Some examples formulated in a 'seeing to it that' logic: Illustrations, observations, problems

Marek Sergot

Department of Computing
Imperial College London
London SW7 2AZ, UK
m.sergot@imperial.ac.uk

March 2013

Abstract

The paper presents a series of small examples and discusses how they might be formulated in a 'seeing to it that' logic. The aim is to identify some of the strengths and weaknesses of this approach to the treatment of action. The examples have a very simple temporal structure. An element of indeterminism is introduced by uncertainty in the environment and by the actions of other agents. The formalism chosen combines a logic of agency with a transition-based account of action: the semantical framework is a labelled transition system extended with a component that picks out the contribution of a particular agent in a given transition. Although this is not a species of the *stit* logics associated with Nuel Belnap and colleagues, it does have many features in common. Most of the points that arise apply equally to *stit* logics. They are, in summary: whether explicit names for actions can be avoided, the need for weaker forms of responsibility or 'bringing it about' than are captured by *stit* and similar logics, some common patterns in which one agent's actions constrain or determine the actions of another, and some comments on the effects that level of detail, or 'granularity', of a representation can have on the properties we wish to examine.

1 Introduction

Logics of 'seeing to it that' or 'bringing it about that' have a long tradition in the analytical study of agency, ability, and action. The best known examples are perhaps the *stit* ('seeing to it that') family associated with Nuel Belnap and colleagues. (See e.g. (Belnap and Perloff, 1988; Horty and Belnap, 1995; Horty, 2001; Belnap et al., 2001) and some of the other papers in this volume.) Segerberg (1992) provides a summary of early work in this area, and Hilpinen (1997) an overview of the main semantical devices that have been used, in *stit* and other approaches. With some exceptions, notably (Pörn, 1977), the semantics is based on a branching-time structure of some kind.

In recent years logics of this kind have also been attracting attention in computer science. They have been seen as a potentially valuable tool in the formal modelling of agent interaction (human or artificial), in distributed computer systems and in the field of multi-agent systems. Works in this area have tended to be quite technical, focussing on various extensions, usually to the *stit* framework, or on connections to other formalisms used in computer science. There are however very few examples to my knowledge of any actual applications and so the usefulness of these formalisms in practice remains something of an open question. Forms of *stit* and Pörn's 'brings it about' have also been used as a kind of semi-formal device in representation languages for regulations and norms and in discussions of the logical form of normative and legal constructs.

In this paper I want to look at a series of simple examples and how they might be formulated in a *stit*-like logic. An element of indeterminism is introduced by the environment—in some examples it may be raining, in others a fragile object might or might not break when it falls—and by the actions of other agents. The aim is, first, to explore something of the expressive power of this framework. An important feature of *stit* is that actions themselves are never referred to explicitly. The semantics abstracts away these details. *stit* thereby sidesteps what remains one of the most contentious questions in the philosophy of action, which is the question of what is action itself. If a man raises his arm, the arm goes up. But what is the *action* of raising the arm? Opinions are divided on this point. In *stit*, actions are not referred to directly and do not have to be named. On the other hand, there is sometimes a price to be paid for this abstraction since it is difficult to do without names for actions in all circumstances. Some of the examples are intended to explore this question. Second, I want to comment on some common patterns that arise, particularly when one agent's actions constrain, or possibly even determine, the actions of another. Relying on informal readings of these patterns can be misleading. And third, I want to identify some of the limitations and inadequacies of the framework as a representational device. These concern the treatment of causality, and questions regarding the effects of granularity, or level of detail, of a representation. I am making no claims of completeness. The treatment of temporal features is rudimentary, I will not touch on topics such as voluntary, deliberative, intentional, purposeful action, and even in these simple examples there are many issues that will not be addressed.

I will not formulate the examples in any form of *stit*-logic exactly, but using a different formalism (Sergot, 2008a,b) that nevertheless has much in common. It combines a logic of 'brings it about' with a transition-based account of action: the semantical framework is a form of labelled transition system extended with an extra component that picks out the contribution—intentional, deliberative but perhaps also unwitting—of a particular agent in a given transition. Although the development was influenced by the constructions used in (Pörn, 1977), it turned out (unexpectedly) to have much greater similarity with *stit*. Indeed, as explained later, it can be seen as a special case of the deliberative *stit*, with a different informal reading and some additional features. Although some aspects of the representations will be specific to the use of my preferred formalism, nearly all the points I want to make will apply equally to *stit*-logics.

2 Syntax and semantics

2.1 Preliminaries: Transition systems

Transition systems A labelled transition system (LTS) is usually defined as a structure $\langle S, A, R\rangle$ where

- S is a (non-empty) set of *states*;
- A is a set of *transition labels*, also called *events*;
- R is a (non-empty) set of labelled *transitions*, $R \subseteq S \times A \times S$.

When (s, ε, s') is a transition in R, s is the initial state and s' is the resulting state, or end state, of the transition. ε is *executable* in a state s when there is a transition (s, ε, s') in R, and *non-deterministic* in s when there are transitions (s, ε, s') and (s, ε, s'') in R with $s' \neq s''$. A *path* or *run* of length m of the labelled transition system $\langle S, A, R\rangle$ is a sequence $s_0\, \varepsilon_0\, s_1\, \cdots\, s_{m-1}\, \varepsilon_{m-1}\, s_m \quad (m \geq 0)$ such that $(s_{i-1}, \varepsilon_{i-1}, s_i) \in R$ for $i \in 1 \ldots m$. Some authors prefer to deal with structures $\langle S, \{R_a\}_{a \in A}\rangle$ where each R_a is a binary relation on S.

It is helpful in what follows to take a slightly more general and abstract view of transition systems. A transition system is a structure $\langle S, R, \mathit{prev}, \mathit{post}\rangle$ where

- S and R are disjoint, non-empty sets of *states* and *transitions* respectively;
- *prev* and *post* are functions from R to S: $\mathit{prev}(\tau)$ denotes the initial state of a transition τ, and $\mathit{post}(\tau)$ its resulting state.

