



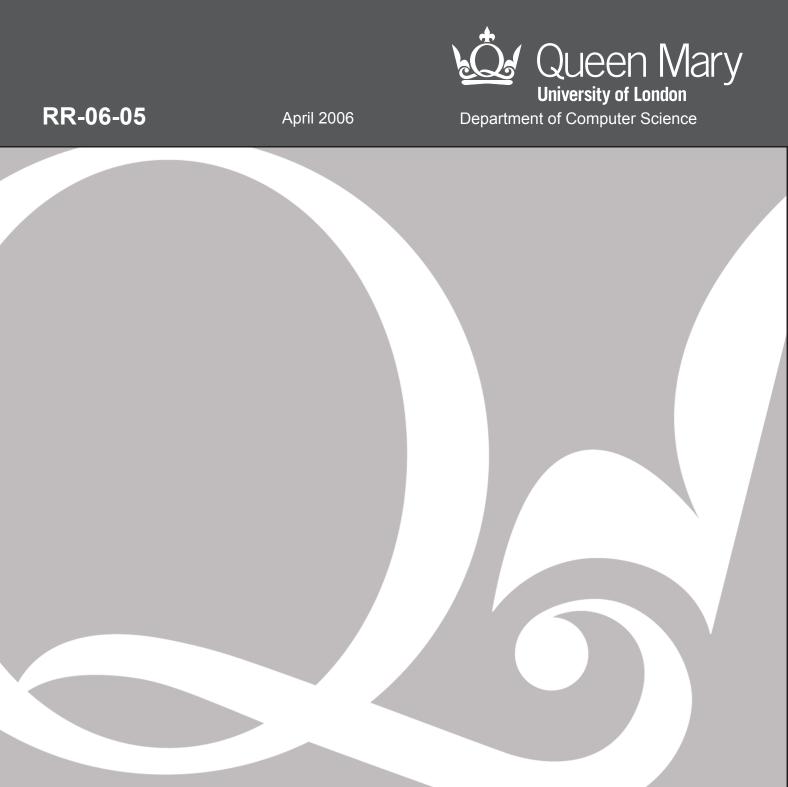
McCain-Turner Theories: Alternative Causal Laws White, Graham

For additional information about this publication click this link. http://qmro.qmul.ac.uk/jspui/handle/123456789/5057

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk

Graham White

Laws



# McCain-Turner Theories: Alternative Causal Laws

Dr. G. White Department of Computer Science Queen Mary University of London London E1 4NS, UK graham@dcs.qmul.ac.uk http://www.dcs.qmul.ac.uk

April 29, 2006

# Contents

Background	1
1.1 The McCain-Turner Theory and its Modal Formulation	1
1.2 The Meaning of the System	3
1.2.1 Terminology $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$	4
Some Laws	4
2.1 The Standard Formulation	4
2.2 Time Forwards	5
2.2.1 The Meaning of These Rules	5
Automata	6
3.1 Automata and Fluents	8
Conclusions: Further Work	10
	1.1 The McCain-Turner Theory and its Modal Formulation   1.2 The Meaning of the System   1.2.1 Terminology   1.2.1 Terminology   Some Laws   2.1 The Standard Formulation   2.2 Time Forwards   2.2.1 The Meaning of These Rules   Automata   3.1 Automata and Fluents

## 1 Background

### 1.1 The McCain-Turner Theory and its Modal Formulation

We recall the theory of [9]; this is a system of predicate modal logic in which we can represent McCain and Turner's "causal reasoning" [6]. In this system, what McCain and Turner describe as a "causal law", and write  $\phi \to \psi$ , will be represented as a modal axiom

$$\phi \vdash \Box \psi \tag{1}$$

Given a set, D, of such causal laws, the system in Table 1 gives a sequent calculus which is sound for these modal axioms. McCain and Turner described their semantics as follows:

