h

$$h \models B_i\phi \text{ iff for all } H', H'_t \in \mathcal{D}_i(h), \quad H', t \models \phi$$

Of course $K_i\phi$ semantically entails $B_i\phi$. In general B_i does not satisfy the veridicality axiom (the truth of $B_i\phi$ at H, t does not necessarily imply that ϕ is true at H, t as H might not be in $D_i(\lambda_i(H_t))$). But it is easy to check that both positive and negative introspection hold. That is

LEMMA 5.2. B_i satisfies both positive and negative introspection. That is the following schemes are valid.

1. $B_i\phi \rightarrow B_i B_i\phi$
2. $\neg B_i\phi \rightarrow B_i \neg B_i\phi$

Proof. Suppose that $H, t \models B_i\phi$. Then for any H' with $H'_t \in \mathcal{D}_i(\lambda_i(H_t))$, $H', t \models \phi$. Let H'' and H''' be arbitrary histories such that $H''_t \in \mathcal{D}_i(\lambda_i(H_t))$ and $H'''_t \in \mathcal{D}_i(\lambda_i(H''_t))$. Since $H'' \in \mathcal{D}(\lambda(H_t))$, $\lambda(H''_t) = \lambda_i(H_t)$ and since $H'''_t \in \mathcal{D}(\lambda(H''_t))$, $\lambda_i(H'''_t) = \lambda_i(H''_t)$. Therefore, $\lambda_i(H'''_t) = \lambda_i(H_t)$ and hence since $I_i(\lambda(H_t)) = I_i(\lambda(H''_t))$, we have $\mathcal{H}_{I_i(\lambda(H_t))} = \mathcal{H}_{I_i(\lambda(H''_t))}$. Therefore, $\mathcal{D}_i(\lambda_i(H_t)) = \mathcal{D}_i(\lambda_i(H''_t))$. Hence, since $H'''_t \in \mathcal{D}_i(\lambda_i(H_t))$, we have $H''', t \models \phi$. Therefore, $H'', t \models B_i\phi$ and since H'' was arbitrary, $H, t \models B_i B_i\phi$. The proof of 2 is similar. □

Thus the logic of the operator B_i is $\mathbf{KD45}_n$ rather than $\mathbf{S5}_n$, but we do *act* as if it were $\mathbf{S5}_n$. We act on the advice of the short story writer Damon Runyon, "The race is not always to the swift, nor the battle to the strong, but that is the way to bet." In short, if a is the best action given ϕ, and $B_i(\phi)$ holds, then we do a.

5.1. *Default Obligations*

The obligation defined in Definition 4.2 is an *absolute obligation for agent i* in the sense that the obligation remains until a required action is performed by the agent. No amount of information, however surprising, can remove the obligation. But this is not the case for Uma's obligation in Example 4. Uma loses the obligation to administer drug d upon learning from nurse Rebecca that the patient is allergic to drug d. In this example, Uma not only gained the obligation to administer drug d' upon learning some surprising information, but also lost an obligation to administer d. Thus Uma's obligation to administer drug d was a *default obligation*, as an absolute obligation could not be lost.

The machinery we developed in this section can be used to formalize this notion. We say that an agent i has a default obligation to perform action a, provided all maximal histories that the agent considers plausible are ones in which a is performed. Formally

DEFINITION 5.3. Agent i has a *default obligation* to perform action a at global history H and time t iff a is an action which i (only) can perform, and i justifiably *believes* that it is good to perform a. Putting this in terms of the agent's local history $h = \lambda_i(H_t)$, all maximal extensions of any $H'_t \in \mathcal{D}_i(h)$ belong to the range of the action a.

Clearly, if agent i is obliged to perform action a, then agent i also has a default obligation to perform action a. There are three notions which are important for this chapter. Let H be a global history, $t \in \mathsf{N}$ and a an action.

1. a is a *good to be performed* at H, t iff every maximal extension of H_t is in the range of a, i.e., $\mathcal{G}(H_t) \subseteq a(H_t)$
2. a is a *knowledge based obligation* at H, t iff a satisfies Definition 4.2
3. a is a *knowledge based default obligation* at H, t iff a satisfies Definition 5.3.

If a is a good action, then a *ought* to be done, but the agent in question might not have any reason to believe that a ought to be done. This framework can now be used to understand the following quite well-known example.

The Kitty Genovese Murder

"Along a serene, tree-lined street in the Kew Gardens section of Queens, New York City, Catherine Genovese began the last walk of her life in the early morning hours of March 13, 1964.....As she locked her car door, she took notice of a figure in the darkness walking towards her. She became immediately concerned as soon as the stranger began to follow her.

'As she got of the car she saw me and ran,' the man told the court later, 'I ran after her and I had a knife in my hand.... I could run much faster than she could, and I jumped on her back and stabbed her several times,' the man later told the cops."