A *path* or *run* of length m of the transition system $\langle S, R, \mathit{prev}, \mathit{post}\rangle$ is a sequence $\tau_1\, \cdots\, \tau_{m-1}\, \tau_m \quad (m \geq 0)$ such that $\tau_i \in R$ for every $i \in 1 \ldots m$, and $\mathit{post}(\tau_i) = \mathit{prev}(\tau_{i+1})$ for every $i \in 1 \ldots m{-}1$.

Two-sorted language Given a labelled transition system, it is usual to define a language of propositional atoms or 'state variables' in order to express properties of states. We employ a *two-sorted* language. We have a set $\mathcal{P}_{\mathrm{f}}$ of propositional atoms for expressing properties of states, and a disjoint set $\mathcal{P}_{\mathrm{a}}$ of propositional atoms for expressing properties of transitions. Models are structures

$$\mathcal{M} = \langle S, R, \mathit{prev}, \mathit{post}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$$

where h^{f} is a valuation function for atomic propositions $\mathcal{P}_{\mathrm{f}}$ in states S and h^{a} is a valuation function for atomic propositions $\mathcal{P}_{\mathrm{a}}$ in transitions R.

Transition atoms are used to represent events and attributes of events, and properties of transitions as a whole. For example, atoms $x{:}\mathit{move}{=}l$ and $x{:}\mathit{move}{=}r$ might be used to represent that agent x moves in direction l and r, respectively. The atom $\mathit{falls}(\mathit{vase})$ might be used to represent transitions in which the object *vase* falls. Transition atoms are also used to express properties of a transition as whole: for instance, whether it is desirable or undesirable, timely or untimely, permitted or not permitted, and so on. So, for example, the formula

$$a{:}\mathit{lifts} \land \neg\, b{:}\mathit{lifts} \land c{:}\mathit{move}{=}l \land \neg\, d{:}\mathit{move}{=}l \land \mathit{falls}(\mathit{vase}) \land \mathit{trans}{=}\mathit{red}$$

might represent an event in which a lifts its end of the table and b does not while c moves in direction l, d does not move in direction l, and the vase falls. The atom *trans=red* might represent that this event is illegal (say), or undesirable, or not permitted.

When a transition satisfies a transition formula φ we say it is a transition of type φ. So, for example, all transitions of type $a{:}\mathit{lifts} \land \neg\, b{:}\mathit{lifts}$ are also transitions of type $a{:}\mathit{lifts}$, and also transitions of type $\neg\, b{:}\mathit{lifts}$.

Formulas We extend this two-sorted propositional language with (modal) operators for converting state formulas to transition formulas, and transition formulas to state formulas.

Formulas are *state formulas* and *transition formulas*. State formulas are:

$$F \ ::= \ \text{any atom } p \text{ of } \mathcal{P}_{\mathrm{f}} \mid \neg F \mid F \land F \mid \underset{\rightarrow}{\Box}\varphi$$

where φ is any transition formula. Transition formulas are

$$\varphi \ ::= \ \text{any atom } \alpha \text{ of } \mathcal{P}_{\mathrm{a}} \mid \neg\varphi \mid \varphi \land \varphi \mid 0{:}F \mid 1{:}F$$

where F is any state formula.

We have the usual truth-functional abbreviations. $\underset{\rightarrow}{\Diamond}$ is the dual of $\underset{\rightarrow}{\Box}$: $\underset{\rightarrow}{\Diamond}\varphi =_{\mathrm{def}} \neg\underset{\rightarrow}{\Box}\neg\varphi$.

Semantics Models are structures

$$\mathcal{M} = \langle S, R, \mathit{prev}, \mathit{post}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$$

where h^{f} and h^{a} are the valuation functions for state atoms and transition atoms respectively. Truth-functional connectives have the usual interpretations. The satisfaction definitions for the other operators are as follows, for any state formula F and any transition formula φ.

State formulas:

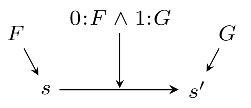
$$\mathcal{M}, s \models \underset{\rightarrow}{\Box}\varphi \quad \text{iff} \quad \mathcal{M}, \tau \models \varphi \text{ for every } \tau \in R \text{ such that } \mathit{prev}(\tau) = s$$

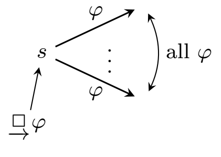
$\underset{\rightarrow}{\Box}\varphi$ is true at a state s when every transition from state s satisfies φ. $\underset{\rightarrow}{\Diamond}\varphi$ says that there is a transition of type φ from the current state.

Transition formulas:

$$\begin{aligned} \mathcal{M}, \tau \models 0{:}F \quad &\text{iff} \quad \mathcal{M}, \mathit{prev}(\tau) \models F \\ \mathcal{M}, \tau \models 1{:}F \quad &\text{iff} \quad \mathcal{M}, \mathit{post}(\tau) \models F \end{aligned}$$

A transition is of type $0{:}F$ when its initial state satisfies the state formula F, and of type $1{:}F$ when its resulting state satisfies F.





As usual, we say a state formula F is *valid* in a model $\mathcal{M}$, written $\mathcal{M} \models F$, when $\mathcal{M}, s \models F$ for every state s in S, and a transition formula φ is *valid* in a model $\mathcal{M}$, written $\mathcal{M} \models \varphi$, when $\mathcal{M}, \tau \models \varphi$ for every transition τ in R. A formula is *valid* if it is valid in every model (written $\models F$ and $\models \varphi$, respectively).

We use the following notation for 'truth sets':

$$\|F\|^{\mathcal{M}} =_{\text{def}} \{s \in S \mid \mathcal{M}, s \models F\}; \quad \|\varphi\|^{\mathcal{M}} =_{\text{def}} \{\tau \in R \mid \mathcal{M}, \tau \models \varphi\}.$$

$\mathcal{M}$ is omitted when it is obvious from context.