Table 1: Sequent Calculus Rules

Ax	— <i>L</i> ,	
$\overline{A \vdash A}$ Ax	$ \_\_ L \bot $	
$\frac{\Gamma\vdash\Delta}{\Gamma,A\vdash\Delta}LW$	$\frac{\Gamma\vdash\Delta}{\Gamma\vdash A,\Delta}RW$	
$\Gamma, A \vdash \Delta$	$\Gamma \vdash A, \Delta$	
$\Gamma, A, A \vdash \Delta$	$\Gamma \vdash A, A, \Delta$	
$\frac{\Gamma, A, A \vdash \Delta}{\Gamma, A \vdash \Delta} LC$	$\frac{\Gamma \vdash A, A, \Delta}{\Gamma \vdash A, \Delta} RC$	
$\frac{\Gamma, A, B \vdash \Delta}{\Gamma, A \land B \vdash \Delta} \land L$	$\Gamma \vdash A, \Delta  \Gamma \vdash B, \Delta$	
$\overline{\Gamma, A \land B \vdash \Delta} \land L$	$\frac{\Gamma \vdash A, \Delta  \Gamma \vdash B, \Delta}{\Gamma \vdash A \land B, \Delta} \land R$	
$\frac{\Gamma, A \vdash \Delta  \Gamma, B \vdash \Delta}{\Gamma, A \lor B \vdash \Delta} \lor L$	$\frac{\Gamma \vdash A, B\Delta}{\Gamma \vdash A \lor B, \Delta} \lor R$	
$\overline{\Gamma, A \lor B \vdash \Delta} \lor^L$	$\overline{\Gamma \vdash A \lor B, \Delta} \lor H$	
$\Gamma \vdash A, \Delta  \Gamma, B \vdash \Delta$	$\Gamma, A \vdash B, \Delta$	
$\frac{\Gamma \vdash A, \Delta  \Gamma, B \vdash \Delta}{\Gamma, A \to B \vdash \Delta} \to L$	$\frac{\Gamma, A \vdash B, \Delta}{\Gamma \vdash A \to B, \Delta} \to R$	
$\Gamma, A[x/t] \vdash \Delta$	$\Gamma \vdash A[x/y], \Delta \cup D^a$	
$\frac{\Gamma, A[x/t] \vdash \Delta}{\Gamma, \forall xA \vdash \Delta}  \forall L$	$\frac{\Gamma \vdash A[x/y], \Delta}{\Gamma \vdash \forall xA, \Delta}  \forall R^a$	
$\frac{\Gamma, A[x/y] \vdash \Delta}{\exists L^b}$	$\frac{\Gamma \vdash A[x/t], \Delta}{\Gamma \vdash \exists xA, \Delta} \exists R$	
$\Gamma, \exists x A \vdash \Delta$	$\frac{1}{\Gamma \vdash \exists xA, \Delta} \exists R$	
$\frac{\Gamma \vdash P_1 \land \ldots \land P_n, \Delta}{}$	$Q_1, \ldots, Q_n \vdash X$	
${\Gamma \vdash \Box X, \Delta} \Box R^{\circ}$		
$\frac{\{\Gamma, P_{i_1}, \dots, P_{i_k} \vdash A, \Delta,  Q_{i_1}, \dots, Q_{i_k} \vdash X\}_{i=1,\dots,n}}{\sqcap L^d}$		
$\Box \Gamma, \Box X \vdash \Box A, \Diamond \Delta$		
$\Gamma \vdash X^m, \Delta  \Gamma', X^n \vdash \Delta'$		

