

The logic of knowledge based obligation

Rohit Parikh¹
Eric Pacuit²
Eva Cogan³

¹ CS, Math and Philosophy
Brooklyn College** and The Graduate Center of CUNY
365 5th Avenue, New York City, NY 10016

rparikh@gc.cuny.edu
www.sci.brooklyn.cuny.edu/~rparikh

² Computer Science
The Graduate Center of CUNY,
epacuit@cs.gc.cuny.edu,

www.cs.gc.cuny.edu/~epacuit

³ Computer Science
Brooklyn College of CUNY
cogan@sci.brooklyn.cuny.edu
www.sci.brooklyn.cuny.edu/~cogan

Abstract. 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, unlike general *commandments*, many obligations are situation dependent, and are only relevant in the presence of the relevant information. This creates the notion of *knowledge based obligation*. We offer an S5, history based Kripke semantics to express this notion. We consider both the case of an absolute obligation (although dependent on information) as well as the notion of an obligation which may be over-ridden by more relevant information. For an example of the latter, 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 *not* use d ; she should use d' instead. Dealing with the second kind of case requires a resort to non-monotonic reasoning and the notion of *weak knowledge* which is stronger than plain belief, but weaker than absolute knowledge in that it can be over-ridden.

Clearly the issue of programming agents (human or other) to address this question of discharging obligations, or informing another agent of *its* obligation to perform some task will arise. A semantics based on "no forgetting" will require unbounded memory for the agents, but the

** 2900 Bedford Avenue, Brooklyn, NY 11210

examples we deal with can also be addressed by finite automata which treat their own local histories as strings of events.

1 Introduction

Suppose we are given two functions α and β over some domain D . Then $\alpha \leq \beta$ iff $\forall x \in D, \alpha(x) \leq \beta(x)$, and moreover $\alpha < \beta$ iff $\alpha \leq \beta$ and $\beta \not\leq \alpha$. If some element d of D is chosen, and we are offered a choice between $\alpha(d)$ and $\beta(d)$ in dollars, then knowing that $\alpha < \beta$, 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. The other context is the game theoretic one where D represents the (already chosen but unknown to us) choices of the other players, and α, β are possible strategies for us. In this context, if $\alpha < \beta$, we will say that β *dominates* α and we will tend to prefer β .

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 $\alpha < \beta$ but 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 $\alpha < \gamma$. Thus in situations where we do not have dominance, we need further information to make a decision. And sometimes that information is possessed by another agent.

Since we are concerned in this paper with obligations, we will interpret such obligations in terms of furthering some general good, and thus we will assume that all agents involved have the same preferences, albeit they may have different information, or different ability to act. This also means that we do not need to address the issue of some agents deliberately misinforming others, as that issue would arise only when the utilities or preferences clash.

Thus our present work has relevance to the situation where the values represent some (individual or) societal good and we ought to do what is best for society. Clearly, knowing the *set* of consequences of action α vs knowing the set of consequences of β will not always tell us how to decide. Rather we need to ask, *given* the current circumstances (possibly unknown or only partially known to us) can we still choose? It has been suggested that action β is preferable to action α if *all* consequences of β are better than any consequence of α . But clearly this requirement is too strict for our purposes.

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 $\{\text{rich} \wedge \text{exercised}, \text{poor} \wedge \text{exercised}\}$ whereas the consequences of being lazy are $\{\text{rich} \wedge \text{lazy}, \text{poor} \wedge \text{lazy}\}$. 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. If choosing between an α and a β , we should choose β if for *our specific circumstance*, β yields a higher value than α .

We illustrate this abstract framework so far with some examples.

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

b) Jill is a physician whose neighbour Sam is ill. The neighbour’s daughter Ann comes to Jill’s house and tells her. Now Jill does have an obligation to treat Sam, or perhaps call in an ambulance or a specialist.

c) Mary is a patient in St. Gibson’s hospital. Mary is having a heart attack. The caveat which applied in case a) 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.

d) Jill has a patient with a certain condition C who is in the St. Gibson hospital mentioned above. There are two drugs d and d' which can be used for C, but d has a better track record. Jill is about to inject the patient with d , but unknown to Jill, 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 Jill is about to administer d . It is then Rebecca’s obligation to inform Jill 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, Jill 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 to express the sorts of issues involved and to point out certain logical properties which will hold.