Examples: transition formulas The following represents a transition from a state where (state atom) p holds to a state where it does not:

$$0{:}p \land 1{:}\neg p$$

von Wright (1963) uses the notation $p\,\mathrm{T}\,q$ to represent a transition from a state where p holds to one where q holds. It would be expressed here in the more general notation as the transition formula:

$$0{:}p \land 1{:}q$$

Let the state atom $\mathit{on\text{-}table}(\mathit{vase})$ represent that a certain vase is on the table. A transition of type $0{:}\mathit{on\text{-}table}(\mathit{vase}) \land 1{:}\neg\,\mathit{on\text{-}table}(\mathit{vase})$, equivalently, of type $0{:}\mathit{on\text{-}table}(\mathit{vase}) \land \neg 1{:}\mathit{on\text{-}table}(\mathit{vase})$ is one from a state in which the vase is on the table to one in which it is not on the table. Let the transition atom $\mathit{falls}(\mathit{vase})$ represent the falling of the vase from the table. Any model $\mathcal{M}$ modelling this system will have the property:

$$\mathcal{M} \models\ \mathit{falls}(\mathit{vase}) \rightarrow (0{:}\mathit{on\text{-}table}(\mathit{vase}) \land 1{:}\neg\,\mathit{on\text{-}table}(\mathit{vase}))$$

There may be other ways that the vase can get from the table to the ground. Some agent might move it, for example. That would also be a transition of type $0{:}\mathit{on\text{-}table}(\mathit{vase}) \land 1{:}\neg\,\mathit{on\text{-}table}(\mathit{vase})$ but not a transition of type $\mathit{falls}(\mathit{vase})$.

The operators 0: and 1: are not normal in the usual sense because formulas F and $0{:}F$ (and $1{:}F$) are of different sorts. However, they behave like normal operators in the sense that, for all $n \geq 0$, if $F_1 \land \cdots \land F_n \rightarrow F$ is valid then so are $0{:}F_1 \land \cdots \land 0{:}F_n \rightarrow 0{:}F$ and $1{:}F_1 \land \cdots \land 1{:}F_n \rightarrow 1{:}F$. Since *prev* and *post* are (total) functions on R, we have

$$\models\ 0{:}F \leftrightarrow \neg 0{:}\neg F \qquad \text{and} \qquad \models\ 1{:}F \leftrightarrow \neg 1{:}\neg F$$

(and 0: and 1: distribute over all truth-functional connectives).

Examples: state formulas $\underset{\rightarrow}{\Diamond}\varphi$ says that there is a transition of type φ from the current state, or in the terminology of transition systems, that φ is 'executable'. $\underset{\rightarrow}{\Diamond}1{:}F$ expresses that there is a transition from the current state to a state where F is true. $\underset{\rightarrow}{\Box}(\varphi \rightarrow 1{:}F)$ says that all transitions of type φ from the current state result in a state where F is true.

There are various relationships between state formulas and transition formulas. For example, the state formula $F \rightarrow \underset{\rightarrow}{\Box}0{:}F$ is valid (true in all states, in all models). Further details are given in the next section.

2.2 Agency modalities

We now extend the language with operators to talk about the actions of agents and sets of agents in a transition. Ag is a finite set of (names of) agents. The account can be generalised to deal with (countably) infinite sets of agents but we will not do so here.

Language Transition formulas are extended with the operators $\Box$, $[x]$ and $[G]$ for every agent x in Ag and every non-empty subset G of Ag. State formulas are unchanged. $\Box\varphi$, $[x]\varphi$ and $[G]\varphi$ are transition formulas when φ is a transition formula. $\Diamond$, $\langle x\rangle$ and $\langle G\rangle$ are the respective duals.

Semantics Models are relational structures of the form

$$\langle S, R, \mathit{prev}, \mathit{post}, \sim, \{\sim_x\}_{x \in Ag}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$$

where $\langle S, R, \mathit{prev}, \mathit{post}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$ is a labelled transition model of the type discussed above, and $\sim$ and every $\sim_x$ are equivalence relations on R.

$$\sim \;=_{\mathrm{def}}\; \{\, (\tau, \tau') \mid \mathit{prev}(\tau) = \mathit{prev}(\tau') \,\}$$

and, for every $x \in Ag$: $\sim_x \subseteq \sim$.

Informally, for any transitions τ, τ' in R, $\tau \sim \tau'$ represents that τ and τ' are transitions from the same initial state, and $\tau \sim_x \tau'$ that τ and τ' are transitions from the same initial state ($\sim_x \subseteq \sim$) in which agent x performs the same action in τ' as it does in τ.

The truth conditions are

$$\begin{array}{lll} \mathcal{M}, \tau \models \Box\varphi & \text{iff} & \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \text{ such that } \tau \sim \tau' \\ \mathcal{M}, \tau \models [x]\varphi & \text{iff} & \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \text{ such that } \tau \sim_x \tau' \end{array}$$

$[x]$ is what some authors (e.g. (Horty, 2001)) call the 'Chellas *stit*'. However, it is important to stress that $[x]\varphi$ is a *transition formula* expressing a property of transitions and that φ is also a transition formula. When $[x]\varphi$ is true at a transition τ, we will say that φ is necessary for how x acts in τ. $\Box$ and each $[x]$ are normal modal operators of type S5. The schema

$$\Box\varphi \to [x]\varphi$$

is valid for all agents x in Ag.

We also have the following relationships between state formulas and transition formulas. All instances of the transition formula $0{:}\underset{\to}{\Box}\varphi \leftrightarrow \Box\varphi$ are valid, as are the state formulas $F \to \underset{\to}{\Box}0{:}F$ and $(\underset{\to}{\Diamond}\top \land \underset{\to}{\Box}0{:}F) \to F$, i.e., $\underset{\to}{\Diamond}0{:}F \leftrightarrow (\underset{\to}{\Diamond}\top \land F)$.