 $\begin{array}{c} \stackrel{a}{ y \text{ not free in } \Gamma \text{ or } \Delta, \text{ and either } y = x \text{ or } y \text{ not free in } A \\ \stackrel{b}{ y \text{ not free in } \Gamma \text{ or } \Delta, \text{ and either } y = x \text{ or } y \text{ not free in } A \\ \stackrel{c}{ where, \text{ for all } i, P_i \triangleright Q_i \\ \stackrel{d}{ where, \text{ for each } i, \text{ we have } P_{i_1} \triangleright Q_{i_1}, \ldots P_{i_k} \triangleright Q_{i_k}, \text{ and where the } \{P_{i_j}\} \text{ and } \{Q_{i_j}\}, \text{ for } i = 1, \ldots n, \text{ are the only such sets of } P_{\text{s and }}Q_{\text{s that there are.}} \\ \stackrel{e}{ where } X^n \text{ stands for } n \text{ occurrences of } X; m, n > 0 \end{array}$ 

1. Suppose that we are given a set of causal laws: call it  $\mathbb{T}$ , formulated in some language  $\mathfrak{L}$ . Let M be a model of  $\mathfrak{L}$ . Define

$$\mathbb{T}^{M} \stackrel{\text{def}}{=} \{\psi | \text{ for some } \phi \in \mathfrak{L}, \phi \triangleright \psi \text{ and } M \vDash \phi\}$$
(2)

- 2. Now we say that M is causally explained (according to  $\mathbb{T}$ ) if it is the only model of  $\mathbb{T}^M$ .
- 3. Finally, we say that  $\phi \in \mathfrak{L}$  is a *consequence* of a causal theory  $\mathbb{T}$  if  $\phi$  is true in every  $\mathbb{T}$ -causally explained model of S.

We can express this in our system as follows:

**Proposition 1.** Given a set  $\mathbb{T}$  of causal laws, consider the corresponding modal system defined by the rules in Table 1. Then M is an explained model iff, for any proposition P,

$$\mathbb{M} \vdash (\Box P \to P) \land (P \to \Box P)$$

where  $\mathbb{M}$  is the theory of the model M.

*Proof.* A straightforward application of the cut elimination result of [9]; we first prove that, if M is a model of  $\mathfrak{L}$ ,

$$\overline{(\mathbb{T}^M)} = \{X | \mathbb{M} \vdash \Box X\}$$

where  $\overline{(\cdot)}$  represents deductive closure.

We can then show that M is an explained model if, when regarded as a world in the canonical Kripke model of the modal logic, it is accessible to itself and to no other worlds; we can also show that the world of the canonical Kripke model are precisely the models of the non-modal language; but this – together with standard correspondence theory results [8] – gives us the proposition.  $\Box$ 

**Corollary 2.** Given a set  $\mathbb{T}$  of explanations, consider the corresponding modal system defined by the rules in Table 1. Then, if A is a non-modal proposition, A is causally explained according to  $\mathbb{T}$  iff there are non-modal propositions  $P_1, \ldots, P_k, Q_1, \ldots, Q_l$  such that, in the modal system,

$$P_1 \to \Box P_1, \ldots, P_k \to \Box P_k, \Box Q_1 \to Q_1, \ldots, \Box Q_l \to Q_l \vdash A.$$

Proof. An easy application of compactness.

### 1.2 The Meaning of the System

This system is quite generic (as is McCain and Turner's formulation): it is welldefined for any collection of "causal laws" at all. However, McCain and Turner also define, associated to each instance of the frame problem, a *particular* set of "causal laws": the goal of this paper is to investigate other sets of laws, and to see what other aspects of the frame problem – or, indeed, of other problems in AI – can be captured thereby.

Once we do this, we begin to realise that this system is not limited solely to causal reasoning: that it is quite a general logic of explanation. A law, such as (1), can be regarded as an explanation pattern: an explanation pattern in which  $\psi$  will be the *explanandum* and  $\psi$  the *explanans*. Now with McCain and Turner's

laws,  $\psi$  was a a fluent, and  $\phi$  was either the action and action precondition which led to that fluent, or it stated that the fluent arose by persistence from the previous instant; laws like these, then, describe causal explanations, and they can thus rightly be regarded as causal.