Many neighbours saw what was happening, but no one called the police. "Mr. Koshkin wanted to call the police but Mrs. Koshkin thought otherwise. 'I didn't let him,' she later said to the press, 'I told him there must have been 30 calls already.' "

"When the cops finished polling the immediate neighbourhood, they discovered at least 38 people who had heard or observed some part of the fatal assault on Kitty Genovese."[3]

Some 35 minutes passed between Kitty Genovese being attacked and someone calling the police. Why?

Analysis: The people who saw Kitty being killed did not have default knowledge that they had the obligation to help her. They all knew that the good histories were ones in which *someone* called the police, but not all these histories were ones where *they themselves* were the caller – someone else could be the caller. Compare this to a situation in a waiting room where a child's mother goes to the bathroom and her daughter starts to cry. Again there is no

one who has a default obligation to comfort the child, but typically, if there is a woman in that waiting room, she will *see* that no one else is taking care of the child and assume responsibility. Unlike the Geneovese case, there will be a common knowledge, *until the child is comforted*, that the child is not being comforted. Well designed social software, (a notion defined originally in (Parikh 2002)) will address such issues.

As we saw earlier there is not only knowledge but justifiable belief, and the justifiable belief of what are the best histories will depend on what one thinks the histories are. A man at the beach alone who sees a boy drowning will surely do something. There *may* be someone watching from a distance who might be a better swimmer than he himself is. But his default is that he is the only one who knows, is present, and therefore has the obligation to help. If on the other hand he is among 50 people at the beach, then he no longer has default knowledge of his obligation. There might well be other people on the beach who are better swimmers than he is, and perhaps among them are the boy's companions. Mrs. Koshkin's admonition to her husband amounted to her saying to him, "You do not have a default obligation."

6. PUTTING EVERYTHING TOGETHER

We have developed quite a bit of machinery in this paper, and so at this point it is worthwhile to summarize our discussion so far. We begin by extending the language $\mathcal{L}_n^{KT}$ to $\mathcal{L}_n^{KTO}$. Formulas in $\mathcal{L}_n^{KTO}$ have the following syntactic form:

$$\phi := p \mid \neg\phi \mid \phi \wedge \phi \mid \bigcirc\phi \mid \phi \; U \psi \mid K_i\phi \mid [a]\phi \mid G(a)$$

where $p \in \mathsf{At}$ and $a \in \mathsf{Act}$. We define the standard boolean operators, L_i and the temporal operators F and G as usual (see Chapter 2). Define $\langle a \rangle \phi$ to be $\neg[a]\neg\phi$. Let $\mathcal{L}_n^{BTO}$ be the language which is just like $\mathcal{L}_n^{KTO}$ but replace each K_i modality with B_i. We now give the intended interpretation of some of the formulas in $\mathcal{L}_n^{KTO}$ ($\mathcal{L}_n^{K^dTO}$).

- $G(a)$: "action a is good"', or "a is a non-informational obligation"
- $\langle a \rangle \top$: "action a can be performed"
- $K_i\langle a_i \rangle \top$: "agent i knows that she can perform action a_i"
- $K_i G(a_i)$: "agent i knows that action a_i is good", i.e., "i is obliged to perform a_i". Note that we will have $H, t \models K_i(G(a_i))$ just in

case a_i is a knowledge based obligation for agent i at H_t (Definition 4.2).
- $B_i\phi$: "agent i (justifiably) believes ϕ"
- $B_iG(a_i)$: "agent i has a default obligation to perform a_i". Note that we will have $H, t \models B_i(G(a_i))$ just in case a_i is a default obligation for agent i at H_t (Definition 5.3).

We will now repeat the four examples from the introduction and show how to formalize each example in the language $\mathcal{L}_n^{KTO}$ ($\mathcal{L}_n^{BTO}$). Let $\mathcal{A} = \{u, s, a, b\}$ be the set of agents (with the obvious interpretations) and suppose that $\mathsf{Act} = \{v, r, m\}$ are the set of actions (the interpretations will be given below).

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. Formally, $\neg K_uG(r)$, where r is the action of treating the neighbour (which only Uma can perform).

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. Formally, $K_u(G(r))$ is true. The interesting thing about this example is that this formula becomes true, because at the previous moment, $K_a(G(m))$ is true, where m stands for the action of telling Uma about Sam's illness (which only Ann can perform) *and* that Ann actually did send the message m. We discuss this example in more detail below, in particular we are interested in capturing Ann's reasoning which allows her to conclude that m is an obligatory action.

EXAMPLE 3. Mary is a patient in St. Gibson's hospital. Mary is having a heart attack. The caveat which applied in Example 1. does not apply here. The hospital has an obligation to *be aware* of Mary's condition at all times and to provide emergency treatment as appropriate. The issue here falls outside of the scope of our discussion thus far. What is important for this example is that the hospital has an obligation to ensure that procedures are setup to guarantee that at each moment $K_u(G(r))$ (here r means treat the next patient). What complicates matters from the hospital's point of view is that the hospital cannot necessarily assume that all agents are using the same value function. Hence, the task of the hospital is to set up

social procedures plus a system of rewards and punishments so that the agents behave as if they are using the same value function. We briefly touch on these issues in the conclusion.