In what follows it is convenient to employ a functional notation. Let:

$$\begin{array}{rcl} \mathit{alt}(\tau) & =_{\mathrm{def}} & \{\tau' \mid \tau \sim \tau'\} \\ \mathit{alt}_x(\tau) & =_{\mathrm{def}} & \{\tau' \mid \tau \sim_x \tau'\} \end{array}$$

alt is for 'alternative'. ($\mathit{alt}(\tau)$ and $\mathit{alt}_x(\tau)$ are thus the equivalence classes $[\tau]^{\sim}$ and $[\tau]^{\sim_x}$ respectively. The alt_x notation is slightly easier to read.)

For every $x \in Ag$ and every $\tau \in R$, we have $\mathit{alt}_x(\tau) \subseteq \mathit{alt}(\tau)$. The truth conditions can be expressed as:

$$\begin{aligned}\mathcal{M}, \tau \models \Box\varphi \quad &\text{iff} \quad \mathit{alt}(\tau) \subseteq \|\varphi\|^{\mathcal{M}} \\ \mathcal{M}, \tau \models [x]\varphi \quad &\text{iff} \quad \mathit{alt}_x(\tau) \subseteq \|\varphi\|^{\mathcal{M}}\end{aligned}$$

$\mathit{alt}(\tau)$ is the set of transitions from the same initial state as τ, and $\mathit{alt}_x(\tau)$ is the set of transitions from the same initial state as τ in which x performs the same action as it does in τ: these are the possible alternative actions that could be performed by x (deliberatively, intentionally, but possibly also unwittingly). $\mathit{alt}_x(\tau)$ is the equivalence class that contains τ, and so, just as in the *stit* framework, it can be regarded as the action performed by x in the transition τ.

For readers familiar with *stit* models, and models for the deliberative *stit* in particular, the set of transitions from any given state s can be seen (some technical details aside) as the set of histories passing through a moment s. (It would be better to speak of mappings from moments to states but I do not want to dwell on technical details here.) Since every transition τ has a unique initial state $\mathit{prev}(\tau)$, every transition can also be thought of as a moment-history pair m/h where the moment m is the initial state $\mathit{prev}(\tau)$ and the history h is the transition τ. Putting aside technical details, one can think of transition system models as the special case of a (deliberative) *stit* model in which there is a single moment-history pair for every history. Evaluating formulas on transitions, as we do, is then like evaluating formulas on moment-history pairs in *stit*-models. Evaluating formulas on states, as we also do, would be like evaluating formulas on moments in *stit*-models. (Mark Brown in his paper in this volume raises the question of whether points of valuation should be moments or moment/history pairs. We want both, which is why we employ a two-sorted language.) Put in these terms, $\tau \sim \tau'$ represents two moment-history pairs $\tau = m/h$ and $\tau' = m/h'$ through the same moment m. The equivalence relations $\sim_x$ determine what in *stit* would be the agent x's choice function. When $\tau = m/h$, $\mathit{alt}(\tau)$ is the set H_m of histories passing through m, and $\mathit{alt}_x(\tau)$ is $\mathit{Choice}_x^m(h)$, i.e., the action performed by x at moment m in history h, or equivalently, the subset of histories H_m in which x performs the same action at moment m as it does at moment m in history h.

Indeed, if we ignore states (or formulas on moments) and look only at transitions (or formulas on moment-history pairs), then models are of the form

$$\langle R, \sim, \{\sim_x\}_{x \in Ag}, h^{\mathrm{a}} \rangle$$

These are exactly the abstract models of the deliberative *stit* discussed in (Balbiani et al., 2008) *except that* there the models have a slightly different, but equivalent, form because they incorporate an extra, very strong 'independence of agents' assumption characteristic of *stit*.

stit-independence says (Horty, 2001, p30) that 'at each moment, each agent must be able to perform any of his available actions, no matter which actions are performed at that moment by the other agents' or (Belnap and Perloff, 1993, p26) 'any combination of choices made by distinct agents at exactly the same moment is consistent.'

Expressed as a condition on alt_x, *stit*-independence would require that, for all pairs of agents x and y in Ag, for all τ_x and τ_y such that $\tau_x \sim \tau_y$,

$$\mathit{alt}_x(\tau_x) \cap \mathit{alt}_y(\tau_y) \neq \emptyset$$

and more generally that, for all transitions τ and all mappings $s'_\tau \colon Ag \to \mathit{alt}(\tau)$:

$$\textstyle\bigcap_{x \in Ag} \mathit{alt}_x(s'_\tau(x)) \neq \emptyset$$

We will not need the more general form in this paper since none of the examples have more than two agents.

I do not understand what the 'independence of agents' assumption is for and why it is adopted without question in works on *stit*. I have not been able to find any convincing justification for it in the literature. (Belnap and Perloff, 1993, p26) remark that '... we do not consider the evident fact that agents interact in space-time' but do not say why. Why *not* consider the evident fact that agents interact in space-time? It is only a matter of dropping the *stit*-independence condition. What purpose does it serve? It is sometimes suggested that *stit*-independence is needed in order to ensure that some combination of actions by individual agents always exists. But that is not so. In the *stit* framework some combination of actions by agents always exists, without the *stit*-independence assumption. The *stit*-independence condition insists that *every* combination of actions always exists, which is much stronger. Further discussion is for another occasion. In what follows, some of the models will satisfy the *stit*-independence condition and some will not.

Group actions Just as in *stit*, the account generalises naturally to dealing with the joint actions of groups (sets) of agents. Let G be a non-empty subset of Ag. $\mathit{alt}_x(\tau)$ represents the action performed by x in the transition τ, which is the set of transitions in $\mathit{alt}(\tau)$ in which x performs the same action as it does in τ. $\bigcap_{x \in G} \mathit{alt}_x(\tau)$ is the set of transitions in $\mathit{alt}(\tau)$ in which every agent in G performs the same action as it does in τ, and is thus a representation of the joint action performed by the group G in the transition τ.

