Reactive Kripke Models and Contrary to Duty Obligations

Dov M. Gabbay

King's College London
Version 1: 31 March 08

This is an intuitive description of our approach to modelling contrary to duty obligations. We shall describe our ideas through the analysis of typical problematic examples taken from Carmo and Jones [6], L. van der Torre [14] and Prakken and Sergot [5].

1 Preliminary Discussion

Contrary to duties (CTD) are dealt with in the framework of standard deontic logic (SDL), and ordinary Kripke possible world models. Given a world t, one associates statically a non-empty set $I(t)$ of ideal worlds for t and $t \vDash Oq$ (q is obligatory for t) if q holds in all the worlds of $I(t)$.

This is a static perception of obligation. If we have to list as Δ_t the set of all obligations for the world t then $I(t)$ would be the set of all models of Δ_t. The contrary to duty examples have some implicit dynamics in them. It is therefore not surprising that there are problems with the formalisation of various CTD examples within SDL. There are currently in the literature various proposals for solutions, however all are still largely within the STL possible world semantics approach or its extensions, with additional operators or preferential ordering. See footnote 2 below and references [18], [15] and [13].

Reactive Kripke models is a stronger version of possible world semantics, affording the semantic characterisation of more modal systems (this is a theorem in [1]. They have a dynamic dimension to them. Therefore using this new semantics might simplify existing solutions to CTD problems as well as offer new sharper solutions.

Note that this new approach does not necessarily abandon or challenge any of the existing solutions, since ordinary Kripke models are a special case of reactive Kripke models. This is an important point to bear in mind. We can proceed on two fronts.

1. Take an existing solution, say the Carmo and Jones model of [6] and view it in the context of the richer reactive Kripke semantics and maybe simplify the models or sharpen the semantics, etc.
2. Offer a new solution of our own, maybe disagree with existing proposals but make our case using the stronger tool of reactive Kripke models.

Either way all benefit and we are in a win–win situation.

Our plan is to give some examples in detail to familiarise the reader with our ideas, leaving the formal machinery and the extensive discussion to the full version of the paper.[1]

[1] We illustrate in this outline how CTD problems can be solved but we do not commit that our examples are the final solution. In the full paper we will offer some final solutions after looking at the problems more thoroughly. The spirit of this paper is correct but the details may change.

R. van der Meyden and L. van der Torre (Eds.): DEON 2008, LNAI 5076, pp. 155–173, 2008.

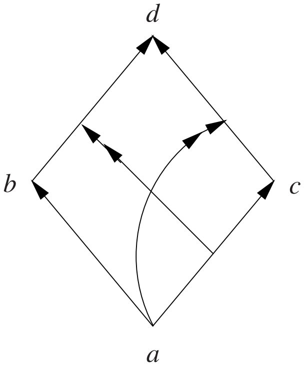

Fig. 1.

2 Reactive Kripke Models

Example 1 (A Reactive Kripke Model). Figure 1 shows such a model.

The single arrows show accessibility relation R. So in this figure we have aRc, aRb, bRd and cRd.

The double arrows are connections which can deactivate accessibility (we can use triple arrows to activate). We have

$aR(c,d)$	double arrow from a to the connection (c,d).
$(a,c)R(b,d)$	a double arrow from the connection (a,c) to the connection (b,d).

The best way to explain how evaluation works in such a model is by actually doing it.

So assume our model is (S,R,h), where

$$S = \{a,b,c,d\}$$
$$R = \{(a,b),(a,c),(b,d),(c,d),(a,(c,d)),((a,c),(b,d))\}$$

If Q is the set of atomic sentences, then h is the assignment. For each $q \in Q$ and $s \in S, h(s,q)$ is a truth value.

The language contains the usual classical connectives and $\Diamond$.

Let us evaluate

$$a \vDash \Diamond\Diamond q, q \text{ atomic}$$

We need to start at point a and move two steps through the accessibility relation and land at a point $x \vDash q$. We can either make our first step to b or to c.

First observe that the minute we leave point a the double arrow from a to (c,d) will cancel the connection cRd. So if we leave a to go to c, then when we get to c the point d will no longer be accessible to c. Furthermore, to go to c we pass along the arc (a,c). The minute we pass through (a,c) the double arrow from (a,c) to (b,d) will cancel the connection bRd.

So when we get to c the model will have changed. Figure 2 shows the model as it is when we go to c from a.

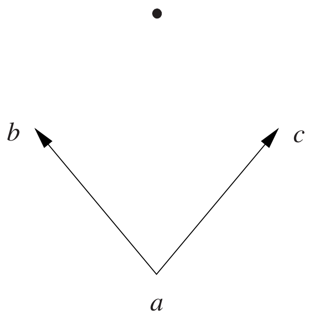

Fig. 2.

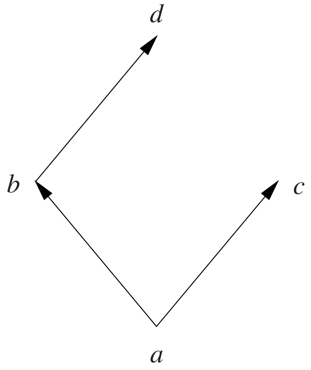

Fig. 3.

On the other hand, if we go from a to b, the connection (c, d) will be cancelled, but the connection bRd is still on and so we can continue to point d.

Figure 3 shows the model as it is when we get to d through the path a, b, d.

If indeed $d \vDash q$, then $a \vDash \Diamond\Diamond q$.

By the way, we have also shown that $a \vDash \Diamond\Box\bot$, because starting from a going to c we get $c \vDash \Box\bot$ as in Figure 2. Note that we need to know how we get to c in order to evaluate at c.

So the correct evaluation metapredicate should be

$$(x, y) \vDash A$$

where there is a unique path from x to y and we are evaluating A at y.

So y is our current evaluation point and x is our starting point. (We assume there is a unique path from x to y, otherwise we have to specify the path.) So we should write:

$$\begin{gathered}(a, a) \vDash \Diamond\Diamond q\\(a, b) \vDash \Box\bot\\(a, b, d) \vDash q.\end{gathered}$$

Note that (a, c, d) is not a path. so really (a, b, d) is a unique path to d.