EXAMPLE 4. Uma has a patient with a certain condition C who is in St. Gibson's hospital mentioned above. There are two drugs d and d' which can be used for C, but d has a better track record. Uma is about to inject the patient with d, but unknown to Uma, the patient is allergic to d and for this patient d' should be used. Nurse Rebecca is aware of the patient's allergy and also that Uma is about to administer d. It is then Rebecca's obligation to inform Uma and to suggest that drug d' be used in this case. Let δ stand for the action of giving drug d to the patient, similarly for δ' and d'. Formally, Uma has the default obligation to give the patient drug d ($B_u(G(\delta))$). However, since Rebecca (b) knows that Uma has this default obligation ($K_b B_u(G(\delta))$), Rebecca has an obligation to inform Uma about the drug ($K_b(G(m_d))$ where m_d means tell Uma about the allergy to drug d). Of course, we can replace each of Rebecca's knowledge operators with a justified belief operator.

Before turning to the semantics, we point out an issue relevant to our analysis. If $\langle a_i \rangle \top$ is true at some finite history H, then this represents that agent i *can* perform action a_i at history H. But this does not mean that agent i actually *does* perform actions a_i. In fact, our formal language does not have any machinery to express such a statement. Thus a question arises as to whether or not an agent will actually perform a given that the agent knows that it is good. This is important for Example 2 as we need not only that Ann knows that she should send a message to Uma, but also that Ann actually does send the message. This is relevant to our discussion because we are assuming that the agents share a utility function. Thus if an agent knows that a is good to perform, then the agent knows that it is in its own best interest to perform a. One is tempted to conclude that *of course* the agent will perform a in this case. Davidson considers these and related issues in (Davidson 1980). These issues are relevant to the question of which protocols are considered plausible, which is not central to the discussion at hand. We now turn to the semantics.

DEFINITION 6.1. Let $\mathcal{F}_K$ be a history based frame. A *knowledge based obligation model* based on $\mathcal{F}_K$ is a structure

$$\mathcal{M}_O = \langle \mathcal{H}, \{E_i\}_{i\in \mathsf{A}}, \{\lambda_i\}_{i\in \mathsf{A}}, \{\mathsf{Act}_i\}_{i\in \mathsf{A}}, \mathsf{val}, V\rangle$$

where

– $\mathcal{H}$ is a protocol closed under finite prefixes
– The sets of actions for the agents are pairwise disjoint and for each $i \in \mathcal{A}$, $\mathsf{Act}_i \subseteq E_i$
– For each $H \in \mathcal{H}$ and each $t \in \mathsf{N}$, there is a *unique* $i \in \mathcal{A}$ such that for each $H' \in \mathcal{F}(H_t)$, there is an action $a_i \in A_i$ such that $H_t a_i \preceq H'$.
– val is a value function (Definition 4.1)
– V is a valuation function

Truth in the model is defined as usual. We only give the definition of the new formulas:

– $H, t \models [a]\phi$ iff for all $H' \in a(H_t)$, $H', t+1 \models \phi$
– $H, t \models G(a)$ iff $\mathcal{G}(H_t) \subseteq a(H_t)$

Note that $G(a) \rightarrow \langle a\rangle \top$ will be valid in any knowledge based obligation model. This follows since the conditions on the val function implies that for any finite history H, $\mathcal{G}(H)$ is non-empty.

A *default knowledge based obligation model* extends a knowledge based obligation model with a system of spheres. That is, a default knowledge based obligation model is a structure

$$\mathcal{M}_{O^d} = \langle \mathcal{H}, \{E_i\}_{i\in \mathsf{A}}, \{\lambda_i\}_{i\in \mathsf{A}}, \{\mathsf{Act}_i\}_{i\in \mathsf{A}}, \mathsf{val}, \mathsf{S}_\mathsf{H}, V\rangle$$

where each component is as above and S_H is a system of spheres on $\mathcal{H}$ (see Definition 5.1).

7. FORMALIZING THE EXAMPLES

In this section, our goal is to show that the formal machinery we have developed in this chapter can be used to capture our intuitions about each of the examples from the introduction. We will only discuss Examples 1, 2 and 4. As stated in the previous section, Example 3 deals with different issues, and so we will not discuss it in this section. The conclusion discusses some issues relevant to Example 3. More specifically, our task is to construct a knowledge based obligation model in which the formulas from the previous section have their requisite truth values.