The truth conditions are:

$$\mathcal{M}, \tau \models [G]\varphi \quad \text{iff} \quad \mathit{alt}_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$$

where

$$\begin{aligned} \mathit{alt}_G(\tau) &=_{\text{def}} \textstyle\bigcap_{x \in G} \mathit{alt}_x(\tau) \\ \sim_G &=_{\text{def}} \textstyle\bigcap_{x \in G} \sim_x \end{aligned}$$

That is, expressed in the relational notation:

$$\begin{aligned} \mathcal{M}, \tau \models [G]\varphi \quad &\text{iff} \quad \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in \textstyle\bigcap_{x \in G} \mathit{alt}_x(\tau) \\ &\text{iff} \quad \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in \mathit{alt}_G(\tau) \\ &\qquad \text{where } \mathit{alt}_G(\tau) =_{\text{def}} \textstyle\bigcap_{x \in G} \mathit{alt}_x(\tau) \\ &\text{iff} \quad \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \text{ such that } \tau \sim_G \tau' \\ &\qquad \text{where } \sim_G =_{\text{def}} \textstyle\bigcap_{x \in G} \sim_x \end{aligned}$$

When $[G]\varphi$ is true at τ we will say that φ is necessary for how the agents G collectively act in τ. (Which is not the same as saying that they act together, i.e., as a kind of coalition or collective agent. We are not discussing genuine collective agency in this paper.) Clearly $\models [\{x\}]\varphi \leftrightarrow [x]\varphi$ for every x in Ag.

Axiomatisation $\Box$ and every $[x]$ and every $[G]$ are normal modal operators of type S5. The logic is the smallest normal logic containing all instances of the following axiom schemas, for all non-empty subsets G and G' of Ag:

$$\begin{array}{lll} \Box & \text{type S5} & \\ [G] & \text{type S5} & \\ \Box\varphi \to [G]\varphi & & \\ [G]\varphi \to [G']\varphi & & (G \subseteq G') \end{array}$$

2.3 Acts differently

We also want to be able speak about alternative transitions from the same initial state in which an agent x, or set of agents G, acts *differently* from the way it acts in a transition τ. We further extend the language of transition formulas with operators $[\overline{x}]$ and $[\overline{G}]$ for every agent x in Ag and every non-empty subset G of Ag: $[\overline{x}]\varphi$ and $[\overline{G}]\varphi$ are transition formulas when φ is a transition formula. $\langle\overline{x}\rangle$ and $\langle\overline{G}\rangle$ are the respective duals.

The truth conditions are:

$$\begin{array}{lll} \mathcal{M}, \tau \models [\overline{x}]\varphi & \text{iff} & (\mathit{alt}(\tau) - \mathit{alt}_x(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \\ \mathcal{M}, \tau \models [\overline{G}]\varphi & \text{iff} & (\mathit{alt}(\tau) - \mathit{alt}_G(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \end{array}$$

Note that $\models\ [\overline{x}]\varphi \leftrightarrow [\overline{\{x\}}]\varphi$, and that:

$$\begin{array}{l} \models\ \langle\overline{G}\rangle\varphi \ \leftrightarrow\ \bigvee_{x \in G}\langle\overline{x}\rangle\varphi \\ \models\ [\overline{G}]\varphi \ \leftrightarrow\ \bigwedge_{x \in G}[\overline{x}]\varphi \end{array}$$

2.4 'Brings it about' modalities

In logics of agency, expressions of the form 'agent x brings it about that' or 'sees to it that' are typically constructed from two components. The first is a 'necessity condition': φ must be necessary for how agent x acts. The second component is used to capture the fundamental idea that φ is, in some sense, caused by or is the result of actions by x. Most accounts of agency introduce a negative counterfactual or 'counteraction' condition for this purpose, to express that had x not acted in the way that it did then the world would, or might, have been different.

Let $\mathrm{E}_x\varphi$ represent that agent x brings it about, perhaps unwittingly, that (a transition has) a certain property φ. $\mathrm{E}_x\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) $\mathcal{M}, \tau \models [x]\varphi$, that is, all transitions from the same initial state as τ in which x acts in the same way as it does in τ are of type φ, or as we also say, φ is necessary for how x acts in τ;

(2) (counteraction) had x acted differently than it did in τ then the transition might have been different: there exists a transition τ' in $\mathcal{M}$ such that $\tau \sim \tau'$ and $\tau \not\sim_x \tau'$ and $\mathcal{M}, \tau' \models \neg\varphi$.

$\mathrm{E}_x\varphi$ is then defined as $\mathrm{E}_x\varphi =_{\mathrm{def}} [x]\varphi \land \langle\overline{x}\rangle\neg\varphi$, or equivalently:

$$\mathrm{E}_x\varphi \ =_{\mathrm{def}}\ [x]\varphi \land \neg[\overline{x}]\varphi$$

The difference modalities $[\overline{x}]$ are useful in their own right (see Section 6), but in order to avoid introducing further technical machinery, we note that if our purpose is only to construct the E_x modalities, then we can simplify. The counterfactual condition (2) can be simplified because of the necessity condition (1): if there is a transition τ' in $\mathcal{M}$ such that $\tau \sim \tau'$ and $\mathcal{M}, \tau' \models \neg\varphi$ but where $\tau \sim_x \tau'$, then the necessity condition (1) does not hold: $\mathcal{M}, \tau \not\models \varphi$. In other words, the following schema is valid, for all x in Ag:

$$\models\ ([x]\varphi \land [\overline{x}]\varphi) \leftrightarrow \Box\varphi$$

So instead of (2) for the counteraction condition we can take simply:

(2′) there exists a transition τ' in $\mathcal{M}$ such that $\tau \sim \tau'$ and $\mathcal{M}, \tau' \models \neg\varphi$.