Before we go on to contrary to duties, let us highlight the "take a walk" point of view of the evaluation. We imagine ourselves as agents standing at point a of Figure 1 and given a formula to evaluate or trying to reach a world (say get to d). We move along the arcs towards our goal worlds and evaluate along the way. Connections change as we walk up the arcs.

The logic is determined by the nature of the connections we allow and by the algorithm which tells us how to walk and evaluate. This point of view is dynamic, not static, and is very compatible with the semantic view of ideal worlds as objects to aspire for and contrary to duties. If we want to satisfy an obligation we must move towards an ideal world. If we deviate we might go in a direction which strays away from the ideal in which case some double arrows will change the connections and steer us in the direction of other subideal worlds. This view is very intuitive. It has implicit dynamics in it even though the model itself is static.

So given a world t and the obligations Δ_t for t, we do not just use semantics to describe Δ_t, i.e. use the set of ideal world $I(t)$ to characterise Δ_t. We actually think of $I(t)$ as worlds spread inside the possible world model and expect our agent to move along the accessibility relation to one of the worlds $I(t)$.

This is an action possible world model. The people at world t take action to move towards an ideal world. In this context contrary to duties become natural.

We must warn the reader that in any model there may be three types of movement.

1. Virtual movement towards the ideal world which is not temporal at all[2]
2. Temporal movement. See, e.g. [12]
3. A combination of (1) and (2).

There are CTD examples of all the above types. In fact some papers solved CTD puzzles of type (2) exploiting temporal operators. See [16] and the thesis of J. Broersen [17] (his use of 'reactive' is not the same as ours).

The model of Figure 1 is by no means the most general reactive model. We can allow double arrows with specific tasks, either to switch on a connection or to switch off a connection. We can also annotate connections and arrows as on or off.

Figure 4 is an example. Double arrows switch connections off. Triple arrows switch connections on.

The annotations 'on', 'off' say which arrows, double arrows and triple arrows are active at the start position before we move out of a. So, for example, (b,d) is off and $((a,b),(b,d))$ is on, and (a,b) is on.

We start at a. Moving out of a to b the double arrow from a to (c,d) switches (c,d) off. The triple arrow from (a,b) to (b,d) switches (b,d) on.

If we carry on from b to d then the double arrow from (b,d) to the double arrow $(a,(c,d))$ switches it off. We also see that triple or double arrows can go to other triple or double arrows, etc.

[2] Some CTD examples are completely static and do not involve time. Still the CTD aspect of the problem does involve movement towards the ideal worlds. How can this be? How can we give a static (non-temporal) model which still involves virtual movement? Well, we have such examples in classical mechanics. We solve a static distribution of forces in a structure by imagining a slight movement. This is called the principle of virtual work. For reference look up "virtual work" in Wikipedia.

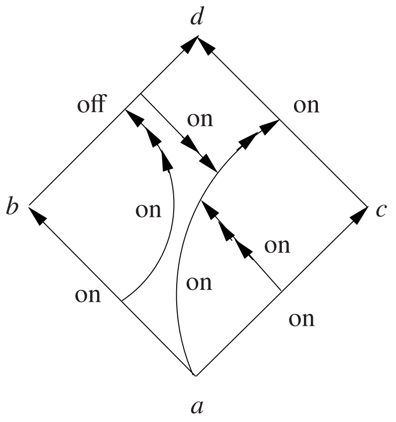

Fig. 4.

We can have suitable modal operators for walking along the arcs in any suitable way.

The next section gives a specific reactive semantics suitable for some analysis of CTD.

3 Contrary to Duty Models

We consider a reactive model of the form (S, R, I, h), where R is a relation say as in Figure 1 (no triple arrows) and I is the ideal world function giving for each $s \in S$ a non-empty set of ideal worlds $I(s) \subseteq S$. So actually $(S, I(s), h)$ is an SDL model for O, R gives us the reactivity. We assume two additional modalities. $\Diamond$ evaluated as above, taking into account the effect of the double arrows and $\Diamond\hspace{-0.55em}|\;$which is an ordinary modality which ignores the double arrows. So for $\Diamond\hspace{-0.55em}|$, Figure 1 becomes Figure 5.

Consider now the model in Figure 6. Its points are

S $=\{a, b, c, w^+, w^-, f^-\}$is as in Figure 6, which also indicates the meaning of the worlds
$R = \{(a, b), (a, c), (c, f^-), (b, w^+), (b, w^-), (a, (c, f^-)), ((a, b), (b, w^-))\}$

The function I satisfies $I(a) = \{f^-\}$. We don't care about the other values of I.

The minute we leave point a the connection with f^- is severed. This means the agent beginning at a is not able, according to this model, to follow a path to the ideal world f^-. Note that double arrows emanating from points are local properties of the model (which we can interpret as having to do with agent's circumstances). Double arrows emanating from connections are systems contrary to duties.

Thus the agent is not able to comply to his obligation and has to go for fence. He can go to point c and get stuck there or go to point b to continue to a world with a fence. As he passes the connection (a, b) the double arrow from (a, b) disconnects his way to the non-white fence world and he has to go to w^+.

Note that we need a starting point an an evaluation point. So the model has the form (S, R, I, h, a, x) (we fix a as the starting point for our example of Figure 6).

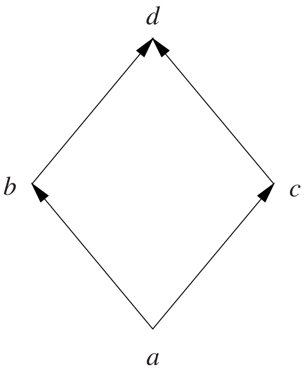

Fig. 5.

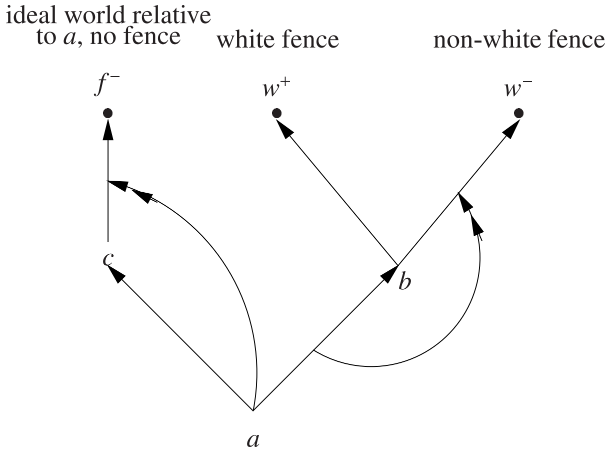

Fig. 6.