We begin by constructing a protocol $\mathcal{H}$. There are four events, v, m, r, c where v stands for Sam vomiting, m stands for Ann telling

Uma, r stands for Uma treating (or offering to treat) Sam and c is a clock tick which, unlike the other three, may occur more than once. Our global histories will consist of sequences in which events occur infinitely often, but v, m, t occur at most once. Moreover, since we assume Ann is truthful, m never occurs without v occurring first. Let $\mathcal{H}$ be the set of all such histories (closed under finite prefixes).

To be more precise, let $\mathcal{A} = \{u, s, a, b\}$ (with the obvious interpretation); and $\mathsf{Act}_u = \{r\}$, $\mathsf{Act}_a = \{m\}$, and $\mathsf{Act}_s = \{v\}$. Assume that the event v is observed by Sam and Ann, m by Ann and Uma, and r and c, let us say, by all three. That is, $E_u = \{r, m, c\}$, $E_a = \{r, v, m, c\}$, $E_s = \{r, v, c\}$. Then let $\mathcal{H}' \subseteq E^\omega = (E_u \cup E_a \cup E_s)^\omega$ and $\mathcal{H}$ be $\mathcal{H}'$ closed under finite prefixes. It is easy to see that, by construction, $\mathcal{H}$ satisfies the conditions from Definition 6.1.

The next set of assumptions concern the values of each global history. In those finite global histories in which v has occurred but not yet r, the best continuations are those in which r now occurs. And if v has not yet occurred then r (in the form of an offer to treat) may occur, but makes the history worse as the doctor is embarrassed by offering to treat a healthy man. Thus we stipulate that all histories in which neither v nor r occurs have value 2, those in which r occurs without v have value 1 as do those in which v is followed by r. Finally those histories in which v occurs but not r have value 0 as they are the worst. Let val be a value function that assigns the global histories these values and let $\mathcal{M}_0$ be the knowledge based obligation model we have just sketched (actually it is only a frame since we have not specified the truth values of the propositional variables).

It is convenient to introduce a propositional variable that can be used to describe properties of the histories (for example whether or not Sam is sick). Let sick be a propositional variable that is true at any finite history in which v has occurred without r. It is worth pointing out that sick is a description of events that have or have not taken place, not a description of how Sam feels. Otherwise, we would be assuming that Uma's treatment *always* cures Sam.

Suppose now that an agent's local history is h and that the agent acquires some knowledge. In that case, the set of global histories H such that $\lambda_i(H_t) = h$ will *decrease*, and universal quantified formulas over all such histories will be more likely to become true. Thus before Uma was told of Sam's illness, the set of global histories compatible with her own local one included many where Sam was not ill. Receiving the information, however, deletes them, and in all global histories still compatible with her knowledge, she must act to help

Sam. Similarly, in Example 2 Ann had an obligation to inform Uma, for before she tells Uma, in many of *Uma's* local histories compatible with Ann's, and in some global histories compatible with these latter, Ann's father is not ill and Uma cannot act. By informing Uma, Ann extends Uma's local history, and creates an obligation for Uma. Moreover, assuming that Ann knows that Uma does what she ought to, Ann herself has the obligation to inform Uma.

We first consider Examples 1 and 2 from Uma's point of view. In a history in which v has occurred but not m, from Uma's point of view there are global histories in which v has not occurred which are compatible with her own local history. So she cannot know that it is good to treat Sam, although it is. She is not yet obligated to treat Sam. Once m occurs, she knows that v must have occurred, it is good to treat, and she knows it. So she is obligated. More formally, we can show that $(K_u\mathsf{sick} \wedge \langle r\rangle\top) \rightarrow K_u G(r)$ is valid in $\mathcal{M}_O$. Furthermore, if we assume that Uma is "ethical" (i.e., her utility function matches the global value function), then we can conclude that if $K_u(G(r))$, then Uma will in fact choose to treat Sam. Finally, the obligation arises to treat Sam *only* because Uma knows that Sam is ill, i.e., $\neg K_u\mathsf{sick} \rightarrow \neg K_u(G(r))$. The following observation makes our claim more precise.

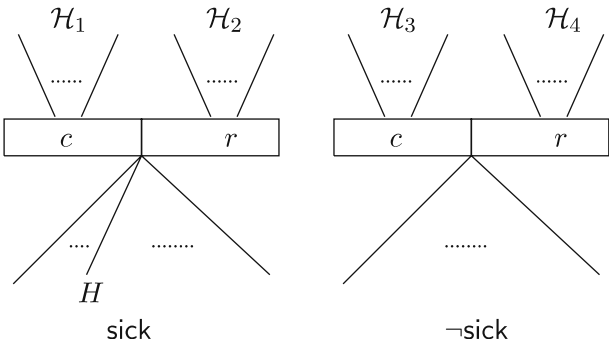
OBSERVATION 7.1. Let $\mathcal{M}_O$ be the knowledge based obligation model sketched above. Then the following formulas are valid in $\mathcal{M}_O$

1. $(K_u\mathsf{sick} \wedge \langle t\rangle\top) \rightarrow K_u G(r)$
2. $\neg K_u\mathsf{sick} \rightarrow \neg K_u(G(r))$

