CHALLENGES WITH OBLIGATIONS IN MULTIAGENT DEONTIC LOGIC

Raik Hipler

Universität zu Lübeck
raik.hipler@student.uni-luebeck.de

Mark Scheibner

Universität zu Lübeck
mark.scheibner@student.uni-luebeck.de

ABSTRACT

With SDL (standard deontic logic) and DSDL (dyadic standard deontic logic) we have frameworks with which we can formally describe how things *ought to be*. In the context of (multi-)agent systems, however, we want to be able to describe the actions that agents may or may not take, i.e. we want to describe what actions agents *ought to do*. If we simply redefine "ought to do" as "ought to be done" – meaning that the obligation to do an action would be represented by the state of that action being done – we will run into various problems. This report will discuss some of these problems and how to deal with them.

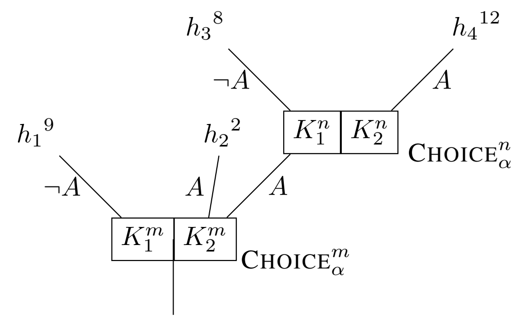

Fig. 1. An example of a STIT-tree

1. INTRODUCTION

When using deontic logic to reason about multi-agent systems the need for a clear definition on how actions of those agents should be modeled arises. The easy solution is to redefine actions as states, where doing an action σ becomes the state "σ has been done". As shown in sections 2, 3 and 4 in [1], this simple redefinition may lead to situations where agents may take counter-intuitive actions, or where sub-optimal situations are encountered. In this context *sub-optimal* means that we reach an outcome that has a low value according to some predefined utility-function.

Before we discuss some of those problems it is necessary to define a logic for modelling (multi-)agent systems in deontic logic.

1.1. STIT-trees

Since we are dealing with choices an agent can make (for example to do or not do an action) and possibly multiple consequences those choices can have, the model for this logic will be trees. An example tree to Horty's STIT (See-To-It-That) logic is given in Figure 1.

In these trees the interior nodes – in our case CHOICE_α^m and CHOICE_α^n – are moments in time where a decision can be made by the agent in question. In this context a decision at moment m is a set of actions, of which the agent must take exactly one. For CHOICE_α^m the agent α can decide between K_1^m and K_2^m, so $\text{CHOICE}_\alpha^m = \{K_1^m, K_2^m\}$. Actions can be non-deterministic, i.e. the same action may lead to multiple different outcomes – this is the case for K_2^m. One could also associate probabilities to each edge, but since the authors of [1] did not do so, we will not as well.

Depending on which action the agent takes, he either will be presented with another decision or reach an outcome. The leafs in those trees are the histories or traces of made decisions, and with each of them is associated a utility value that is based on the state reached, e.g. h_1 has a utility of 9. Finally, each edge is assigned a set of propositions that are true at this point in time.

1.2. Semantic on STIT-trees

For the semantics we always look at a moment-history pair m, h and check whether $m, h \models \varphi$ holds. If our formula is simply an atomic proposition A, we get $m, h \models A$ iff A is true at the outgoing edge of m on the path to h. For example, this is the case for m, h_3 but not for m, h_1 and n, h_3. A formula $\mathcal{F}A$ is fulfilled iff A holds at some point on the path from m to h (e.g. $m, h_3 \models \mathcal{F}A$ but also $m, h_3 \models \mathcal{F}\neg A$). The relation $m, h \models \bigcirc A$ holds iff A holds on all edges of m that reach histories having the highest utility value for those histories reachable from m. Note how h is irrelevant. For example, this is the case for $m, h_2 \models \bigcirc A$ since A holds in the best history reachable from m, namely, h_4.

To deal not only with obligations but also agents, a new operator is introduced: $[\alpha\,\text{cstit}\colon A]$ meaning that α sees to it that A (*stit* for *sees to it that* and the *c* from the name of the logician *Chellas*). The relation $m, h \models [\alpha \,\mathrm{cstit}\colon A]$ is fulfilled iff A is true for every pair m, h' that belongs to the same action as m, h. This is the case for m, h_3 since A holds at every outgoing edge of K_2^m.

2. NON-DETERMINISTIC ACTIONS AND THE GAMBLING PROBLEM

In order to reapply the known deontic logic semantics in the world of agent systems, one may intuitively think that the obligation to do something is the same as the obligation that something will be done or that a certain history will be reached. This, however, becomes challenging if an action may lead to multiple histories in an unpredictable manner.