But in general, there is no need to restrict ourselves to explanations merely of that sort; indeed, Lifschitz hinted as much in [5]. Now explanations can be viewed as answers to questions (and specifically to why-questions) [7]; there are, however, other questions that we could well think of asking, and, consequently, there are other sorts of explanation that we could usefully apply this logic to. Specifically, one of the most interesting applications will be to the question, not of what caused a particular fluent, but of what that fluent gave rise to: the question of what happened next, rather than the question of what went before.

#### 1.2.1 Terminology

There is, unfortunately, the question of terminology and notation to consider. Terminology of "causal laws" and "causal theory" is, by now, quite well entrenched (and I have also been quite careful to use it in quotation marks, so as not to prejudice the issue). In order to keep as close as possible to established terminology, I will tend to use terms like 'explanatory law' and 'explanatory theory', or often just neutrally 'law'. When I am feeling non-traditional I may also refer to a law as an explanation.

So much for terminology. We will also need a metatheoretical notation for laws (distinct from their representation in the theory as modal entailments such as (1)): I shall use

$$\phi \triangleright \psi. \tag{3}$$

This is, it is true, different from McCain and Turner's notation (that is,  $\cdot \Rightarrow \cdot$ ): however,  $\Rightarrow$  seems to have too many entailment-like connotations to be entirely convincing as a notation for explanation.

# 2 Some Laws

#### 2.1 The Standard Formulation

Suppose that we have fluents f, actions a, and possibly domain constraints C, all formulated in our usual language  $\mathfrak{L}$ . McCain and Turner's laws are of this form:

- 1.  $f_0 \Rightarrow \Box f_0$  and  $\neg f_0 \Rightarrow \Box \neg f_0$ , for any fluent f at time 0;
- 2.  $a_t \Rightarrow \Box a_t$  and  $\neg a_t \Rightarrow \Box \neg a_t$ , for any action a at any time t;
- 3.  $f_{t-1} \wedge f_t \Rightarrow \Box f_t$  and  $\neg f_{t-1} \wedge \neg f_t \Rightarrow \Box \neg f_t$ , for any fluent f and any time t;
- 4.  $f_{t-1} \wedge a_{t-1} \Rightarrow \Box g_t$ , for any time t, where f is the precondition and g is the postcondition of action a.
- 5.  $\neg P \Rightarrow \perp$ , for any domain constraint *P*.

Now interesting uses of these laws will be right rules of the form

$$\frac{\Gamma \vdash f_{t-1} \land f_t, \Delta}{\Gamma \vdash \Box f_t, \Delta} \Box \mathbf{R}$$

or of the form

$$\frac{\Gamma \vdash f_{t-1} \land a_{t-1}, \Delta}{\Gamma \vdash \Box f_t, \Delta} \Box \mathbf{R}$$

or, of course, the corresponding left rules: and, in any of these cases, application of the sequent calculus rules works backwards in time, from t to t - 1. Thus, both proof search, and the resulting proofs, will look like temporal regression from a goal: this is, of course, a well established tradition in planning, and it is reassuring to see that the standard form of these theories leads to it.

#### 2.2 Time Forwards

It would, then, also be interesting to investigate a system which would give the opposite temporal direction to proof search: that is, in which we would start at time 0, and extend trajectories into the future. So we would consider laws of the form

- 1.  $f_0 \Rightarrow \Box f_0$  and  $\neg f_0 \Rightarrow \Box \neg f_0$ , for any fluent f at time 0;
- 2.  $a_t \Rightarrow \Box a_t$  and  $\neg a_t \Rightarrow \Box \neg a_t$ , for any action a at any time t;
- 3.  $f_{t-1} \wedge f_t \Rightarrow \Box f_{t-1}$  and  $\neg f_{t-1} \wedge \neg f_t \Rightarrow \Box \neg f_{t-1}$ , for any fluent f and any time t;
- 4.  $g_t \wedge a_{t-1} \Rightarrow \Box f_{t-1}$ , for any time t, where f is the precondition and g is the postcondition of action a.
- 5.  $\neg P \Rightarrow \perp$ , for any domain constraint P.