First consider formula 2. This represents the situation in Example 1. That is Uma does not know that Sam is ill, so she does not have the obligation to treat Sam. Let H be an arbitrary global history and $t \in \mathsf{N}$ an arbitrary moment such that $H, t \models K_u\mathsf{sick}$. Now, by the construction of $\mathcal{H}$, for any H' such that $H_t \sim_i H'_t$, m does not occur in H'_t. This follows since we assume Uma is aware of m and m only occurs in histories in which v has occurred. Furthermore, if r cannot be performed at H_t, then trivially $H, t \models \neg K_u(G(r))$ (since in this case $H, t \not\models G(r)$). Finally, it is not hard to see that in the construction of $\mathcal{H}$ we have assumed that Uma can *choose* whether or not to perform action r. As such we can assume the following. There are four subsets of global histories $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ and $\mathcal{H}_4$ such that

– $\mathcal{H}_1 = \{H' \mid H'_t c \preceq H' \text{ and } H_t \sim_u H'_t \text{ and } H', t \models \mathsf{sick}\}$
– $\mathcal{H}_2 = \{H' \mid H'_t t \preceq H' \text{ and } H_t \sim_u H'_t \text{ and } H', t \models \mathsf{sick}\}$
– $\mathcal{H}_3 = \{H' \mid H'_t c \preceq H' \text{ and } H_t \sim_u H'_t \text{ and } H', t \models \neg\mathsf{sick}\}$
– $\mathcal{H}_4 = \{H' \mid H'_t t \preceq H' \text{ and } H_t \sim_u H'_t \text{ and } H', t \models \neg\mathsf{sick}\}$

For simplicity, assume that $H \in \mathcal{H}_1$. This situation can be pictured as follows:



The above picture shows all the global histories that are equivalent from Uma's point of view at time t. These global histories can be grouped into two disjoint sets: the ones in which v has occurred and the ones in which v has not occurred. Each of the sets can be further divided into ones in which Uma performs action r and those in which Uma performs the (in)action c. Now, the definition of the value function implies that $\max(\mathsf{val}[\mathcal{H}_1]) = 1$, $\max(\mathsf{val}[\mathcal{H}_2]) = 2$, $\max(\mathsf{val}[\mathcal{H}_3]) = 2$ and $\max(\mathsf{val}[\mathcal{H}4]) = 1$. In other words, if the neighbour is sick then it is strictly better to treat the neighbour than to not treat the neighbour; however if the neighbour is not sick, then treating the neighbour for an illness he does not have is worse than not treating the neighbour. Let $H' \in \mathcal{H}_3$ be a history with maximal value (with respect to the histories in $\mathcal{H}_3$). Then since $H' \notin r(H'_t)$, we have $\mathcal{G}(H'_t) \not\subseteq r(H'_t)$ and so $H', t \not\models G(r)$. Therefore, since $H_t \sim_u H'_t$, $H, t \not\models K_u G(r)$. Thus Uma is not obliged to perform action r. Essentially, we are comparing the functions r and c on a domain D of histories compatible with Uma's local history. On this domain r and c are not comparable, neither dominates the other.

Returning to the formula 1, above, For the first formula, suppose that Ann informs Uma that her father is sick (as in Example 2). Actually all that is needed to be assumed is that Uma can rule out the right half of the above picture, i.e., all the histories in which v has not occurred (it does not matter *how* she came upon

this information). However, we will focus on Example 2. The message from Ann changes Uma's local view so that the sets of histories $\mathcal{H}_3$ and $\mathcal{H}_4$ are no longer possible for her. Uma's local view restricts the set of possible global histories to $\mathcal{H}_1$ and $\mathcal{H}_2$. And so, Uma *is* obliged to perform action a, since for any history on the new domain of histories compatible with Uma's updated local view, r is strictly better than the (in)action c. Formally, if we assume that event m has occurred, then Uma rules out all global histories in which v has not occurred. Notice that for Uma this effect is achieved by assuming that in $\mathcal{H}$ there are no global histories that contain m alone. This amounts to assuming that Ann is honest, i.e., if she sends a message about her father's illness it is only because v has occurred (and this is common knowledge). Thus if H is a global history in which m has occurred (at time $t-1$), then for all H' with $H_t \sim H'_t$, since v must have occurred in H', $H', t \models \mathsf{sick}$ and so $H, t \models K_u \mathsf{sick}$. Now, we have that for each H' such that $H'_t \sim_u H_t$, $\mathcal{G}(H_t) = \{H' \mid H'_t t \preceq H'\} = t(H_t)$ and so $H', t \models G(r)$. Hence $H, t \models K_u G(r)$. Thus Uma has the (knowledge based) obligation to treat Sam.