2.1. The gambling problem

The first problem concerning agents the authors talk about is the *gambling problem* [1]. In this scenario an agent is presented with two options – he gambles and either doubles or loses his bet or he does not gamble and thus preserves his money. An example tree for which the agent starts with 5€ is given in Figure 2. Here the utility value describes how much money he has in the end and the proposition G means "he gambles". The question is whether he should gamble hoping for the sweet win but also risking to lose everything or rather refrain from doing so.

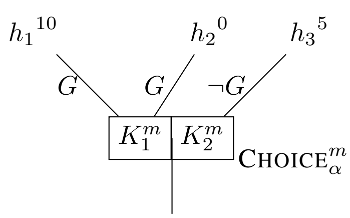

Fig. 2. An example STIT-tree for the gambling problem

2.2. Optimality of non-deterministic actions

The optimal path would end in h_1, so in order to get to this best outcome the agent should "ought to see to it that G". Intuitively, one would try to formalize this with $\bigcirc[\alpha \,\mathrm{cstit}\colon G]$, however, the action K_1^m is not necessarily the better option because, in the end, the outcome of the gamble comes down to luck.

This problem arises from the fact that optimality of a history is defined only by its utility value disregarding that the actions that lead to this history are non-deterministic. Or in other words, while an optimal history concerns ought-to-be and optimal actions concern ought-to-do, we cannot infer the optimal action from the optimal history because an action may lead non-deterministically to multiple (and potentially worse) histories. Rather, one should correspond actions to sets of histories and compare the optimality of all possible histories to determine the optimal action.

Horty suggested that an action K_1 should be considered more optimal than K_2 if each of its histories have at least the same utility value than any of the histories reachable from K_2, while the other way is not the case. This leads to a new operator $\odot[\alpha \,\mathrm{cstit}\colon G]$ meaning "α ought to see to it that G". In contrast to $\bigcirc[\alpha \,\mathrm{cstit}\colon G]$ this formula is fulfilled iff for each action not leading to G there is a better action that does lead to G. Intuitively the difference lies in how to optimize the result:

The $\bigcirc$-operator will go for the best history disregarding the fact that an action taken along the way may non-deterministically lead to a worse history, while the $\odot$-operator allows to take "sub-optimal" actions (not gambling in the example) if there is an outcome that is still better than the worst outcome of the other decision.

Since K_1^m and K_2^m correspond to equally good sets of histories, neither $\odot[\alpha \,\mathrm{cstit}\colon G]$ nor $\odot[\alpha \,\mathrm{cstit}\colon \neg G]$ hold for m.

3. MORAL LUCK

In some cases the utility of an agents decision might be dependent on another agents decision. This becomes particularly problematic if this dependence is circular and the involved agents have no way of communicating or reasoning about the decisions of the others.

3.1. The driving problem

In [1] the authors discuss this problem with the example of the *driving problem*:

Suppose an agent α is driving on a small road and another driver, agent β, is travelling towards α. Both agents have the option of swerving to the side or continuing to drive straight. However, should they decide on the same action they will collide. It's obvious that there are two good and two bad outcomes: the good ones are, where only one of the persons moves to the side, the bad ones are where both move to the side or both do not. Figure 3 illustrates how α has to decide between K_1^m(swerve) and K_2^m(do not swerve) and β between K_3^m(swerve) and K_4^m(do not swerve). Depending on the combination of their decisions m either leads to an optimal solution h_1 or h_3 or they crash in h_2 or h_4. The proposition S stands for "α swerves".

This problem is somewhat similar to the gambling problem since the utility of an action of an agent is based on external influences. In this case, however, it is not a random event, but the action of another independent agent whose actions α cannot reason about.

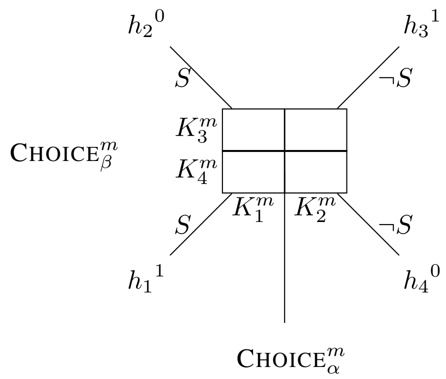

Fig. 3. A STIT-tree for the driving problem

3.2. Dominance act utilitarianism and orthodox perspective

We now face the problem whether $\odot[\alpha \, \mathrm{cstit}\colon S]$ should be true or not at the moment of the decision. There are generally two established views on this problem.