So by looking at a, w^+ we know the agent is not able to comply to his obligation and has to go for fence and he has a white fence because of CTD.

So a set of CTD sentences determines a class of reactive models. Given a model we can read from it the CTD sentences it suggests. We are saying ‘suggests’ rather than ‘holds’ because existence of double arrows suggests STD sentences, see section 5.

Let us look at the analysis of the scholarly work of Carmo and Jones [6, p. 305].

Statement of the Problem

(d1) *There must be no fence*
In the model this is implemented by f^- being ideal world relative to a, $f^- \in I(a)$ with a as the starting point.

(d2) *But if there is a fence it must be white*
This is implemented by the fact that any arc in the model where there is no continuation to an ideal world has double arrows emanating from it cutting access

to any non-white fence world ahead, i.e. in Figure 6 this is the double arrow $((a,b),(b,w^-))$.

We can add

(d3) *There is a fence*

This is implemented by saying we are at point $x \in \{w^+, w^-\}$. So, for example, the model of Figure 6 with starting point a and evaluation point x, $x \in \{w^+, w^-\}$ does model (d1)–(d3).

Carmo and Jones [6] did not have reactive models. They added additional modalities of actually possible and potentially possible and used them to make case analysis. We now quote their case analysis and show how to express the cases in our reactive models. The reader should note that we did not construct our model to simulate Carmo and Jones. Had we done so systematically we probably would have come up with a slightly different model, which implements additional modalities using reactivity.

Case 3.1

(f1) *There is no fence and it is still actually possible not to erect a fence and actually possible to erect a fence, white or not.*

In the model we look at point a without the double arrows emanating from it, i.e. to model Case 3.1, we take a model without double arrows at all. The starting point is a (which allows for all the possibilities) and the evaluation point is f^- (which allows for the fact that there is no fence). Alternatively, we can say $a \vDash$ no fence[3]

Case 3.2

(f1) *There is a white fence and it is actually fixed that there will be a fence, possibly white or another colour.*

To model this take a as the starting point and w^+ as the evaluation point. The double arrow from a makes it actually not possible to have no fence and the double arrow $((a,b),(b,w^-))$ blocks w^-.

(f2) *It is potentially possible to have no fence.*

This is clear since there is a connection path to f^-, if we ignore the double arrows (i.e. use $\Diamond\!\!\!\mid$ for potentially and $\Diamond$ for actually).

Let us do another example:

Example 2 (Chisholm paradox). The following is from Carmo and Jones [6, p. 299]

(d1) *It ought to be that a certain man go to help his neighbours*
(d2) *It ought to be that if he goes he tell them he is coming*
(d3) *If he does not go, he ought not to tell them he is coming*
(d4) *he does not go.*

Consider Figure 7. (d1) is modelled by $d \in I(a)$, i.e. d is an ideal world for a. (d4) is modelled by taking e as an evaluation point.

[3] Our purpose here is not necessarily to model and simplify Carmo and Jones [6] but to show we have the power to offer our own models or to model other approaches. See Remark 11. In fact, we do not need to make such a case analysis. In the full paper, we shall study in detail the case analysis of [6] as well as the works in [12], [18] and [19].

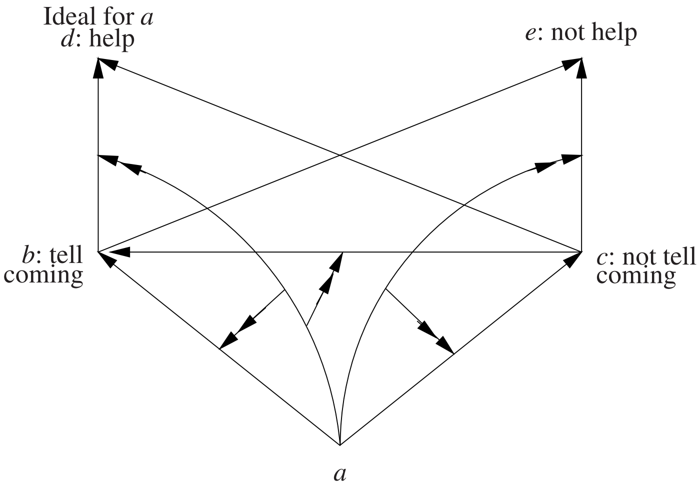

Fig. 7.

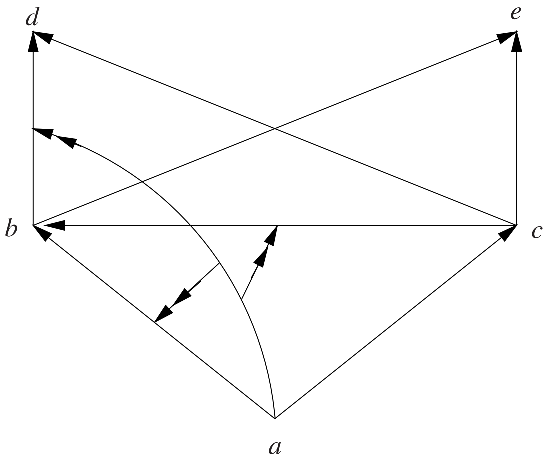

Fig. 8.

Modelling (d2) and (d3) is a bit challenging, because both help and tell are in the future and tell comes before help. See Example 6 below for a discussion. To model (d2) and (d3) first note that the double arrow from a to (b, d), triggers the system to send a double arrow from $(a, (b, d))$ to (a, b) and to (c, b). This models (d3). Second note that to model (d2) we have the double arrows $(a, (c, e))$ and $((a, (c, e)), (a, c))$. However, putting them both in the same model means that the man decides to block his path from going anywhere.

The way to solve it is to split Figure 7 to Figure 8 one with the single arrows of Figure 7 and one with only the double arrows $\{(a, (b, d))((a, (b, d)), (c, b)), ((a, (b, d), (a, b))\}$ modelling (d3) and Figure 9 with all the single arrows plus only the double arrows $\{(a, (, e)), ((a, (c, e)), (a, c))\}$, modelling (d2).