These are, of course, much the same as the previous set: the only rules which differ are the temporally directed ones, which simply exchange the temporal direction. And, technically, these work as expected: proof search amounts to extending trajectories from time 0 towards the future, and we can appropriately formulate problems of temporal progression using rules of this form.

#### 2.2.1 The Meaning of These Rules

The merely *technical* adequacy of these rules is, perhaps, no surprise: we are taking a calculus and reversing the direction of time in order to get the desired effect. But what does it *mean*?

As I indicated earlier, the causal reading of this system may not always be the most appropriate one. A more general reading would be to regard it as, simply, a logic of explanation: McCain and Turner's explanations would be causal, but these time-reversed laws would not have a causal reading. Even though we had reversed time, the direction of causality would still be the same: the past would still influence the future, and not the other way round. However, we could regard these laws as describing the logic of a different explanatory task: rather than asking for answers to questions like "what gave rise to this situation?", we would be looking for answers to the question "what came after this situation?". Indeed, the idea of questions and answers is quite appropriate here. According to Hintikka [3, 2], a question can be regarded as denoting its set of possible answers (out of which an appropriate answer selects one). Here we have two rules for  $\Box$ , left rules and right rules; when we apply a left rule to the necessitation of a given fluent, we get the set of possible answers to the appropriate question ("what went before?" or "what next?"), as appropriate. When we apply a right rule, we have to select an answer from the set of appropriate ones. The duality of left rules and right rules, then, corresponds to a duality of questions and answers.

# 3 Automata

We have a further modification to make to the form of the rules. We have been looking at temporally indexed, fluent-based, formalisms: the fundamental object of investigation is a temporally indexed sequence of conjunctions of fluents and actions. However, there are alternative approaches to the problem of state change: in particular, we have the theory of automata [1]. There is no reason why we should not also apply a formalism of this sort to the theory of automata: it has no inherent notion of fluents (and an automaton in itself has no notion of temporal indexing), and we can try to use our calculus to handle the theory of automata in its own terms.

We will start with a simple definition of automata:

**Definition 1.** An *automaton* will be a finite set  $\Sigma$  of states, a finite set A of actions, an initial state  $\tilde{\sigma} \in \Sigma$ , and, for each  $\sigma \in \Sigma$ ,

- 1. a set  $A_{\sigma} \subseteq A$ , which represents the actions possible at  $\sigma$ , and
- 2. a map  $\rho_{\sigma}: A_{\sigma} \to \Sigma$ , representing the next state.

Note that these automata are finite, and the actions are deterministic. Notice also that the same action can occur in more than one state: this is basic for the theory of automata. For the sake of simplicity, we do not mention termination: this, together with nondeterminism, could easily be accommodated, if necessary.

Suppose we have such an automaton: let us consider the following explanatory laws. The language – which will describe the set of execution traces of the automaton, rather than the automaton itself – will have propositions indexed by times: the propositions will be

- 1.  $a_t$ , for  $a \in A$ ,  $t \in \mathbb{N}$ , representing the occurrence of action a at time t;
- 2.  $\sigma_t$ , for  $\sigma \in \Sigma$ ,  $t \in \mathbb{N}$ , representing the state  $\sigma$  at time t.

We also allow ourself (by abuse of notation) to write, for  $\sigma$  a state and a an action,  $(\rho_{\sigma_t}(a_t))_{t+1}$ ; this will mean the t+1-subscripted proposition corresponding to whatever state it is that is the result of action a acting in state  $\sigma$ . It is definitely *not* the application of a higher-order function in the language: the constructs here are explicitly metatheoretical.