The first is dominance act utilitarianism, where an action is only optimal, if each of its outcomes is optimal. More formally the formula $\odot[\alpha \, \mathrm{cstit}\colon S]$ is true only if all actions that lead to optimal outcomes also guarantee S. Since this is not the case in the driving problem, we have that $\odot[\alpha \, \mathrm{cstit}\colon S]$ does not hold at moment m. Note that this is also the case for $\odot[\alpha \, \mathrm{cstit}\colon \neg S]$ since both actions (swerving and staying in lane) may lead to optimal outcomes, which agent α ought to strive for.

In the second view, the orthodox perspective, the actual circumstances the agent will find themselves in after making the decision are also taken into consideration. Intuitively α should ask themselves what *would* the best decision have been. This means that instead of looking at $\odot[\alpha \, \mathrm{cstit}\colon S]$, the question of whether $\bigcirc[\alpha \, \mathrm{cstit}\colon S]$ should be true at moment m for the history h that occurred can be answered. In this case $\bigcirc[\alpha \, \mathrm{cstit}\colon S]$ is true for h_1 and h_4, and likewise $\bigcirc[\alpha \, \mathrm{cstit}\colon \neg S]$ is true for h_2 and h_3.

4. PROCRASTINATION

Another problem that arises is the problem of procrastination. If an agent agrees to take on an obligation, the agent may indefinitely put off following through with that obligation until a deadline is reached. This is especially a problem if the highest utility is achieved by that specific agent taking on that obligation and then following through.

4.1. Professor Procrastinate

The example given in [1] describes a professor – aptly named Professor Procrastinate – who is asked to write a review. As he is the best person available for this, the optimal situation is him agreeing to write the review and then actually doing it. However, this professor is known to procrastinate and even though he has the choice to actually write the review, he will not in fact do that, thus the review will never be written. Since this is the worst case it would be better for him to decline the request so that someone else can write the review.

This leads us to the contradiction that the professor ought to accept the request and write the review since this is the optimal course of action, but at the same time he ought to decline the request since doing otherwise will lead to the worst possible outcome.

4.2. Fields and actualism

The problem that leads to the contradiction is the existence of a possible choice that is in a way already settled. A possible solution to this problem is to narrow down the choices that are taken into consideration to those that are actual choices. This view on the problem is fittingly named actualism.

For a STIT-tree this means only a subset of the possible choices will be considered, which gives us a sub-tree of the problem that may be a more appropriate model of reality. This sub-tree, on which the formula is evaluated, is called *a field*.

For the example of Professor Procrastinate we can safely take into consideration only the first choice as him accepting or declining the request will not change the fact that he will not write the review. So our field will contain only the first decision. Since we are now no longer allowed to reason about the possibility of the professor writing the review, it is easy to see that he ought to decline.

Let us take a look at what this means for the semantics of the formula $\odot[\pi \, \mathrm{cstit}\colon A]$, where π is the professor and A is the decision to accept the request. If our field contains all possible choices – a view which is also called *possibilism* – then $\odot[\pi \, \mathrm{cstit}\colon A]$ should be true, since this is the only choice leading to the optimal outcome. But using the field containing only the first choice we have that $\odot[\pi \, \mathrm{cstit}\colon A]$ is false, since with the background assumption that Professor Procrastinate will not write the review, the actual optimal outcome can only be reached if the professor declines.

5. SUMMARY

Despite seeming like a straightforward conclusion to talk about actions by talking about histories inferred by them, it is not as simple as to reduce *ought-to-do* to *ought-to-be-done*. Rather, one has to accommodate for the specifics that come with the usage of agency, including, among other things, different histories caused by non-deterministic actions, multiple agents influencing each other and the difference between what *could* happen and what *actually* happens.

For the gambling problem we can circumvent the problem by modifying the semantics to consider every possible outcome instead of only the best one.

In the case of the driving example two ways of looking at the problem were given, both of which do not actually solve the problem but rather work around it.

The problem of procrastination was solved by using additional information – in the example that is the information that the professor is not going to write the review – and restricting ourselves to those histories that are still feasible given that information.

As we can see, some of these solutions can give rise to real procedures for automating decision-making while others only allow us to view the problem from different perspectives.

6. REFERENCES

[1] Gabriella Pigozzi and Leon van der Torre, "Multiagent deontic logic and its challenges from a normative systems perspective," *IfCoLog Journal of Logics and Their Applications*, pp. 2929–2993, 2017.