This is just $\mathcal{M}, \tau \models \Diamond\neg\varphi$, or equivalently, $\mathcal{M}, \tau \models \neg\Box\varphi$.

The following simpler definition is thus equivalent to the original:

$$\mathrm{E}_x\varphi \ =_{\mathrm{def}}\ [x]\varphi \land \neg\Box\varphi$$

This is exactly the construction used in the definition of the 'deliberative *stit*' (Horty and Belnap, 1995)

$$[x \ \mathit{dstit}\colon \varphi] \ =_{\mathrm{def}}\ [x]\varphi \land \neg\Box\varphi$$

except of course that we are reading φ as expressing a property of a transition.

The notation $\mathrm{E}_x\varphi$ is from (Pörn, 1977) (though the semantics are different). It is chosen in preference to the *dstit* notation because it is more concise, and in order to emphasise that we do not want to incorporate the very strong *stit*-independence assumption that is built into *dstit*.

Notice that $\mathrm{E}_x\varphi \land \mathrm{E}_y\varphi$ is satisfiable even when $x \neq y$. Indeed

$$\models\ \mathrm{E}_x\varphi \land \mathrm{E}_y\varphi \leftrightarrow\ [x]\varphi \land [y]\varphi \land \neg\Box\varphi$$

It is possible to define a stronger kind of 'brings it about' modality which represents a sense in which it is agent *x and x alone* who brings it about that φ. We will not need that stronger form in this paper since none of the examples has more than two agents. See (Sergot, 2008a,b) for details and for discussion of some forms of collective action by groups (sets) of agents.

Note that adding the *stit*-independence condition validates, among other things, the following schema, for all distinct x and y in Ag:

$$\neg\mathrm{E}_x\mathrm{E}_y\varphi \qquad (x \neq y)$$

Finally, in many of the examples that follow we will be interested in expressions of the form $\mathrm{E}_x(0{:}F \land 1{:}G)$. We note for future reference that:

$$\models\ \mathrm{E}_x(0{:}F \land 1{:}G) \leftrightarrow (0{:}F \land \mathrm{E}_x 1{:}G)$$

3 Example: Vase (one agent)

We begin with a very simple example containing just a single agent a. Agent a can move a certain (precious) vase between indoors and outdoors. An element of indeterminism is introduced by allowing that it might be raining or not raining in any state, which is something that is outside the control of the agent a. Further, for the sake of an example, suppose it is forbidden, illegal, wrong for the vase to be outside in the rain.

Let state atoms *in* represent that the vase is indoors, *rain* that it is raining, and *red* that the state is forbidden/illegal. *out* is shorthand for $\neg in$; *green* is shorthand for $\neg red$.

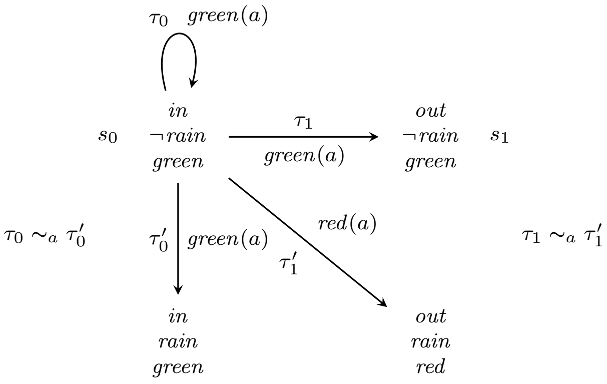


Figure 1: Transitions from state s_0 ($in \wedge \neg rain$)

Figure 1 shows a fragment of a transition system modelling this example, depicting the transitions from state s_0 ($in \wedge \neg rain$). The labels $green(a)$ and $red(a)$ on transitions will be explained presently. Figure 2 shows the transitions from state s_1 ($out \wedge \neg rain$). They are shown in a separate diagram simply to reduce clutter. Not shown in the diagrams are the transitions from the other two states in the model, where it is raining.

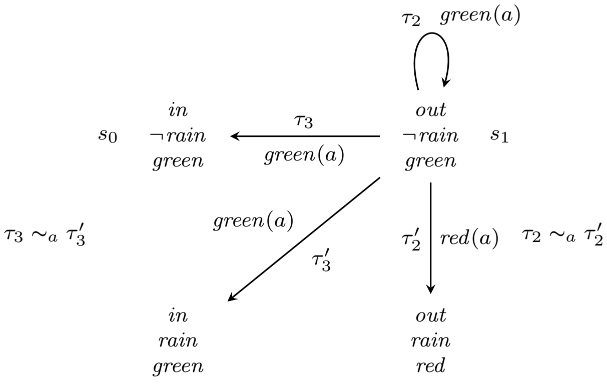


Figure 2: Transitions from s_1 ($out \wedge \neg rain$)

I have deliberately not included any transition atoms to name the actions by a. A perceived advantage of the *stit* treatment of action is that we are not forced to say exactly what action is performed by a when the vase is moved or left where it is. We need only say (in the example as I am thinking of it) that, whatever these actions are, the actions by a are the same in the two transitions τ_0 and τ_0' ($\tau_0 \sim_a \tau_0'$); they differ only in whether it is raining or not in the resulting state and not in what agent a does when the vase stays in place. And similarly, $\tau_1 \sim_a \tau_1'$. The possible actions by a in state s_0 are thus $\{\,\{\tau_0, \tau_0'\},\ \{\tau_1, \tau_1'\}\,\}$, and those in state s_1 are $\{\,\{\tau_2, \tau_2'\},\ \{\tau_3, \tau_3'\}\,\}$. From the diagram, one can see that they can be characterised in various ways, including:

$$\begin{aligned}
\{\tau_0, \tau_0'\} &= \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land 1{:}\mathit{in}\| \\
&= \|\neg 0{:}\mathit{rain} \land \mathrm{E}_a(0{:}\mathit{in} \land 1{:}\mathit{in})\| \\
&= \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land \mathrm{E}_a 1{:}\mathit{in}\|
\end{aligned}$$

$$\begin{aligned}
\{\tau_1, \tau_1'\} &= \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land 1{:}\mathit{out}\| \\
&= \|\neg 0{:}\mathit{rain} \land \mathrm{E}_a(0{:}\mathit{in} \land 1{:}\mathit{out})\| \\
&= \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land \mathrm{E}_a 1{:}\mathit{out}\|
\end{aligned}$$

And similarly for a's possible actions in state s_1. $\{\tau_2, \tau_2'\} = \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{out} \land 1{:}\mathit{out}\|$ and $\{\tau_3, \tau_3'\} = \|\neg 0{:}\mathit{rain} \land 0{:}\mathit{out} \land 1{:}\mathit{in}\|$, and so on.