The Framework: our main tool will be the distinction between global histories and local histories as in [PR’85,HMV,PR’03]. 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 Jill 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' 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, Jill is unaware that Sam is ill, but at 11 AM she knows.” The time t (e.g. 10 AM) allows us to talk simultaneously about a moment for Jill 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* time t .

2 An abstract model

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 $[n] = \{1, 2, \dots, n\}$, and events from a fixed (possibly infinite⁴) set E . 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', \dots 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_1H_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 $\geq t$, we let H_t denote the finite prefix of H consisting of the first t elements. For a set of histories \mathcal{H} , let $\mathcal{P}(\mathcal{H})$ 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 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 21 *A system is a tuple $S = (\mathcal{H}, E_1, \dots, E_n)$, where $\mathcal{H} \subseteq E^\omega$ (our protocol) is the set of all (infinite) possible global histories of S , and for $i \in [n]$, $E_i \subseteq E$ 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 Jill, but the event m , which Jill *does* observe, of Ann saying “My dad is throwing up,” creates in Jill the knowledge of the event v which she did not observe, for every global history H in \mathcal{H} which includes an event like m also includes a previous event like v . If the protocol ‘allowed’ Ann to lie, then clearly Jill could not infer v from m .

⁴ Typically, when the set E is infinite then it has some structure. For instance E could be the set of strings on some finite alphabet. It is not intuitively plausible that an infinite set without any such structure could be part of a system of communication. This issue was addressed by Turing in his classic paper where he defined Turing machines.

Local histories are got by ‘projecting’ global histories to local components. For $i \in [n]$, let $\lambda_i : \mathcal{P}(E^\omega) \rightarrow E^*$ be the *projection map* for i , such that $\lambda_i(H)$ 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 . $\mathcal{H}_i \stackrel{\text{def}}{=} \{\lambda_i(H) \mid H \in \mathcal{P}(\mathcal{H})\}$ is the set of local histories of i .

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. Thus hc will be longer than h but will not have any additional non-trivial events.

Definition 22 Let $H, H' \in \mathcal{P}(\mathcal{H})$. For $i \in [n]$, define $H \sim_i H'$ iff $\lambda_i(H) = \lambda_i(H')$. For $H, H' \in \mathcal{H}$ let $H \sim_{i,t} H'$ iff $\lambda_i(H_t) = \lambda_i(H'_t)$.

$\sim_i, (\sim_{i,t})$ is an equivalence relation, and gives the *indistinguishability relation* for i (for i at time t). We can consider this relation as giving the information partition for i in the system S ; that is, given the information available to i , the histories H and H' cannot be distinguished; i can only know properties *common* to H, H' . Note that we are tacitly assuming a “no forgetting” condition, i.e. that agent i does not forget any of his local events. In practice we can often get away with agent i being a finite automaton with limited memory.

The properties of such systems can be studied in a logical language. Let L be a language which has formulae expressing (time dependent) properties of global histories. Then we can write $H, t \models \phi$, for ϕ belonging to L , to mean that the history H satisfies formula ϕ at time t . If the truth value of ϕ does not depend on t , then it is timeless. If ϕ has the property that once true it remains true, then it is *persistent*. We expand L to a larger language LK by closing under boolean connectives and operators K_i . Thus if ϕ is a formula of LK and i is an agent, then $K_i(\phi)$, meaning i knows ϕ , is also in LK . We can then define $H, t \models K_i(\phi)$ to hold if for all $H' \in \mathcal{H}$, if $H'_t \sim_i H_t$ then $H', t \models \phi$. What the agent i knows at time t depends on its local history. Moreover, the laws of logic $LK5$ (the $S5$ version of the logic of knowledge) are valid.