We now consider the situation from Ann's point of view. Suppose again that v has occurred but not m yet. Then according to Ann, Uma's local history is compatible with v not having occurred and in fact we will have that $K_a(\neg K_u(\mathsf{sick}))$ (Ann knows that Uma does not know about the vomiting). This formula will be true provided Ann *knows* Uma's local history. Of course it is unrealistic that Ann knows all of Uma's local events, but it is enough for Ann to know enough about Uma's histories so that Ann knows that probably Uma considers it possible that Sam has not vomited, i.e., $B_a(\neg K_u(\mathsf{sick}))$ (or perhaps $B_a(\neg B_u(\mathsf{sick}))$).

Since the vomiting *has* happened, all good histories now are those in which Sam has been treated, and those are included in the ones in which Ann has told Uma. So Ann ought to inform Uma about v, i.e., cause the event m. Formally, we have for any infinite global history H and time $t \in \mathsf{N}$, $H, t \models K_a \mathsf{sick} \wedge \langle m \rangle \top \rightarrow K_a(G(m))$. The proof of this fact is analogous to the argument concerning Uma. Let H be a fixed global history and $t \in \mathsf{N}$. The idea is that in our model, the maximal histories that extend H'_t, where $H_t \sim_a H'_t$ all contain the event m. In fact, more can be said about this situation. The analysis so far does not explain *why* Ann concludes that she should send the message m to Uma. We will discuss this in more detail in the next section, but for now we show

that the following formula is valid in our model: $[m]K_u G(r)$. Essentially, the reason is that we only consider histories such that if they contain m then they must contain v, i.e., Ann is truthful (and this fact is common knowledge). So if F is an arbitrary history and $t \in \mathsf{N}$, then for each global history $F' \in a(F_t)$, F' is a history in which both m and v have taken place. Then using the above argument, $F', t \models K_u G(r)$. Hence $F, t \models [m]K_u G(r)$. Since it is true for arbitrary global histories, then it will certainly be true at histories which are equivalent to H_t according to Ann. Hence, $H, t \models K_a[m]K_u G(r)$.

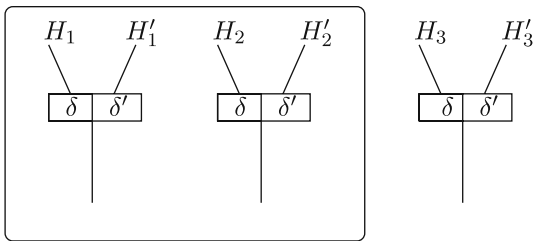
In a more complex scenario, with other agents, it could of course be that someone other than Ann had informed Uma of Sam's illness, but that Ann does not know this. We would say that Ann still has a default obligation to inform Uma, and this can easily be expressed in our language. Also note that in our scenario, once the obligation to treat arises, it remains until treatment has taken place.

The case of the nurse Rebecca is a bit more tricky. The reason is that acquiring knowledge may create an obligation as we saw before, but it cannot erase an absolute one. The existence of an obligation is a universally quantified formula whose truth value can only go from *false* to *true* as the domain shrinks. Thus if Uma had an absolute obligation to administer drug d before being informed by Rebecca of Mary's allergy, then she would still have it. How, then do we represent the fact that on learning of the allergy she *acquires* the obligation to administer d' but *loses* the obligation to administer d?

As discussed in Section 2.3, to deal with this case we will use the notion of a default history. Those histories in which patients do not have this allergy may be regarded as the usual kind and those in which they do are unusual. Typically, obligations are evaluated in terms of histories of the usual kind and when we say "good" history, we mean a good history of the usual kind. Learning about the allergy deletes these usual histories, and then the action contemplated is re-evaluated in terms of the unusual variety. Thus d is better than d' when we consider the usual sort of history, but the opposite happens when we consider the unusual variety.

The following picture illustrates the above discussion. Suppose that δ is the action 'give drug d to Mary' and δ' is the action 'give drug d' to Mary'. Suppose that according to Uma's information, each of the histories H_i is indistinguishable from H_j for $i, j = 1, \ldots, 3$ and similarly for the H'_i, H'_j. Also that $\mathsf{val}(H_i) > \mathsf{val}(H'_i)$ for $i = 1, 2$, but $\mathsf{val}(H'_3) > \mathsf{val}(H_3)$. In this case Uma is not absolutely obliged to per-

form δ since $\mathsf{val}(H_3') > \mathsf{val}(H_3)$. However, if the histories H_3 and H_3' are only *remotely* possible, then Uma has a default obligation to perform action δ, i.e., administer drug d. In the figure below, the histories inside the innermost rectangle are the "usual" histories. Once Rebecca informs Uma about Mary's allergy, the histories inside this rectangle are no longer possible; and so Uma is now obliged to perform action δ' and no longer obliged to perform δ.