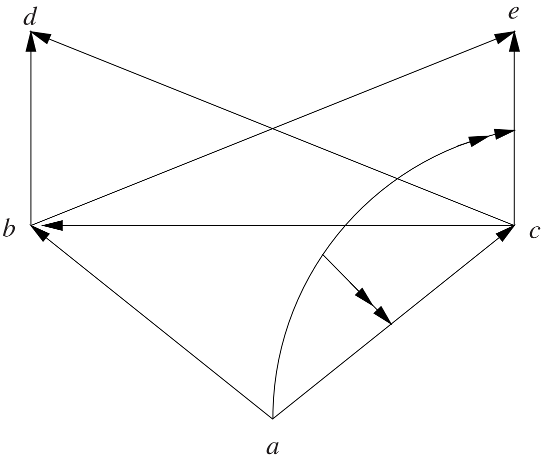

Fig. 9.

The actual modelling of the Chisholm paradox is the pair of reactive models done in parallel disjunctively. See Definition 8 and also [19].

We now follow the case analysis of Carmo and Jones [6, pp. 300].

Case 1.1

(f1) *The man decides not to go to help.*
This is modelled by Figure 8.

(f2) *It is potentially possible for the man to help and to tell and potentially possible for the man to help and not to tell*
This is modelled in Figure 8 by choosing the evaluation point as a. The beginning point is always a, so we have (a, a) as our pair. The decision in (f1) is the choice of Figure 8. The potentiality comes from the fact that the man has not started yet (evaluation point a) and he potentially can change his mind and choose the model of Figure 9.

(f3) *The man has not in fact told that he is coming to help although it is still actually possible that he does tell and actually possible that he does not tell.*
This is modelled by taking the evaluation point as point c. The man can actually move either from c to e or from c to b and then to c.

The other cases of Carmo and Jones, namely case 1.2 and case 1.3, [6, pp. 300 and 301] can be done similarly.

Case 1.4 contains the fact that the man helped but that it was potentially impossible for the man to tell his neighbour. For this we need a variation of Figure 9 where there is no connection from a to b. The man took the path a to c to d. See Figure 10.

Case 1.5 is where it is not potentially possible to help but the man does tell he is coming but he might have not told.

This is covered by Figure 11.

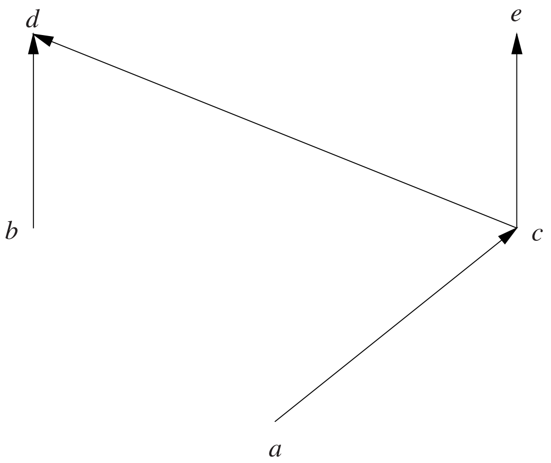

Fig. 10.

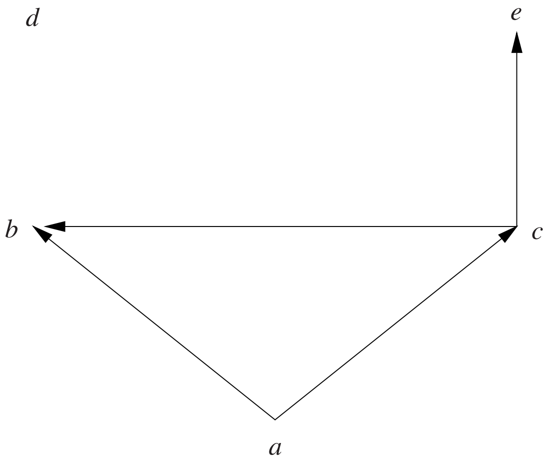

Fig. 11.

In the full version of this paper we shall analyse the Carmo and Jones examples in detail. Note that since we are using reactive models we do not need any case analysis. Our two diagrams 8 and 9 actually solve the paradox in principle.

4 Concluding Remarks

Remark 1 (The Idea of Reactivity). The idea of reactivity is a general one applicable across research areas. Whenever we have a system with states and algorithms involving these states we can turn it reactive by allowing signals from states or transitions which are being used to some other states causing change in the other states.

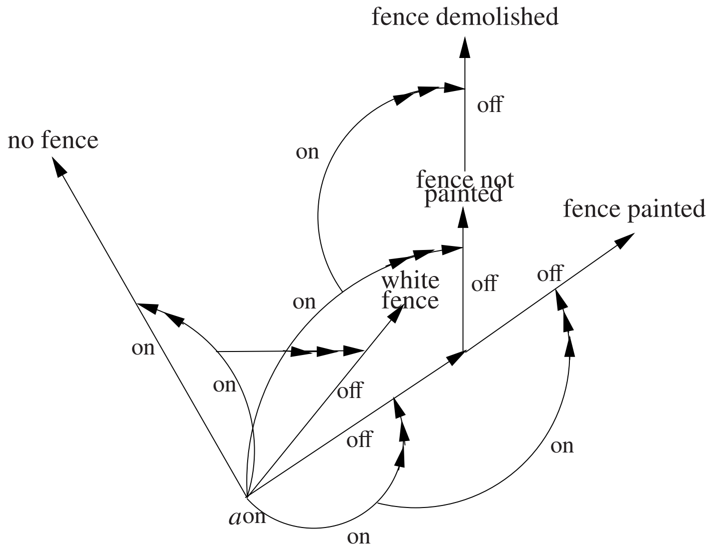

Fig. 12.

This change can be due to faults and overuse of the system, or feedback in the system or object level implementation of norms regimentation in the system or just efficiency shortcuts embedded in the system by design. We are now systematically studying reactive automata, reactive grammars, reactive conditionals, reactive proof theory and more.

Remark 2 (Proof Theory for Reactive Semantics). Proof theory can be provided for reactive Kripke models in the methodology of LDS (Labelled Deductive Systems). This means we can propose models for CTD systems as well as proof theory. This also means that we can provide LDS proof theory for existing CTD systems such as, for example, the Carmo and Jones proposals [6]. More on this in the full version of the paper.

Remark 3 (Multiple Level CTD). We have no inherent difficulties with multiple level contrary to duty.