For definiteness, we fix a specific language \mathcal{L} so that the semantics of $H, t \models \phi$ is also fixed. Since the basic elements of the model are sequences, a linear time temporal logic suggests itself. Let $At = \{p_0, p_1, \dots\}$ be a finite set of atomic propositions. Formally, the syntax of \mathcal{L} is given by:

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

Here P stands for “in the past”, F for “in the future”, and K_i for “ i knows that”.

A **model** is a pair $M = (S, V)$, where $V : \mathcal{P}(\mathcal{H}) \rightarrow 2^P$ is a valuation map on finite prefixes of global histories which gives the truth values of some atomic predicates at the states. We can now inductively define the notion $H, t \models \phi$, for $H \in \mathcal{H}$, $t \geq 0$ and $\phi \in \mathcal{L}_0$:

1. $H, t \models p$ iff $p \in V(H_t)$, for $p \in P$.
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. $H, t \models F\phi$ iff for some $m > t$, $H, m \models \phi$.
5. $H, t \models P\phi$ iff for some $m < t$, $H, m \models \phi$.
6. $H, t \models K_i\phi$ iff for all $H' \in \mathcal{H}$ such that $H_t \sim_i H'_t$, $H', t \models \phi$.

For simplicity we do include the connective U , *until*, as none of our current examples need it. Of course there are other examples, like *keep up mouth to mouth resuscitation until the patient breathes on his own*, which do need this connective.

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 formula ϕ is said to be *satisfiable* if there exists a model M , a global history $H \in \mathcal{H}$ in M and $t \geq 0$ such that $M, t \models \phi$. ϕ is said to be *valid* iff $\neg\phi$ is not satisfiable. The following laws of the logic $LK5$ are easily seen to be valid:

- $K_i(\phi \supset \psi) \supset (K_i\phi \supset K_i\psi)$.
- $K_i\phi \supset \phi$.
- $K_i\phi \supset K_iK_i\phi$.
- $\neg K_i\phi \supset K_i\neg K_i\phi$.

There are of course other laws which connect K_i with the temporal connectives. However, we shall not attempt to give a complete axiomatization in this paper. See [HMY] for related results. See also [PaPa] for a logic of learning from other agents.

2.1 Actions and Values

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). Then an action $a \in Act$ can be understood as a partial function from the set of finite histories to global histories. Given a finite global history H ,

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

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, $a(H)$ will be the set of global histories H' such that there is an $H_1 \in E^*$ and $HH_1a \preceq H'$. However, for now, we will use the above simpler definition of action performance.

We assume that each agent knows *when* it can perform an action. Thus if $H \sim_{i,t} H'$ and i can perform a at H_t then it can also perform a at H'_t . Moreover,

for simplicity we assume that only one agent can perform some action at any moment. If no agents perform an action, then nature performs a ‘clock tick’.

We can introduce a *PDL* style operator into our language in order to represent executing an action. If $a \in Act$, then $[a]\phi$ is intended to mean that in all histories in which a is performed, ϕ is true. I.e., all executions of a makes 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 F and P modal operators are linear time operators, i.e., they range over moments on a single global history, the dynamic modalities just defined are best understood as branching time operators.

We now have enough machinery to formalize the notion of a knowledge based obligation. All global histories will be presumed to have a *value* and of course so will those global histories which extend H_t and in which a has been performed. Under natural assumptions, (e.g. that the set of values is finite or compact) there will be a set \mathcal{V} of extensions of H_t which have the highest possible value. We will refer to this set as the H_t -good histories and denote it as $\mathcal{V}(H_t)$. Since we are not dealing with lotteries, our notion of value is weak, and rather close to being a mere representation of preferences. But we *will* assume that if the same preferences are represented by two value functions V, V' , then each is an increasing, continuous function of the other.