7.1. *Common Knowledge of Ethicality*

Note that many of the arguments in the previous section boiled down to assumptions about which strings of events belong to the protocol under consideration. As such, the analysis may have appeared *ad hoc*. In this section we argue that the assumptions we made about the protocol in the previous section were not *ad hoc*, but rather follow from a general principle. We call this assumption *Common Knowledge of Ethicality*. Before discussing this principle, we go into some more details about Ann's reasoning.

At this point it is convenient to introduce some propositional variables which will make the discussion easier to follow. Recall that sick is a propositional variable which is true at all finite histories in which v has occurred. Similarly, define treat to be true at exactly those histories in which r has occurred and msg to be true at those histories in which m has occurred. We argued in the previous section that Ann has the (knowledge based) obligation to tell Uma about her father's illness. Clearly, Ann will not be under any obligation to tell Uma that her father is ill, if Ann justifiably believes that Uma would not treat her father even if she knew of his illness. Thus, to carry out a deduction we will need to assume

$$K_a(K_u\mathsf{sick} \leftrightarrow \bigcirc\mathsf{treat})$$

This says that Ann knows that Uma will treat (at next moment) iff she knows that Sam is ill. A similar assumption is needed to derive

that Uma has an obligation to treat Sam. Obviously, if Uma has a good reason to believe that Ann always lies about her father being ill, then she is under no obligation to treat Sam. In other words, the following formula must be true

$$K_u(\mathsf{msg} \leftrightarrow \mathsf{sick})$$

This formula says that Uma knows that a message is sent iff Ann's father is ill.

These formulas can all be derived from one common assumption which we call *Common Knowledge of Ethicality*. Analogous to the common knowledge of rationality in the game theory literature, this assumption assumes that all the agents are ethical, everyone knows that they are ethical, everyone knows that everyone knows that they are ethical, and so on. Here of course, "ethical" simply means that the agent's personal utility function matches the social value function.

Although this assumption of common knowledge of ethicality is needed in order to fully understand our examples, we do not need to include it explicitly as it is tacitly included in the set $\mathcal{H}$ which are considering. E.g. we simply *leave out* histories in which Uma knows about Sam's illness but fails to treat him. A more ambitious analysis would start with a *larger* $\mathcal{H}'$ and then use the common knowledge of ethicality to *cut down* to the sort of $\mathcal{H}$ we are using.

Notice that in the above discussion we assume that the agents are reasoning about the pre and post conditions of an action. The reason is that the formula $\langle r \rangle \top$ represents the statement "Uma *can* treat Sam". Thus $K_a(K_u\mathsf{sick} \leftrightarrow \langle r \rangle \top)$ represents the statement that "Ann knows that Uma knows Sam is sick if and only if she *can* treat Sam". However, an important part of Ann's reasoning is that Uma *will* treat Sam (provided she knows he is sick); and this is the intended meaning of the above formula.

8. CONCLUSIONS

A central issue when designing a social procedure is how to ensure that the agents will perform the required actions. One may suspect that the situation is trivial if we can assume that the agents all share the same utility function. The examples discussed in the introduction show that this is not the case. The information state of the agent is crucial when determining whether or not the agent

is obliged to perform an action. This paper provides a formal framework for reasoning about agents in social situations that are assumed to share a utility function. We start with the intuition that agents should not be faulted for not performing actions that they do not know about, and develop a formal language and semantics for reasoning about obligations, actions and knowledge in a multi-agent setting. The main contribution of this paper is conceptual, and indeed a number of technical questions remain. Nonetheless, we have showed that the formal machinery developed in this paper can be used to formalize the four illustrative examples from the introduction and so provides a powerful framework for reasoning about social software.

We first note that there are a number of issues related to Example 3 which we have not discussed. The most important is centered around the following question. Given a set of histories and values assigned to each history, we can ask, "Is it possible to program the agents in such a way that *if the agents do what the know they ought to do*, then one of the best histories is produced? We first must decide how much computational power we will ascribe to the agents. Assuming that agents have perfect recall requires that they have unbounded memory, and we will need to model them as Turing machines; however, we may want to assume that the agents only need to remember a bounded amount of information. In this case we will assume that the agents are finite automata. Essentially, the idea is to show that we can design finite automata that will generate a knowledge based obligation model which satisfies the appropriate knolwedge based obligation formulae. Thus, the problem can be reduced to model checking an appropriate interpretted system (see (Fagin et al. 1995; Lomuscio and Sergot 2003; Wozna et al. 2004)).

The main technical issue which remains is a sound and complete axiomatization. Of course, we should begin with the axiomatization from (Halpern et al. 2004) for the knowledge and temporal modalities and add the required axioms that correspond to the agent's reasoning capabilities (in this case perfect recall and common knowledge of the global clock). Finding the right axioms that connect our obligation formulas and our knowledge formulas requires making the common knowledge of ethicality more explicit. One obvious solution is to introduce a common knowledge operator into our language together with the standard axiomatization. However, as shown in (Halpern and Vardi 1989) this greatly increases the complexity of the validity problem and in some cases even makes

the validity problem Π^1_1 complete. In particular, if the agents are assumed to have perfect recall and have access to the global clock, then if we add a common knowledge operator to our language (with the standard interpretation), no recursive axiomatization is possible. Thus we need to find a way to bring in the assumption of common knowledge of ethicality without explicitly introducing a common knowledge operator. This will be left for further research.

