E. Pacuit, R. Parikh & E. Cogan
The Logic of Knowledge Based Obligation

Helen Kwong

January 15, 2010

Introduction

- Pacuit, Parikh & Cogan combine deontic logic with logic of knowledge and belief.
- Motivation: many obligations depend on what the agent knows.
  - **Example 1:** Uma is a physician whose neighbor is ill. Uma does not know and has not been informed. Uma has no obligation (as yet) to treat the neighbor.
  - **Example 2:** Uma is a physician whose neighbor Sam is ill. Sam'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.
- Difference: Uma's **knowledge** of her neighbor's sickness.

Another motivating example

Example 3: Uma has a patient with a certain condition C who is in the St. Gibson hospital. 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.

Here, Uma has the **default obligation** to administer drug *d*, based on her **justified belief** that *d* is the right drug. But this default obligation can be over-ridden by new information.

Outline

- Introduction and motivating examples
- History-based framework for knowledge, actions, and values
- Definition of knowledge based obligation
- Formalizing the examples
- Informal definition of justifiable beliefs and default obligation
- Discussion

Main idea: An agent is obliged to perform action *a* if he knows (based on events he has observed) that it is good to perform *a*.

History-based knowledge framework

Events and global histories

- Fixed set of **events** *E*
  - Can be actions by agents, or system events
- A **global history** *H* is a (finite or infinite) sequence of events from *E.*
  - $H \preccurlyeq H'$ denotes that *H* is a finite prefix of *H'*
  - For any finite *t*, H_t denotes the prefix of *H* consisting of the first *t* elements, i.e., the history up to time *t*.
- The **protocol** $\mathcal{H}$ is the set of all possible global histories
  - Limits the possible histories an agent may consider

History-based knowledge framework, cont.

Agents and local histories

- Fixed set of **agents** $A = \{1, 2, \ldots, n\}$.
- Each agent i observes a set of **local events** $E_i \subseteq E$.
  - E.g., Ann observes Sam vomiting, but Uma does not
- For each agent i, the **local view function** λ_i maps any finite global history H to the observed local history for agent i. Non-observed events are mapped to a system clock tick c.
  - E.g., if $H = wxyzx$ and $E_i = \{w, y\}$, then $\lambda_i(H) = wcycc$.
- Equivalence relation: $H \sim_i H'$ iff $\lambda_i(H) = \lambda_i(H')$
  - $H \sim_i H'$ means agent i cannot distinguish the two histories

Logic for knowledge and time

Syntax

$$\varphi := p \in \mathrm{At} \mid \neg\varphi \mid \varphi \vee \psi \mid \mathrm{O}\varphi \mid \varphi\mathrm{U}\psi \mid K_i\varphi$$

Semantics

Given $\mathcal{H}$, $E_1, \ldots, E_n$, and a valuation V, which maps any finite global history H to the set of atomic propositions true at H:

- $H, t \vDash p$ iff $p \in V(H_t)$, for $p \in \mathrm{At}$
- $H, t \vDash \neg\varphi$ iff $H, t \nvDash \varphi$
- $H, t \vDash \varphi \vee \psi$ iff $H, t \vDash \varphi$ or $H, t \vDash \psi$
- $H, t \vDash \mathrm{O}\varphi$ iff $H, t + 1 \vDash \varphi$ ("φ holds at the next moment")
- $H, t \vDash \varphi\mathrm{U}\psi$ iff for some $m > t$, $H, m \vDash \psi$ and for all k such that $t < k < m$, $H, k \vDash \varphi$ ("φ until ψ")
- $H, t \vDash K_i\varphi$ iff for all $H' \in \mathcal{H}$ such that $H_t \sim_i H'_t$, $H', t \vDash \varphi$

Actions

- Each agent *i* has a set $\mathrm{Act}_i \subseteq E$ of actions he can perform.
  - Act_i and Act_j are disjoint if $i \neq j$.
- For any action *a* and finite history *H*, define

  $$a(H) = \{H' \in \mathcal{H} \mid Ha \preccurlyeq H'\}.$$

  Another possibility: histories in which *a* is done *eventually*.
- Add a PDL-style modal operator $[a]\varphi$ to our language:

  $$H, t \vDash [a]\varphi \quad \text{iff} \quad \text{for all } H' \in a(H_t),\ H', t+1 \vDash \varphi$$
- Assumption 1: At any moment, only one agent can perform any action. If he does nothing, then nature does a clock tick.
- Assumption 2: Each agent knows *when* he can perform an action, i.e., $\langle a_i \rangle \mathrm{T} \rightarrow K_i \langle a_i \rangle \mathrm{T}$.