We will say that a is *necessary* to be performed at H at time t , $\mathcal{G}(a, H_t)$, if $\mathcal{V}(H_t) \subseteq a(H_t)$, i.e., there are no H_t -good histories which do not involve the performing of a . And we say that a *may* be performed at H_t if $\mathcal{V}(H_t) \cap a(H_t)$ is non-empty.⁵

Let \mathcal{H} be a set of global histories and $H \in \mathcal{H}$ a global history. For each $t \in \mathbb{N}$, let $\mathcal{F}(H_t) = \{H' \in \mathcal{H} \mid H_t \preceq H'\}$. That is, $\mathcal{F}(H_t)$ is the ‘fan’ of global histories (in \mathcal{H}) that contain H_t as an initial segment. Recall that if \mathcal{F} is any set of histories, $val[\mathcal{F}] = \{val(H) \mid H \in \mathcal{F}\}$. We require for each global history $H \in \mathcal{H}$,

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

⁵ 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 the formula, $(\mathcal{V}(H_t) \cap a(H_t) \neq \emptyset) \wedge (\mathcal{V}(H_t) \cap b(H_t) \neq \emptyset)$ rather than the more obvious interpretation $(\mathcal{V}(H_t) \cap (a(H_t) \cup b(H_t)) \neq \emptyset)$ which does not 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 $(\mathcal{V}(H_t) \cap a(H_t) \neq \emptyset)$ expresses permission to post the letter. It does imply $(\mathcal{V}(H_t) \cap (a(H_t) \cup b(H_t)) \neq \emptyset)$ but, in our view, the latter formula does not express the intent of the English sentence ‘‘You may post the letter or burn it.’’

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.

Since $val[\mathcal{F}(H_t)]$ is closed and bounded for all t , there are maximal and minimal elements. Thus we define, $\mathcal{V}(H_t) = \{H' | H' \in \operatorname{argmax}(val[\mathcal{F}(H_t)])\}$. Thus $\mathcal{V}(H_t)$ is the set of maximally good, (or just maximal) extensions of H_t .

We can now define knowledge based obligation.

Definition 23 *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 necessary to perform a , i.e. $K_i(\mathcal{G}(a, H))$, or $(\forall H')(H \sim_{i,t} H' \wedge H' \in \mathcal{V}(H_t) \rightarrow H' \in a(H_t))$. I.e., 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 .*

We can formalize the above notion as follows. For each $a \in Act$, we define a primitive proposition $G(a)$. We say that $H, t \models G(a)$ iff all good global histories $H \in \mathcal{H}$ which extend H_t are such that $H_t a \preceq H$. Then we say that i is obliged to perform action a if $K_i(G(a))$.

2.2 Comparison with Horty

This above definition of a necessary action generalizes Horty’s dominance of actions ([Ho’01]). In [Ho’01] actions are 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 $Value(H) \geq Value(H')$ for each $H \in a$ and $H' \in a'$. In such a case we will write $a \geq a'$ ($\leq, <, >$ can then be defined in similar ways). However, using the *sure-thing principle* 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 . That is

$$State_i^m = Choice_{\mathcal{A}-\{i\}}^m$$

where \mathcal{A} is the set of all agents⁶. Horty can now compare actions as follows

Definition 24 (Horty [Ho’01]) *Let i be an agent and a and a' be two members of $Choice_i^m$. Then (a' weakly dominates a) $a \preceq a'$ if and only if $a \cap S \leq a' \cap S$ for each $S \in State_i^m$; and $a \prec a'$ if $a \preceq a'$ and not $a' \preceq a$.*

⁶ 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 [Ho’01] for more details.

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.

2.3 Applications

Suppose now that an agent acquires some knowledge. In that case, the set of global histories H such that $\lambda_i(H, t) = h$ will *decrease*, and the universal quantifier over all such histories will be more likely to become true. Thus before Jill 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 b) Ann had an obligation to inform Jill, for before she tells Jill, in many of *Jill's* local histories compatible with Ann's, and in some global histories compatible with these latter, Ann's father is not ill and Jill cannot act. By informing Jill, Ann extends Jill's local history, and creates an obligation for Jill. Moreover, assuming that Ann knows that Jill does what she ought to, Ann herself has the obligation to inform Jill.