For example:

1. It is obligatory to have no fence.
2. If there is a fence it should be white.
3. If it is not white it should be painted white.
4. If it cannot be painted (some plastics cannot take paint, I have some in my office at King's) then it should be demolished.

Figure 12 illustrates a possible model.

The beginning position is that all arrows are off except the one leading to no fence. The arrows emanating from *a* block the path to no fence and clear a path to fence not painted. The arrows from the agent's arrows force him to demolish.

Remark 4 (Conflicting Norms). We can cope more easily with conflicting norms. The modern world is full of them. Think of

1. There should be no fence
2. There should be no dog
3. If there is a dog there should be a fence
4. If there is a fence it should be demolished
5. There is a dog

In the reactive model we can loop, repeatedly erecting and demolishing a fence and thus fulfill our obligations, that is assuming we insist on a dog.

Remark 5 (Expressing Ideal Worlds using Double Arrows). The additional power of the double arrows and triple arrows can be used to eliminate the ideal world function $\lambda t I(t)$. In a model with connections capable of being on or off, we can characterise $I(t)$ as all those worlds accessible to t by an active direct connection and the non-ideal worlds as those not accessible. The minute we make our first move we can activate and deactivate connections to bring us back to whatever connections we want. So in fact the ideal worlds are recognised by the way we do our on and off switches. Figure 7 becomes the new Figure 13 below.

Here we use arrows, double arrows and triple arrows.

The starting position is that this access (on) only to the ideal world f^-. The minute we move from a, we cancel access to it and activate access in the $\{b, w^+, w^-\}$ direction. We know f^- is ideal because disconnecting access to it makes the difference.

There may be better ways to do the coding. We just want to illustrate the principle involved.

Remark 6 (Solving CTD Problems using Reactive Proof Theory). Reactive proof theory can be used directly to solve problems of CTD. The idea of reactive proof theory is that using a rule can activate or de-activate other rules. So, for example, we may have

1. $O\neg f$. There must be no fence.
2. $f \to Ow^+$ If there is a fence then it must be a white fence.
3. $\vdash w^+ \to f$. White fence is a fence

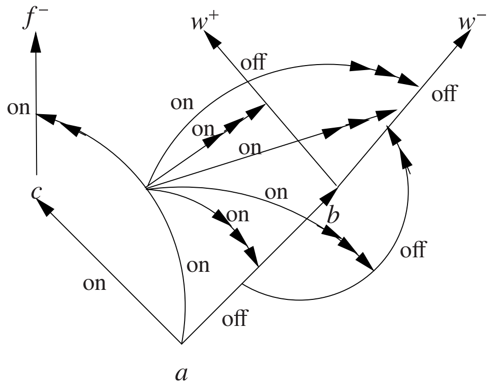

Fig. 13.

4. f. There is a fence.
5. Reactivity rules: if we use 6.2 we cannot use item 6.1.

This is crude but effective.

Before the time when we understood the reactive proof theory, such rules might have looked (to the traditional logicians) as a hack, a trick without semantical meaning. However, with reactive semantics, a competently crafted system with such rules may actually be characterised by a class of reactive models.

The above example may be too crude and some balanced refinement may be needed, but it does illustrate the idea. Note that the use of 6(2) as a ticket corresponds the move from point a to point b in Figure 8. The ideal world f^- is no longer accessible so 6(1) cannot be used. Ow^+ is derived. This corresponds to the CTD rule.

I have studied rule cancellations and deletion in logic extensively. I assure you this is workable. See my papers [8,9] and [10].

Remark 7 (Comparison with Dyadic Obligations O_AB). We now comment on the use of the binary operator O_AB to express CTD B if A (also denoted $O(B/A)$. Semantically this is a powerful operator on the set of possible worlds corresponding to a binary relation xR_Ay indexed by subsets (the truth set of A). If you look at Figure 8 again, you see that the double arrow depends on a pair of points. So, for example, we formally can write $\neg bR_{(a,b)}w^-$, i.e. $R_{(a,b)}$ represents the double arrow. So formally we need relations indexed by pairs of points at the most, not all subsets. O_AB is too powerful. I also think in addition to the above technical points that O_AB as a different concept and should not be used as a hacking coding ground for solving the CTD problems. We leave it at that, more in the full paper. Note that Carmo and Jones [6] whose a dyadic connective O_AB use for it a semantical function from subsets to families of subsets of the form

$$ob : w^S \mapsto 2^{2^S}$$

which is a very high level function.

Remark 8 (Input Output Logic). There is affinity between our reactive Kripke models and the work of Makinson and Torre on input output logics. I refer to [11]. Given a CTD of the form If A then obligatory that B, we can feed A as input to an input output node and get B as output. This idea can be incorporated into reactive Kripke models if we allow the worlds in the model to have input output facilities.

We shall elaborate on this in the full paper.

5 Technical Definitions

This section supplies the technical definitions supporting the intuitive ideas presented in this paper.

Definition 1 (Language). *Our propositional language contains a set of atomic propositions Q, the classical connectives and the modal connectives OA, O_AB, $\Diamond A$, $\Diamond\!\!\!\mid A$, and possibly more.*

Definition 2 (Reactive Arcs)

1. *Let S be a non-empty set. The set $\mathcal{A}$ of arcs on S is defined as the smallest set $\mathcal{A}$ containing $S \times S$ and closed under the following operation.*
   - *If $x \in S \cup \mathcal{A}$ and $y \in \mathcal{A}$ then $(x, y) \in \mathcal{A}$.*

   The above condition says that we can have for example $(t, (a, b)) \in \mathcal{A}$ but not $((a, b), t) \in \mathcal{A}$. We do not allow $(x, y) \in \mathcal{A}$ with $y \in S$.
2. *A set $\mathcal{A}_0 \subseteq \mathcal{A}$ of arcs is said to be well founded if whenever $(x, y) \in \mathcal{A}_0$, then $x \in S \cup \mathcal{A}_0$ and $y \in \mathcal{A}_0$.*

Definition 3 (Reactive Accessibility)