Not shown in Figures 1 and 2 are the transitions from the two states where it is raining. It is for that reason that the actions by a in state s_0 are not just $\|0{:}\mathit{in} \land 1{:}\mathit{in}\|$ and $\|0{:}\mathit{in} \land 1{:}\mathit{out}\|$ but $\|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land 1{:}\mathit{in}\|$ and $\|\neg 0{:}\mathit{rain} \land 0{:}\mathit{in} \land 1{:}\mathit{out}\|$. The example as formulated leaves open the possibility that moving-when-it-is-raining-now is not the same action as moving-when-it-is-not-raining-now, and not-moving-when-it-is-raining-now is not the same action as not-moving-when-it-is-not-raining-now.

Suppose however that we *do* want to say that the actions performed by a are the same irrespective of whether it is raining or not in the initial or final states: suppose the actions performed by a are the same in all transitions $\|0{:}\mathit{in} \land 1{:}\mathit{out}\|$, the same in all transitions $\|0{:}\mathit{out} \land 1{:}\mathit{in}\|$, and the same in all transitions $\|(0{:}\mathit{in} \land 1{:}\mathit{in}) \lor (0{:}\mathit{out} \land 1{:}\mathit{out})\|$ where the vase stays where it is.

That would require an adjustment to the model structures. We could add a relation $=_x$ for every agent x in Ag, using $\tau =_x \tau'$ to represent that the action performed by x is the same in any transitions τ and τ' not just those that have the same initial state. We would then have:

$$\sim_x \ =_{\mathrm{def}}\ \sim \cap =_x$$

A strong argument could be made that, for modelling purposes, this would be a useful and natural extension. It is easy to accommodate but I will not do so in the rest of this paper. It would not fit so well in the *stit*-framework since that would require relating actions/choices across moments in different, incompatible histories which does not seem so natural.

One final remark: I am thinking here of 'moving' as a basic, simple kind of act, such as moving an arm while it grasps the vase or pushing the vase in one movement from one location to another. I am not thinking of 'moving' as an extended process of some kind requiring the vase to be packed up, transported somehow to the new location, and unpacked (say). In the latter case, the transitions in the diagrams would correspond to executions of this more elaborate 'moving' process. In that case we might well *not* want to say that $\tau_1 \sim_a \tau_1'$, since the moving process might be different if it happens to be raining as the vase reaches the *out* location. Indeed there might be many different 'moving' transitions between *in* and *out*, each corresponding to a different combination of actions by a. We will return to this point later under discussions of granularity of representations.

Example: obligations

There is an obligation on a that the vase is not outside in the rain. Let the transition atom $\mathit{red}(a)$ represent a transition in which a fails to comply with this obligation. $\mathit{green}(a)$ is shorthand for $\neg\mathit{red}(a)$ and so is satisfied by transitions in which a does comply. Figures 1 and 2 show these labels on transitions. (It is an open question whether the transitions from a *red* state where the vase is already out in the rain should be $\mathit{green}(a)$ or $\mathit{red}(a)$ transitions. We will ignore that question here.)

One sense of 'it is obligatory for agent x to 'do' φ' in a state s can be defined as follows:

$$\mathrm{O}_x\varphi \;=_{\mathrm{def}}\; \underset{\rightarrow}{\Box}(\mathit{green}(x) \rightarrow \varphi)$$

or equivalently $\mathrm{O}_x\varphi =_{\mathrm{def}} \underset{\rightarrow}{\Box}(\neg\varphi \rightarrow \mathit{red}(x))$. It follows that $\models\; \mathrm{O}_x\,\mathit{green}(x)$.

But *can* x comply with its obligations?

One sense of agent ability is that discussed by Mark Brown (1988); it is expressed in the *stit* framework by the formula $\Diamond[x]\varphi$. In the present framework where we distinguish between state formulas and transition formulas, that sense of x can 'do' φ in state s would be expressed:

$$\mathit{Can}_x\varphi \;=_{\mathrm{def}}\; \underset{\rightarrow}{\Diamond}[x]\varphi$$

In the example:

$$s_1 \models\; \mathit{Can}_a\,\mathit{green}(a) \qquad (\underset{\rightarrow}{\Diamond}[a]\,\mathit{green}(a))$$

But *what* should a do to ensure $\mathit{green}(a)$? We are looking for transitions from s_1 of type $[a]\,\mathit{green}(a)$. There are such transitions: those in which the vase is moved from *out* to *in*. a might also comply with its obligation by leaving the vase outdoors but compliance then is a matter of chance, outside a's control.