To see this more precisely we consider global histories consisting of four events, v, m, t, c where v stands for Sam vomiting, m stands for Ann telling Jill, t stands for Jill treating (or offering to treat) Sam and c is a clock tick which, unlike the other three, may occur more than once. Thus our global histories will consist of sequences in which events occur infinitely often, but v, m, t occur at most once. Moreover, since Ann is truthful, m never occurs without v occurring first. In those finite global histories in which v has occurred but not yet t , the best continuations are those in which t now occurs. And if v has not yet occurred then t (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 t occurs have value 2, those in which t occurs without v have value 1 as do those in which v is followed by t . Finally those histories in which v occurs but not t have value 0 as they are the worst.

There are three agents, Sam, Ann, and the doctor, Jill. The event v is observed by Sam and Ann, m by Ann and Jill, and t , let us say, by all three. In a history in which v has occurred but not m , from Jill'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 necessary 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 necessary to treat, and she knows it. So she is obligated.

Suppose again that v has occurred but not m yet. Then from Ann's point of view, Jill's local history is compatible with v not having occurred and in fact we will have $K_a(\neg K_j(V))$ (Ann knows that Jill does not know about the vomiting) where V denotes that vomiting has occurred. 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 Jill. So Ann ought to inform Jill about v , i.e. cause the event m , and then hope for t to take place. Ann has the obligation to tell Jill.

In a more complex scenario, with other agents, it could of course be that someone other than Ann had informed Jill of Sam's illness, but that Ann does not know this. We would say that Ann still has an obligation to inform Jill, and this can easily be expressed in our language.

Note that in our scenario, once the obligation to treat arises, it remains until treatment has taken place.

Formal Example: We can formalize the above discussion as follows. Suppose that t is the action 'treat the neighbour', c is the action 'do not treat the neighbor', and **sick** is the sentence 'the neighbor is sick'. Suppose that there are four global histories H_1, H_2, H_3, H_4 . The situation described in example (a) can be represented as follows:

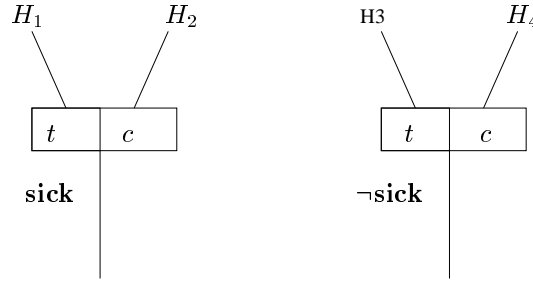


Figure 1

Jill cannot distinguish these four situations. Technically $H_1 \sim_{Jill} H_2 \sim_{Jill} H_3 \sim_{Jill} H_4$. Thus Jill does not know that her neighbor is sick ($\neg K_j(\mathbf{sick})$). Since $H_1 > H_2$, (i.e., $Value(H_1) > Value(H_2)$) $H_1 > H_3$, $H_4 > H_3$, if the neighbor is sick then it is strictly better to treat the neighbor than to not treat the neighbor; however if the neighbor is not sick, then treating the neighbor for an illness he does not have is worse than not treating the neighbor. Thus Jill is not obliged to perform action t , since given her local history, even though $H_1 > H_2$, $H_4 > H_3$. We are comparing the functions t and c on a domain D of histories compatible with Jill's local history. On this domain t and c are not comparable, neither dominates the other.

Now suppose that Ann informs Jill that her father is sick (as in example (b)). This event changes Jill's local view so that the H_3, H_4 are no longer possible for her. Jill's local view is now restricted to the left two histories (H_1 and H_2). And so, Jill *is* obliged to perform action a , since on the new domain D' of histories compatible with Jill's updated local view, t is strictly better than the action c .

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 (a persistent) 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 Jill had the 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 ?

Dealing with this case will require a resort to 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.

Thus we will assume that each history fragment H_t has a set $\mathcal{D}(H, t)$ of default extensions such that not all members of \mathcal{H} which extend H_t are in $\mathcal{D}(H, t)$. Now we can define the notion of an action which is necessary as a *default*, replacing \mathcal{H} by $\mathcal{D}(H, t)$ in our original definitions.