1. *An element $(x, y) \in S \times S$ is called a single arrow. We can also write $x \rightarrow y$.*
2. *Let $(x, y) \in \mathcal{A}$. (x, y) can be used either as a negative switch or as a positive switch. We regard (x, y) as a negative switch by writing it as a double arrow $x \twoheadrightarrow y$. We regard it as a positive switch by writing it as a triple arrow $x \mathrel{\to\!\!\!\!\twoheadrightarrow} y$.*
3. *A single arrow, double arrow or triple arrow can be either on (we put $+$ in front of it) or it can be off (we put $-$ in front of it).*
4. *An accessibility relation R is obtained from a well founded base of arcs $\mathcal{A}_R$ as follows:*
   (a) *If $(x, y) \in \mathcal{A}_R$ and $x, y \in S$ then either $+(x, y) \in R$ or $-(x, y) \in R$, but not both.*
   (b) *If $(x, y) \in \mathcal{A}$ with $y \notin S$ then exactly one of the following must be in R:*
   either $+(x \twoheadrightarrow y)$
   or $-(x \twoheadrightarrow y)$
   or $+(x \mathrel{\to\!\!\!\!\twoheadrightarrow} y)$
   or $-(x \mathrel{\to\!\!\!\!\twoheadrightarrow} y)$
5. *A reactive Kripke model has the form (S, I, R, a, e, h), where $S \neq \varnothing$ is the set of possible worlds, $I : S \mapsto (w^S - \varnothing)$ is the ideal world function, R is a reactive accessibility relation and $a \in S$ is the beginning world and $e \in S$ is the evaluation world. h is the assignment to the atoms. For each $q \in Q, h(q) \subseteq S$.*

Remark 9 1. The meaning of $x \twoheadrightarrow y$ is that as we pass through x the arc y is put in an off position if it is on. The meaning of $x \mathrel{\to\!\!\!\!\twoheadrightarrow} y$ is that as we pass through x the arc y is put in an on position if it is off.
2. An arc y is in off position in R if $-y \in R$. It is in an on position in R if $+y \in R$. The formal definition is Definition 4 below.

Example 3 Let us describe the model introduced in Figure 4.

$$\begin{aligned} S &= \{a, b, c, d\} \\ R &= \{+(a, c), +(a, b), -(b, c), +(c, d)\} \cup \\ &\quad \{+(a \twoheadrightarrow (c, d)), +((a, c) \mathrel{\to\!\!\!\!\twoheadrightarrow} (a, (c, d))), +((a, b) \mathrel{\to\!\!\!\!\twoheadrightarrow} (b, c)), +((b, d)) \mathrel{\to\!\!\!\!\twoheadrightarrow} (a, (c, d))\} \end{aligned}$$

I is not specified in the figure, neither is h or e.

Definition 4 (Movement in a Reactive Model). *Let (S, R, e) be part of a model. Assume $+(e, e') \in R$. We explain what it means to move along the arc (e, e') from (S, R, e) to $(S, R_{(e,e')}, e')$.*

$R_{(e,e')}$ is obtained from R by executing the following actions:

1. *For every $+(e \twoheadrightarrow y) \in R$ such that $+y \in R$, replace $+y$ by $-y$.*
2. *For every $+(e \mathrel{\twoheadrightarrow\mkern-12mu\twoheadrightarrow} y) \in R$ such that $-y \in R$, replace $-y$ by $+y$.*
3. *For every $+((e,e') \twoheadrightarrow y) \in R$ such that $+y \in R$ replace $+y$ by $-y$.*
4. *For every $+((e,e') \mathrel{\twoheadrightarrow\mkern-12mu\twoheadrightarrow} y) \in R$ such that $-y \in R$ replace $-y$ by $+y$.*
5. *$R_{(e,e')}$ is the set obtained from R by doing exactly the above actions.*

Define R_a as the set obtained from R by executing actions (1) and (2) only.

Definition 5 (Reachability). *Let (S,R,a) be a part of a model. Let $x \in S$. We define the notion of 1x is reachable from a in (s,R,a)' by induction:*

1. *If $+(a,x) \in R$ then x is reachable in one step from a in (S,R,a).*
2. *x is reachable in $n+1$ steps in (S,R,a) if for some $a' \in S$ $+(a,a') \in R$ and x is reachable in n steps from a' in $(S,R_{(a,a')},a')$.*
3. *x is reachable from a in (S,R,a) if for some n, x is reachable from a in (S,R,a) in n steps.*

Definition 6 (Contrary to Duties). *Let (S,I,R,a,e) be part of a model. We now define the contrary to duties suggested by this model relative to a.*

The ideal worlds are $I(a)$, and assume that none of $I(a)$ is reachable from a in (S,R_a,a)

Let a' be such that$+(a,a') \in R$. Let $CTD_{(a,a')}$ be the points reachable from a' in $(S,R_{(a,a')},a')$. Let $\mathrm{CTD}_a = \bigcap_{+(a,a')\in R} \mathrm{CTD}_{(a,a')}$. Then CTD_a are the contrary to duty worlds relative to a. In words:

> *It is obligatory to go to $I(a)$ but if not go to CTD_a.*

Remark 10. The double and triple arrows emanating from a (i.e. $+(a \twoheadrightarrow y)) \in R$ or $+(a \mathrel{\twoheadrightarrow\mkern-12mu\twoheadrightarrow} y) \in R$) can be seen to indicate the intention of the agent, or the restrictions on the user imposed by the model (if we do not want to ascribe intentions to users).

If activated the agent will move to a model with R_a. If $I(a)$ world are no longer accessible then our agent is not able to execute his obligations at a. The contrary to duties are the adjustments (activation and cancellation of arcs) firing as the agent passes through to any a' such that $+(a,a') \in R$. The CTD_a are the worlds which are always accessible from any a' the agent goes to. These are the contrary to duty worlds. Note that $I(a')$ are the ideal worlds of a'. This is not the same as the contrary to duties at a as imposed from node a onto node a'.

The reader may ask what if at a' some $x \in I(a)$ is still reachable? The definition of CTD_a still works. What is the meaning of it? The answer is in the next example. We call these *preventive* CTD, PCTD.

Example 4. At a world where you have a fence, it is obligatory to point the fence white within seven days. Say after five days nothing has been done. There are two options. Do nothing and the fence will not be painted. Hire extra hands and the fence will be painted. A preventive contrary to duty is to put pressure on the agent by blocking some of his moves. See Figure 14.