Values and goodness

- Each global history H is assigned a **value**, val(H)
  - All agents share a social utility function
- For any finite history H, the ***H-good* histories** (denoted $\mathcal{G}(H)$) are the extensions of H with the highest value:

$$\mathcal{G}(H) = \mathrm{argmax}(\mathrm{val}[\{H' \in \mathcal{H} \mid H \preccurlyeq H'\}]).$$

- For each action a, add **G(a)** to our language, with the intended meaning that "**action *a* is good**". Truth definition:

$$H, t \vDash G(a) \quad \text{iff} \quad \mathcal{G}(H_t) \subseteq a(H_t).$$

  All the H_t-good histories have a as the next action.

Knowledge-based obligation

An 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, i.e., $H, t \vDash K_iG(a)$. The truth condition is

$$(\forall H')(H_t \sim_i H'_t \text{ and } H' \in \mathcal{G}(H'_t) \Rightarrow H' \in a(H'_t)).$$

Formalizing the examples

Agents: $A = \{u, s, a\}$ (Uma, Sam, Ann)

$\mathrm{Act}_u = \{\mathrm{r}\}$ (Uma treating Sam)

$\mathrm{Act}_a = \{m\}$ (Ann telling Uma)

$\mathrm{Act}_s = \{v\}$ (Sam vomiting)

Observed events: $E_u = \{r, m, c\}$, $E_a = \{r, v, m, c\}$, $E_s = \{r, v, c\}$.

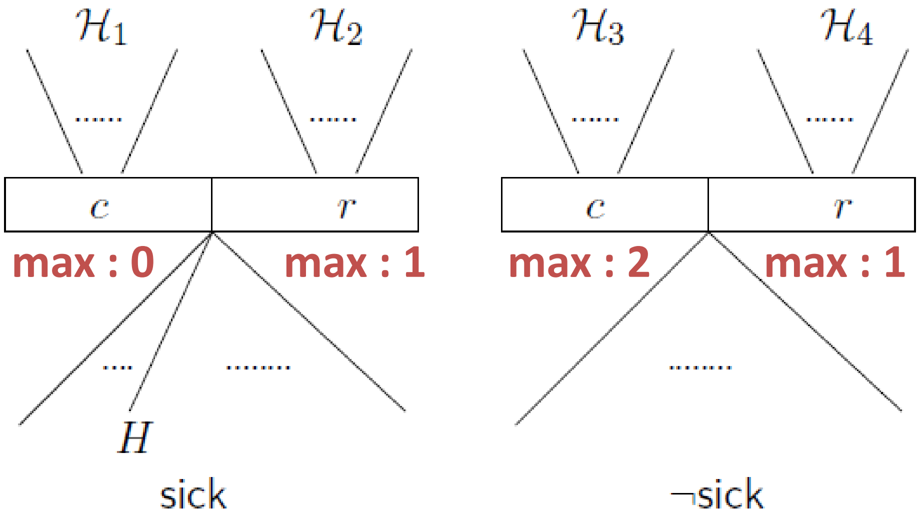
Protocol $\mathcal{H}$: Assume that in each possible global history, *v*, *m*, and *r* occur at most once. Also assume that *m* never occurs without *v* occurring first (Ann is truthful).

Values: Histories in which neither *v* nor *r* occurs have value 2. *v* occurs, followed by *r*: 1. *r* occurs without *v* occurring: 1. *v* occurs without *r* occurring afterwards: 0.

Let *sick* be a propositional variable true at any finite history in which *v* has occurred without *r*.

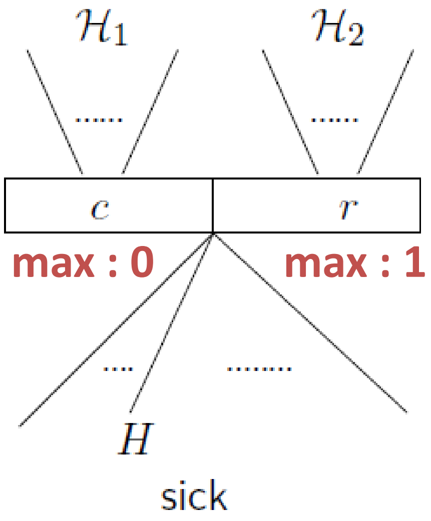
Example 1 formalized

Uma does not know that Sam is sick. That is, $H, t \vDash \neg K_u sick$. Then there must be some possible history H' such that $H_t \sim_i H'_t$ and v has not occurred in H'_t. The maximally good extension of H' does not involve r, i.e., Uma treating Sam (if Sam is not sick, the value is greater if Uma does not offer to treat him). So Uma does not know that it is good to perform r, and thus have no obligation to perform r.