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 Jill's information, all of the histories H_i, H'_i for $i = 1, \dots, 4$ are indistinguishable; and that $H_i > H'_i$ for $i = 1, 2, 3$, but $H'_4 > H_4$. In this case Jill is not obliged to perform δ since $H'_4 > H_4$. However, if we assume that the histories H_4 and H'_4 are only *remotely* possible, then Jill is obliged to perform action δ , i.e., administer drug d . In the figure below, the histories inside the innermost rectangle are the "usual" histories. Once Rebeca informs Jill about Mary's allergy, the histories inside the rectangle are no longer possible; and so Jill is now obliged to perform action δ' and not obliged to perform δ .

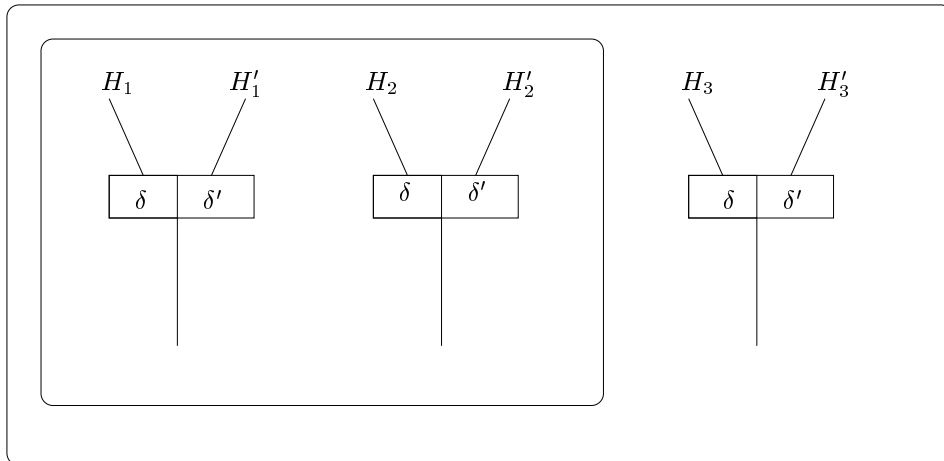


Figure 2

To deal with such cases we introduce the notion of *weak knowledge* or more prosaically, *justified belief*. For each H_t , divide its extensions into two (there could be more than two) parts, the normal extensions (of which there must be some) and the unusual extensions. Now we say that ϕ is justifiably believed by i at H, t iff for all normal extensions H' of some H'_t which are i, t -equivalent to H_t , $H', t \models \phi$. Justified belief no longer implies truth as H itself might not be one of these normal extensions. It is possible for Jill to justifiably believe that the patient does not have allergy although he does. Moreover, after nurse Rebecca learns of the patient's allergy, but before she tells Jill, the two have disjoint normal histories. Rebecca will now think in terms of 'typical patients with allergy', patients which, for Jill, are atypical. After Jill learns of the allergy, their views are again compatible.

Finally we come to case c) where we talk of the hospital's obligation to keep track of a patient's condition. Suppose that every heart attack, after a certain amount of time, results in death, unless treated, and such treatment can only follow an observation of the patient. Then it is clear that it is the obligation of the hospital to observe the patient periodically. We postpone details to the full version of the paper.

3 Programming the Agents

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 they know they ought to do*, then one of the best histories is produced?"

We first must decide on 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 whereas assuming that agents are finite automata means that agents have bounded memory.

In the following example the agents are finite automata.

Example: Consider the example where Ann is obliged to inform Jill about her father's vomiting which induces Jill to have the obligation to treat Sam (Ann's father). We assume $E = \{v, m, t, c\}$, where v stands for vomiting, m for Ann telling Jill about her father's illness, t for Jill treating Sam and c for a clock tick. Thus histories are strings over E . For the conditions placed on these strings, refer to Section 2.3.