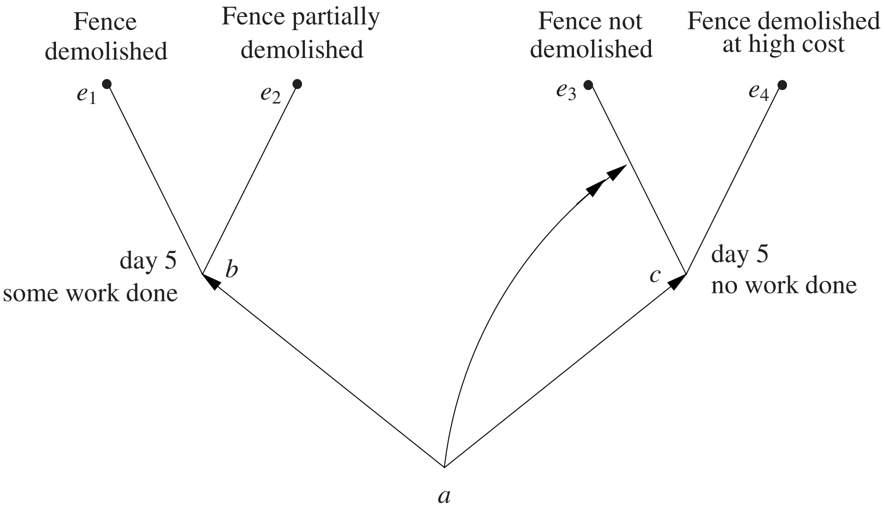

Fig. 14.

It is not correct to say that a CTD is that if after five days no work is done then the agent has a duty to bring extra workers to do the job. His duty remains to demolish the fence within seven days and he can still do it.

So the double arrow $(+((a, c), (e, e_3))$ is a preventive measure, a PCTD.

Definition 7 (Evaluation of Modalities in a Model). *Consider a model* (S, R, I, a, e, h). *We define the notion of* $e \vDash A$ *for a wff* A.

1. $e \vDash q$, *for* q *atomic, if* $e \in h(q)$.
2. *We adopt the usual definition for the classical connectives.*
3. $e \vDash OA$ *iff for all* $x \in I(e)$, $x \vDash A$ *in the model* (S, R, I, a, x, h).
4. $e \vDash \Diamond\kern-0.6em|\kern0.35em A$ *iff for some* x *such that* $+(e, x) \in R$, $x \vDash A$ *in* (S, R, I, a, x, h).
5. $e \vDash \Diamond A$ *iff for some* x *such that* $+(e, x) \in R$, $x \vDash A$ *in* $(S, R_{(e,x)}, I, a, x, h)$.

Definition 8 (Suggested CTDs). *A CTD multimodel is a family* $\mathcal{M}$ *of of models of the form* $\mathcal{M}_i = (S, I, R_i, a, e, h)$ *where all* (S, I, a, e, h) *are the same for all models and only* R_i *change. The CTD rules suggested by the family are all the* CTD^i_a *suggested by each model* M_i.

Remark 11. Note that for formulas without modalities, we have $x \vDash_i A$ iff $x \vDash_j A$ for any i, j, since they all agree on (S, I, h). Write $x \vDash A$ if A holds in any i. So for such formulas we can extract a syntactical CTD. Let $\varDelta_a = \{A \mid x \vDash A \text{ for all } x \in I(a)\}$.

Let $\varTheta^i_a = \{B \mid y \vDash B \text{ for all } y \in \mathrm{CTD}^i_a\}$.

Then our syntactical CTDs suggested by M_i are $\varDelta_a$ and if not then $\varTheta^i_a$.

Example 5 (The Reykjavic Paradox). We show how to handle this paradox, see [14]. we have

1. X should not tell the secret to Reagan.
2. X should not tell the secret to Gorbachev.

3. If X tells Reagan, then X should tell Gorbachev.
4. If X tells Gorbachev, then X should tell Reagan.
5. X tells Reagan and Gorbachev.

(1)–(5) is the paradox. It is easy to model it in our system.

So, just to add to the problem, let me add (6) as a challenge

6. If X insists on telling exactly one of them, then it should be Reagan.

Figure 15 is a model for the above.

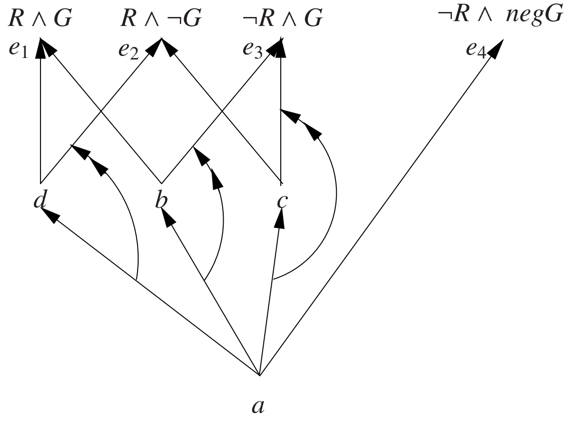

Fig. 15.

Point e_4 is the ideal world for a, $e_4 \in I(a)$. This models (1) and (2). Point d forces telling Reagan. Point b forces telling Gorbachev. Point c forces telling exactly one of them.

The double arrows are the contrary to duties. They model (3), (4) and (6). The arc to point d for example, means X is going to point d which forces telling Reagan. We want him to be under CTD to tell Gorbachev. So we disconnect the arc (d, e_2). So we need the double arrow $+((a, d) \twoheadrightarrow (d, e_2))$. Similarly, we need $+((a, b) \twoheadrightarrow (b, e_3))$ and also $+((a, c) \twoheadrightarrow (c, e_3))$.

The beginning point is a, the evaluation point is e_1 which models (5).

The reader may ask how did we construct the model? Well, there are some heuristics.

Remark 12 (Heuristics for Building a Reactive Model in Cases where there is no Temporal Element Involved). Let $A_1, \ldots, A_n$ be obligations. Let If $\neg A_i$ then $B_i, i = 1, \ldots, n$ be CTDs.

First let $\Theta = \{X_1, \ldots, x_{2^{2n}}\}$ be the set of all Boolean combinations of $\{A_1, \ldots, A_n, B_1, \ldots, B_n\}$.

The contrary to duties say that if $\neg A_i$ then B_i. When we move along the arc $+(a, t_i)$, we are getting to a point where $\neg A_i$ is committed. So we must force B_i. So any t_j such that $X_j \vdash \neg A_i \wedge \neg B_i$ must be disconnected. So we include in R all double arrows of the form $+((a, t_i) \twoheadrightarrow (t_i, e_j))$ such that $X_j \vdash \neg A_i \wedge \neg B_i$.