The laws will be

- 1.  $\sigma_0 \triangleright \sigma_0$  for any  $\sigma$ ,
- 2.  $\sigma_{t+1} \triangleright \sigma_t$  for any  $\sigma$  and any t,

- 3.  $a_t \triangleright a_t$  and  $\neg a_t \triangleright \neg a_t$ , for any a and any t, and
- 4.  $\sigma_t \wedge a_t \wedge (\rho_{\sigma_t}(a_t))_{t+1} \triangleright \sigma_t$ , for any  $\sigma$ , any  $a \in A_{\sigma}$ , and any t.

Finally, the domain constraints will be:

1. actions exclude each other (i.e. for two distinct actions a and a', and for any t,

$$a_t \to \neg a'_t \text{ and } a'_t \to \neg a_t$$

- 2. states exclude each other, and
- 3. actions can only occur in an appropriate state: that is, if a is an action,

$$a_t \to \bigvee_{a_t \in A_\sigma} \sigma_t$$

We have, then, the following:

**Proposition 3.** Let M be a model of this theory such that, for some state  $\sigma$ ,  $M \vDash \sigma_0$ . M is causally explained iff it makes true exactly one sequence of state propositions, starting with the initial state:  $\tilde{\sigma}_0, \sigma_1, \ldots$ , such that, for each i, either  $\sigma_i = \sigma_{i+1}$ , or there is an action  $a_i$  with  $M \vDash a_t$  and such that a, when executed in  $\sigma_i$ , yields  $\sigma_{i+1}$ .

First a lemma:

**Lemma 4.** Under the above assumptions, if  $\sigma$  is a state and t a time, then  $\mathbb{M} \vdash \Box \sigma_t$  iff either

- 1.  $M \vDash \sigma_t \wedge \sigma_{t+1}$ , or
- 2.  $M \models \sigma_t \land a_t \land \sigma'_{t+1}$

where  $a_t$ , acting in  $\sigma_t$ , yields  $\sigma'_{t+1}$ .

*Proof.* The left and right rules imply that

$$\Box \sigma_t \cong \sigma_t \land \left( \sigma_{t+1} \lor \bigvee_{a_t \in A_{\sigma_t}} a_t \land (\rho_{\sigma_t}(a_t))_{t+1} \right)$$

The domain constraints imply that at most one of the disjuncts on the right hand side can be true in a model: this gives the result.  $\Box$ 

Proof of Proposition 3. The only if direction is easy: we proceed by induction on t. We start with the  $\sigma_0$ , which is given by the assumption. Given that  $M \models \sigma_t$ , we know that, since M is explained,  $M \models \Box \sigma_t$ ; the lemma gives us a suitable  $\sigma_{t+1}$  and, if necessary, an action at t.

For the if direction, note that, if we have such a sequence of propositions and actions, their truth, together with the domain constraints, determines the truth value of every other atom. It is easy to see that actions are equivalent to their own necessitations; the lemma shows that a fluent is true in our model iff its necessitation is.  $\hfill \Box$ 

We thus have a McCain-Turner theory for which the explained models correspond to execution traces of our automaton.

#### 3.1 Automata and Fluents

We have, then, a logical theory of automata: the theory of automata has closely linked advantages and disadvantages. The advantages are that it provides a synoptic view of the behaviour of a system: we have a set of states, which are not inherently temporally indexed, together with transitions between them, and we can recover from this all of the possible temporal behaviours of the system.

However, these advantages are closely coupled to the disadvantages: it represents very much a God's eye view of the system. It is temporally synoptic (that is, specifying it involves knowing all of the possible states of the system), and the states themselves are unstructured: they are simply nodes of a graph, without internal structure, and we have to characterise what a state is by talking about what transitions are possible from it. We need, beyond the notion of an automaton as given here, some notion of what one can observe about a state; we could call this an *observable*, but one could also call it an *output*.

*Remark* 1. The idea that states themselves are inscrutable has a quite precise technical formulation in the theory of automata: equality between states is never defined, and we consider states (and, correspondingly, automata) as equivalent if they are *bisimilar* (that is, if their input-output behaviours are identical); see [4].