ACKNOWLEDGEMENTS

All authors would like to thank members of the Knowledge, Games and Beliefs group of CUNY. Research for this work was supported in part by CUNY FRAP grants. Partial support for Eva Cogan was provided by PSC-CUNY Award #66171-00-35. The authors would also like to thank John Horty and two anonymous referees for comments.

NOTES

* Earlier versions of this paper were presented at the conferences *SEP-2004*, and *DALT-2004*.

[1] A very informal explanation of Savage's sure-thing principle says the following. If α is better than β provided P is true and α is better than β if P is false, then the agent may as well do α without bothering with the truth value of P. The reader is referred to (Horty 2001) for a more detailed discussion and the relevant references.

[2] We have only defined the set $Choice_i^m$ for one agent, so the above definition only makes sense if there are only two agents. However, this definition can be extended to multiple agents, see (Horty 2001) for more details.

[3] This quote is from the article 'A cry in the night: the Kitty Genovese murder', by a police detective, Mark Gado, and appears on the web in *Court TV's Crime Library*.

REFERENCES

Belnap, N., M. Perloff and M. Xu: 2001, *Facing the Future*, Oxford.

Davidson, D.: 1980, *Essays on Actions and Events*, Oxford University Press, New York.

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

Grove, A.: 1988, Two Modellings for Theory of Change, *Journal of Philosophical Logic* **17**, 157–179.

Halpern, J. and M. Vardi: 1989, The Complexity of Reasoning about Knowledge and Time, *Journal of Computer and System Sciences* **38**, 195–237.

Halpern, J., R. van der Meyden, and M. Vardi: 2004, Complete Axiomatizations for Reasoning about Knowledge and Time, *SIAM Journal of Computing* **33:2**, 674–703.
Hilpinen, R.: 2001, Deontic Logic, In Lou Goble (ed.), *Blackwell Guide to Philosophical Logic*, Blackwell, pp. 159–182.
Hintikka, J.: 1962, *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Cornell University Press.
Horty, J.: 2001, *Agency and Deontic Logic*, Oxford University Press.
Lomuscio, A. and M. Sergot: 2003, Deontic Interpreted Systems, *Studia Logica* **75** 63–92.
Mally, E.: 1926, *Grundgesetze des Sollens: Elemente der Logik des Willens*, Leuschner & Lubensky, Graz.
van der Meyden, R.: 1996, 'The Dynamic Logic of Permission', *Journal of Logic and Computation*, **6**(3), 465–479.
Moore, R.C.: 1990, 'A Formal Theory of Knowledge and Action'. In J. F. Allen, J. Hendler, and A. Tate (eds.) *Readings in Planning*, Morgan Kaufmann Publishers, San Mateo, CA., pp. 480–519.
Pacuit, E.: 2005, Topics in Social Software: Information in Strategic Situations, Doctoral Dissertation, City University of New York Graduate Center.
Pacuit, E. and R. Parikh: 2004, 'A Logic for Communication Graphs', to appear in the post-proceedings of DALT 2004.
Parikh, R.: 1995, Knowledge Based Computation (Extended Abstract), In *Proceedings of AMAST-95*, Montreal, July 1995, Edited by Alagar and Nivat, Lecture Notes in Computer Science no. 936, pp. 127–42.
Parikh, R.: 2002, 'Social Software', *Synthese* **132**, 187–211.
Parikh, R.: 2003, 'Levels of Knowledge, Games, and Group Action', In *Research in Economics* **57**, 267–281.
Parikh, R. and R. Ramanujam: 1985, 'Distributed Processes and the Logic of Knowledge', In *Logic of Programs*, LNCS #193, Springer pp. 256–268.
Parikh, R. and R. Ramanujam: 2003, 'A Knowledge based Semantics of Messages', In *Journal of Logic, Language and Information* **12**, 453–467.
Wozna, B., A. Lomuscio, and W. Penczek: 'Bounded Model Checking for Deontic Interpreted Systems'. *Proceedings of the second Workshop on Logic and Communication in Multi-Agent Systems (LCMAS04)*. Nancy, July, 2004.

Eric Pacuit
Department of Computer Science, The Graduate Center of CUNY
E-mail: epacuit@cs.gc.cuny.edu

Rohit Parikh
Departments of Computer Science, Mathematics and Philosophy, Brooklyn College and the Graduate Center of CUNY
E-mail: rparikh@gc.cuny.edu

Eva Cogan
Department of Computer Science, Brooklyn College
E-mail: cogan@sci.brooklyn.cuny.edu