The ideal worlds for a are all e_j such that $X_j \vdash \bigwedge_{i=1}^{n} A_i$.

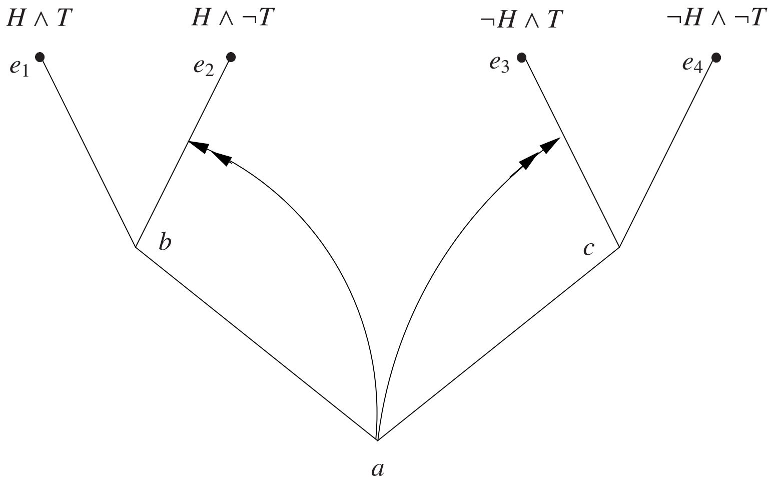

Fig. 16.

The above construction is a Henkin-like type of construction. It works when there is no temporal element.

Example 6 (Chisholm Paradox Revisted). Let us try and model the Chisholm paradox, using the Henkin-like method as in the previous example 5. We get Figure 16

Figure 16 does the job but it does not take into account the temporal aspect of the problem. See also [10] and [19].

Evaluation point is $x \in \{e_3, e_4\}$, $x = e_4$ if the agent complies with the CTD and $x = e_3$ if he does not.

This solution is more appropriate if instead of 'Tell' we have 'Wear his overalls'. So

(d1) It ought to be that a certain man go to help his neighbour.
(d2) It ought to be that if he goes he wear his overalls.
(d3) If he does not go he ought not wear his overalls.
(d4) He does not go.

We need to develop special methods to deal with the temporal aspects of CTD.

We can improve the situation in the Chisholm example by reading 'T' as 'T''

T' = having told in the past.

This would help. However, we do not get the best model. We do need to develop a general theory for time dependence. Again I ask the reader to wait for the full paper.

Let us stop here. Full analysis in the expanded full version of the paper possibly with more co-authors.

Acknowledgements

I am grateful to L. van der Torre and O. Rodrigues for valuable comments.

References

1. Gabbay, D.M.: Reactive Kripke Semantics and Arc Accessibility. In: Avron, A., Dershowitz, N., Rabinovich, A. (eds.) Pillars of Computer Science: Essays Dedicated to Boris (Boaz) Trakhtenbrot on the Occasion of His 85th Birthday. LNCS, vol. 4800, pp. 292–341. Springer, Berlin (2008); Earlier version published. In: Carnielli, W., Dionesio, F. M., Mateus, P. (eds.) Proceeding of CombLog04, Centre of Logic and Computation University of Lisbon, pp. 7–20 (2004), http://www.cs.math.ist.utl.pt/comblog04/ ftp://logica.cle.unicamp.br/pub/e-prints/comblog04/gabbay.pdf
2. Gabbay, D.M., Barringer, H., Rydeheard, D.: Reactive Grammars. Draft
3. Gabbay, D.M., Crochemore, M.: Reactive Automata. Draft
4. Gabbay, D.M., D'Agostino, M.: Reactive Conditionals. Draft
5. Prakken, H., Sergot, M.J.: Contrary-to-duty obligations. Studia Logica 57(1), 91–115 (1996)
6. Carmo, J., Jones, A.J.I.: Deontic Logic and Contrary-to-Duties. In: Gabbay, D.M., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 8, pp. 265–343. Springer, Heidelberg (2002)
7. Gabbay, D.M.: Reactive Proof Theory. Draft
8. Gabbay, D.M., Rodrigues, O., Woods, J.: Belief Contraction, Anti-formulas, and Resource Overdraft: Part I. Logic Journal of the IGPL 10, 601–652 (2002)
9. Gabbay, D.M., Rodrigues, O., Woods, J.: Belief Contraction, Anti-formulae and Resource Overdraft: Part II. In: Gabbay, D.M., Rahman, S., Symons, J., van Bendegem, J.-P. (eds.) Logic, Epistemology and the Unity of Science, pp. 291–326. Kluwer, Dordrecht (2004)
10. Gabbay, D.M., Reyle, U.: N-Prolog: An Extension of Prolog with Hypothetical Implications I. Journal of Logic Programming 1, 319–355 (1984)
11. Makinson, D., van der Torre, L.: Constraints for input-output logics. Journal of Philosophical Logic 30(2), 155–185 (2001)
12. van der Torre, L.W.N., Tan, Y.-H.: The Temporal Analysis of Chisholm's Paradox. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence and the Ninth Innovative Applications of Artificial Intelligence Conference (1998)
13. Hansson, B.: Standard Dyadic Denotic Logic. Noûs 3, 373–398 (1969)
14. van der Torre, L.: Violated obligations in a defeasible deontic logic. In: Proceedings of ECAI 1994, Amsterdam, pp. 371–375 (1994)
15. Makinson, D.: Five faces of minimality. Studia Logica 52, 339–379 (1993)
16. Broersen, J., van der Torre, L.: Reasoning About Norms, Obligations, Time and Agents. In: Proceedings of PRIMA 2007. LNCS. Springer, Heidelberg (to appear)
17. Broersen, J.: Modal Action Logics for Reasoning about Reactive Systems, Jan Broersen, PhD-thesis Vrije Universiteit Amsterdam (January 2003)
18. Hansen, J., Pigozzi, G., van der Torre, L.: Ten Philosophical Problems in Deontic Logic. In: Boella, G., van der Torre, L., Verhagen, H. (eds.) Normative Multi-agent Systems, Dagstuhl Seminar Proceedings, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany (2007)
19. Broersen, J., van der Torre, L.W.N.: Semantic Analysis of Chisholm's Paradox. In: BNAIC 2005, pp. 28–34 (2005)