So, here is a definition of such an automaton:

**Definition 2.** An *automaton with fluent observables* is an automaton, as above, together with a set of fluent symbols F, and, for all states  $\sigma$ , a map

$$eval_{\sigma}: F \to \{\top, \bot\}.$$

We could accomodate this notion in our logic by treating the eval relation as entailment: we add axioms to the system so that, for all t, all  $\sigma$ , and all fluents  $f \in F$ ,

$$\sigma_t \vdash f_t$$
 iff  $eval_{\sigma}(f) = \top$ .

Although this would be a useful first step, it falls victim to the frame problem: it involves explicitly assigning truth values to each fluent in each situation. What we want is to be able to assign preconditions and postconditions to each action, to assign a complete evaluation to the initial state, and to work out the values of fluents at subsequent states by a combination of the effects of actions with inertia: this is somewhat reminiscent of Foo and Zhang [10], who do much the same thing but start with dynamic logic rather than the theory of automata.

**Definition 3.** An *automaton with fluent observables and inertia* is an automaton, together with a set of fluents, together with

- 1. an evaluation function for the initial state,  $eval_{\tilde{\sigma}}: F \to \{\top, \bot\}$
- 2. for each action a, a conjunction pre(a) of fluents and negations of fluents, which are its preconditions
- 3. for each action a, a conjunction post(a) of fluents and negations of fluents, its postconditions.

Given such an automaton, we can evaluate observables in all of its accessible states by starting with the initial state, and then assigning truth values to subsequent states by first assigning the fluents in the postconditions of the relevant action, and then assigning the other fluents by inertia. We must, however, stipulate that this process is consistent:

**Definition 4.** The *coherence condition*, for automata with fluent observables and inertia, stipulates that, if a state is reachable from the initial state by two different sequences of actions, then the assignments corresponding to each sequence of actions should be the same.

*Remark* 2. This is a sensible condition to put on these automata. Recall from Remark 1 that states do not have a good notion of identity: rather, two states are equivalent if their input-output behaviour is identical (in technical terms, if they are *bisimilar*). Now if an automaton fails the coherence condition, this means that allegedly identical states have different observable behaviours: that is, the results of the observations that we can make on them (by evaluating fluents) depend on the causal history by which we approach the states, rather than the states themselves. Such a failure of coherence can always be cured by constructing a new set of states, with several states in the new set corresponding to a single state in the old set: and the automaton constructed with the expanded set of states will have the same observable input-output behaviour as the old one. What matters will be the same.

So now, given an automaton with fluent observables and inertia, how do we adapt our theory so that it will deliver, not states, but fluents? As suggested above, we would like to do this by entailment: that is, we would like to set up entailments between states and fluents, so that a state entailed a fluent iff that that fluent was true in that state. We can do this by an explanatory theory of the following form: here (since the values of fluents are determined solely by states) we have fluents indexed by states. Intuitively,  $f_{\sigma}$  is true in a model iff f is true in state  $\sigma$ .

- 1.  $f_{\tilde{\sigma}} \triangleright f_{\tilde{\sigma}}$  if  $\operatorname{eval}_{\tilde{\sigma}} f = \top$ ,
- 2.  $\neg f_{\tilde{\sigma}} \triangleright \neg f_{\tilde{\sigma}}$  if  $eval_{\tilde{\sigma}} = \bot$
- 3.  $f_{\rho_{\sigma}(a)} \triangleright f_{\rho_{\sigma}(a)}$  for any state  $\sigma$  and any action  $a \in A_{\sigma}$  whenever  $\mathsf{post}_a \vdash f$ ,
- 4.  $\neg f_{\rho_{\sigma}(a)} \triangleright \neg f_{\rho_{\sigma}(a)}$  for any state  $\sigma$  and any action  $a \in A_{\sigma}$ , whenever  $\mathsf{post}_a \vdash \neg f$ ,
- 5.  $f_{\sigma} \wedge f_{\rho_{\sigma}(a)} \triangleright f_{\rho_{\sigma}(a)}$  for any state  $\sigma$  and any action  $a \in A_{\sigma}$ ,
- 6.  $\neg f_{\sigma} \land \neg f_{\rho_{\sigma}(a)} \triangleright \neg f_{\rho_{\sigma}(a)}$  for any state  $\sigma$  and any action  $a \in A_{\sigma}$ .