Example 2 formalized

Ann tells Uma that Sam is sick, i.e., *m* occurs. Uma knows from the protocol that if *m* occurs, then *v* must have occurred (Ann does not lie). So she eliminates those histories in which *v* did not occur. Now for all histories *H'* such that $H_t \sim_i H'_t$, *r* is the next action in all the best-case extensions of H'_t. So Uma knows that it is good to perform *r*, which means she has the knowledge based obligation to treat her neighbor.

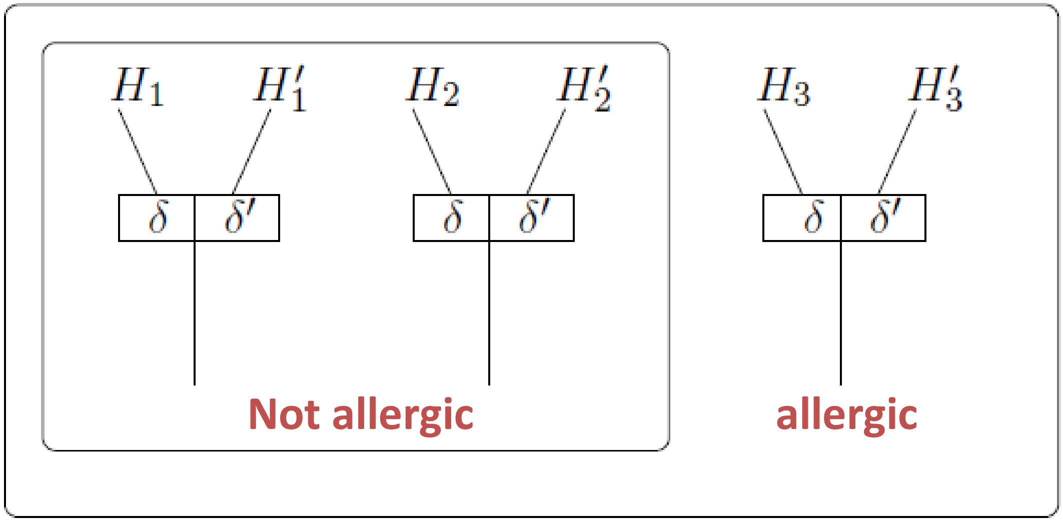


Default obligation

- Agent i **justifiably believes** φ at H, t, denoted $B_i\varphi$, if φ is true in the histories i considers the most plausible, given the events that i has seen.
- Define a system of spheres $\mathcal{H}_1$, $\mathcal{H}_2$, ..., where $\mathcal{H}_i$ is "more plausible" than $\mathcal{H}_j$ if $i < j$, and the union of the spheres is equal to $\mathcal{H}$. Take the least sphere that contains a history i considers possible (given the events he has seen). The i-plausible histories are the histories in this sphere that are possible to i.
- Agent i has a **default obligation** to perform action a at global history H at time t iff a is an action which i can perform, and i **justifiably believes** that it is good to perform a, i.e., $B_iG(a)$.

Example 3 formalized

Originally, histories in which the patient is allergic to drug *d* and those in which the patient is not are both possible to Uma. But those in which the patient is not allergic are the most plausible to Uma. So she **justifiably believes** that the best histories involve administering drug *d*, which means she has the **default obligation** to give drug *d*. Learning that the patient is allergic eliminates those most plausible histories as possible histories and thus the default obligation.



Kitty Genovese example

A woman was attacked and murdered in her neighborhood, and many neighbors saw what was happening, but no one called the police until some 35 minutes after the attack.

- No one had the (default) obligation to call the police, because to the witnesses, it was possible (more plausible) that someone had already called.
- Contrasting example: A child is crying in a waiting room. No one has a default obligation to comfort the child, but someone will see that no one else is taking care of the child and assume responsibility (there is **common knowledge** that the child is not being comforted).

Discussion

- Common knowledge leads to better outcomes
  - Bad outcomes can be avoided if there is more common knowledge
  - Common knowledge of ethicality: that everyone is ethical is common knowledge. For Ann to carry out deduction that he has the obligation to tell Uma, we need to assume $K_a(K_u sick$ <-> O treat)
- Agent *i* is obliged to do *a* if he knows *a* is performed in the best possible histories.
  - It seems *i* should be allowed to do an action that is not as good as *a* but does not lead to bad outcomes. Maybe more appropriate to oblige *i not* to do something that he knows is bad?

Discussion, cont.

- Obligation to know: For example, the hospital has an obligation to be aware of the patient’s condition.
  - Individual agents’ obligations to know or have somewhat accurate beliefs.
- Computational complexity: How hard is it to determine what is true in all the worlds consistent with what agent *i* has seen?
  - If it is too difficult, we may not expect *i* to actually know that *a* is good. Then he should not have the obligation?
- Formalizing the Kitty Genovese example
  - Current model not enough, if we assume that the later someone calls, the lower the value.