Since in this example, Sam has no control over whether or not he vomits, we only consider Jill and Ann. We can ascribe the following finite automata to Jill and Ann. For Jill, suppose that the input alphabet is $\Sigma_J = \{m, t\}$, the states are $Q_J = \{j_0, j_1, j_2\}$ with j_0 being the start state. As for the transitions, we need to consider two types of transitions. The first is a transition induced by an action of another agent. For example, when Jill is in state j_0 , and she "sees" a m , she moves to state j_1 . Since m is not an action that Jill can perform, we think of this transition as being forced or caused by another agent (Ann in this case). Now once Jill is in state j_1 , it is her turn to act. She can move to state j_2 by

performing action t or simply stay in state j_1 by doing nothing. But in any case *knowing* of v corresponds to being in state j_1 .

Ann's automaton will be similar. Let $\Sigma_A = \{v, m\}$ and $Q_A = \{a_0, a_1, a_2\}$. Ann's initial state is a_0 , when her father vomits she transitions from a_0 to a_1 . While in a_1 she can choose to do nothing or perform action m to move to state a_2 . But she *knows* of v as she is in state a_1 .

The following figure depicts the above finite automaton. The dashed line represents transitions induced by other agents or the environment, and the solid line represents the choices that each agent can make.

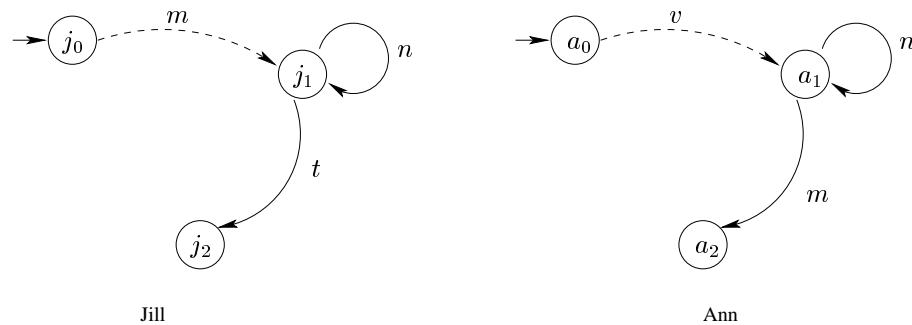


Figure 3

Define values as follows. All histories in which neither v nor t occurs have value 2, Those histories in which v occurs but not t have value 0 as they are the worst. Those histories in which v is followed by t are assigned as follows. Let H be a history in which v is followed by t , $val(H) = \frac{1}{N+1} + \frac{1}{M+1}$, where N is the number of clock ticks between the occurrences of v and m , and M is the number of clock ticks between the occurrences of m and t . Those in which t occurs without v have value 1 as do those in which v is followed by t . This valuation not only means that both Ann and Jill have to act, but that they should act speedily, for any delay leads to histories with lower values.

References

- [BPX] Belnap, N., Perloff, M., and Xu, M., *Facing the Future*, Oxford 2001.
- [Hi] Hilpinen, R., Deontic Logic, in *Blackwell guide to philosophical Logic*, Ed. Lou Goble, Blackwell 2001, 159-182.
- [Ho'01] Horty, J., *Agency and Deontic Logic*, Oxford 2001.
- [HMV] Halpern, J., R. van der Meyden, and M. Vardi, Complete axiomatizations for reasoning about knowledge and time, *SIAM journal of computing*.
- [LS] Lomuscio, A., and M. Sergot, Deontic interpreted systems, *Studia Logica*, **75** (2003) 63-92.
- [P95] Parikh, R., 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, 127-42.

- [P03] Parikh, R., Levels of knowledge, games, and group action, in *Research in Economics*, **57**, (2003) 267-281.
- [PaPa] Pacuit, E., and R. Parikh, A Logic for communication graphs, to be presented at the *Association for Symbolic Logic* annual meeting in Pittsburgh, May 2004.
- [PR'85] Parikh, R., and R. Ramanujam, Distributed processes and the logic of knowledge, in *Logic of Programs*, LNCS #193, Springer 1985, pp. 256-268.
- [PR'03] Parikh, R., and R. Ramanujam, A Knowledge based Semantics of Messages, in *J. Logic, Language and Information*, **12**, (2003) 453-467.
- [VM] van der Meyden, R. The Dynamic Logic of Permission, *Journal of Logic and Computation*, Vol 6, No. 3 pp. 465-479, 1996.