Now these actions assign the correct truth values to fluents in situations:

**Proposition 5.** Consider the language whose atoms are fluents indexed by situations, and use the above explanatory theory to axiomatise the modality: then any explained model assigns the correct truth values to the accessible states.

*Proof.* This is exactly similar to the proof of the correctness for normal McCain-Turner theories – that is, we can do it by a descending induction – except that the induction is over the graph of accessible states of the model, rather than over the time line.  $\Box$  So we now have two separate explanatory theories: one is a forward chaining theory, and describes the evolution of the automaton, whereas the other is a backward chaining theory and assigns truth values to fluents in states. How do we link the two together? That is, how do we find a theory which delivers sequences of temporally indexed fluents? The answer is straightforward: we need a further set of explanations, namely explanations which explain the value of  $f_t$  by referring to the state at t and the value of f in  $\sigma$ . We need, that is,

- 1.  $f_{\sigma} \wedge \sigma_t \triangleright f_t$
- 2.  $\neg f_{\sigma} \land \sigma_t \triangleright \neg f_t$

With these additional explanations, everything works as expected.

# 4 Conclusions: Further Work

We have shown that McCain-Turner style causal theories are considerably more adaptable than has been realised: we can use them for non-causal explanations, both forwards and backwards chaining, and we can also – as the last example makes clear – combine several different styles of explanation in the same theory.

However, there are still gaps to be filled. One of the most conspicuous is this: we have to handle time by an index set external to the language. We do not, that is, have time as a variable in the language, and we cannot, in the language, quantify over it. A proper treatment of quantification, however, would exceed the bounds of this paper.

### References

- [1] Michael A. Arbib. Theories of Abstract Automata. Prentice-Hall, 1969.
- [2] Jaakko Hintikka. Semantics for propositional attitudes. In Leonard Linsky, editor, *Reference and Modality*. Oxford University Press, Oxford, 1972.
- [3] Jaakko Hintikka. The Semantics of Questions and the Question of Semantics. North-Holland, Amsterdam, 1976.
- [4] Bart Jacobs and Jan Rutten. A tutorial on (co)algebras and (co)induction. EATCS Bulletin, 62:229–259, 1997.
- [5] Vladimir Lifschitz. On the logic of causal explanation. Artificial Intelligence, 96:451-465, 1997.
- [6] Norman McCain and Hudson Turner. Causal theories of action and change. In Proceedings of AAAI97, pages 460–465, Providence, RI, 1997. AAAI.
- [7] Joseph C. Pitt, editor. *Theories of Explanation*. Oxford, 1988.
- [8] Johan van Benthem. Correspondence theory. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic II*, volume 165 of *Synthese Library*, pages 167–247. Reidel, Dordrecht, first edition, 1984.

- [9] G. Graham White. A modal formulation of McCain and Turner's theory of causal reasoning. In Sergio Flesca, Sergio Greco, Nicola Leone, and Giovambattista Ianni, editors, *Logics in Artificial Intelligence: 8th European Conference, JELIA 2002*, number 2424 in Lecture Notes in Artificial Intelligence, pages 211–222. Springer, 2002.
- [10] Dongmo Zhang and Norman Foo. EPDL: A logic for causal reasoning. In Bernhard Nebel, editor, *IJCAI-01: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 131–137, San Francisco, CA, August 2001. International Joint Conferences on Artificial Intelligence, Inc. and The American Association for Artificial Intelligence, Morgan Kauffman.