'Absence of moral luck' (Craven and Sergot, 2008) is an (optional) rationality constraint that we might often want to check for when considering sets of regulations or specifications for computer systems. It reflects the idea that, for practical purposes, whether actions of agent x are in accordance with the norms directed at x should depend only on x's actions, not on the actions of other agents, nor actions in the environment, nor other extraneous factors. It can be expressed

- 'absence of moral luck' (in a model $\mathcal{M}$)

$$\mathcal{M} \models\; \mathit{green}(x) \rightarrow [x]\,\mathit{green}(x)$$

- 'absence of moral luck' (locally, in a state s)

$$\mathcal{M}, s \models\; \underset{\rightarrow}{\Box}(\,\mathit{green}(x) \rightarrow [x]\,\mathit{green}(x)\,)$$

In the example, if agent a leaves the vase outside, it is a matter of luck whether it complies with its obligation or not, for this will depend on whether it rains, and that is an extraneous factor outside of a's control. Thus:

$$\tau_2 \not\models \mathit{green}(a) \rightarrow [a]\,\mathit{green}(a)$$
$$s_1 \not\models \underset{\rightarrow}{\Box}(\mathit{green}(a) \rightarrow [a]\,\mathit{green}(a)) \qquad \text{(no 'absence of moral luck')}$$

'Absence of moral luck' is a rather strong form of 'Ought implies Can'. Other, weaker forms can also be expressed. For instance, 'Ought implies Can' (1) (at state s in model $\mathcal{M}$)

$$\mathcal{M}, s \models \mathrm{O}_x\varphi \rightarrow \underset{\rightarrow}{\Diamond}\varphi \qquad \text{for all } \varphi$$

This is equivalent (it turns out) to $\mathcal{M}, s \models \underset{\rightarrow}{\Diamond}\mathit{green}(x)$ and to $\mathcal{M}, s \models \neg\mathrm{O}_x\bot$.

Compare 'Ought implies Can' (2) (at state s in model $\mathcal{M}$):

$$\mathcal{M}, s \models \mathrm{O}_x\varphi \rightarrow \mathit{Can}_x\varphi \qquad \text{for all } \varphi$$

This is equivalent (it turns out) to $\mathcal{M}, s \models \mathit{Can}_x\,\mathit{green}(x)$. It is easy to check that 'Ought implies Can' (2) is stronger than (implies) 'Ought implies Can' (1). 'Absence of moral luck' is stronger still: it implies 'Ought implies Can' (2).

In the example,

$$s_1 \models \mathit{Can}_a\,\mathit{green}(a)$$

and so 'Ought implies Can' (2) at s_1. But there is no 'absence of moral luck' at s_1, as demonstrated earlier.

States where the vase is *in* are similar. Refer to Figure 1. Here again

$$s_0 \not\models \underset{\rightarrow}{\Box}(\mathit{green}(a) \rightarrow [a]\,\mathit{green}(a)) \qquad \text{(no 'absence of moral luck')}$$
$$s_0 \models \mathit{Can}_a\,\mathit{green}(a) \qquad (\underset{\rightarrow}{\Diamond}[a]\,\mathit{green}(a))$$

In contrast, suppose that the obligation on a is not to ensure that the vase is not outdoors in the rain but instead that the vase is to be *moved* indoors if it is outdoors. In that case, transition τ_2 in Figure 2, which was labelled $\mathit{green}(a)$ would be labelled $\mathit{red}(a)$. In that modified form of the example we have:

$$s_1 \models \mathrm{O}_a(0{:}\mathit{out} \wedge 1{:}\mathit{in})$$
$$s_1 \models \underset{\rightarrow}{\Box}(\mathit{green}(a) \rightarrow [a]\,\mathit{green}(a)) \qquad \text{('absence of moral luck')}$$
$$s_1 \models \mathit{Can}_a\,\mathit{green}(a) \qquad (\underset{\rightarrow}{\Diamond}[a]\,\mathit{green}(a)) \qquad \text{(which follows from the above)}$$

4 Example: Vase (two agents)

Let us now introduce another agent, b. Suppose that the vase can be in one of three possible, mutually exclusive, locations, *in*, *out*, and *over*, say. Agent a can move the vase between *in* and *out*, and b can move it between *out* and *over* (but not simultaneously). There is an obligation on a to move the vase to *in* if it is *out*. There is no obligation on b to move the vase.

Figure 3 shows the transitions from the state s_1 where the vase is *out* and it is not raining.

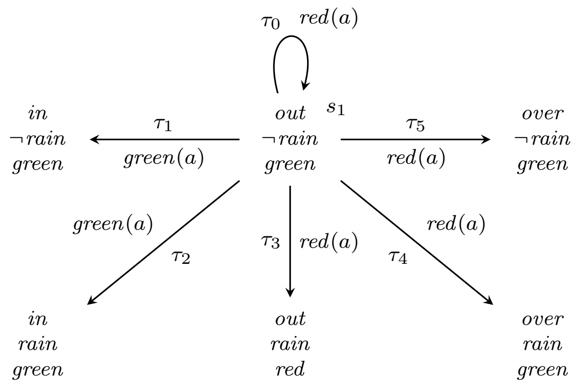


Figure 3: Transitions from state s_1 ($out \wedge \neg rain$)

The possible actions by a in state s_1 are (as we conceive the example) $\{\{\tau_0, \tau_3, \tau_4, \tau_5\}, \{\tau_1, \tau_2\}\}$. From the diagram:

$$\begin{aligned} \{\tau_0, \tau_3, \tau_4, \tau_5\} &= \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}\neg in\| \\ \{\tau_1, \tau_2\} &= \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}in\| \end{aligned}$$

The possible actions by b in state s_1 are $\{\{\tau_0, \tau_1, \tau_2, \tau_3\}, \{\tau_4, \tau_5\}\}$.

$$\begin{aligned} \{\tau_0, \tau_1, \tau_2, \tau_3\} &= \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}\neg over\| \\ \{\tau_4, \tau_5\} &= \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}over\| \end{aligned}$$

This model does not satisfy *stit*-independence:

$$\{\tau_1, \tau_2\} \cap \{\tau_4, \tau_5\} \;=\; \emptyset$$

That is as it should be: the actions of a and b are not independent. If a moves the vase to *in* then b cannot simultaneously move it to *over*, and vice-versa.

a still has an obligation to move the vase *out* from *in*: the transitions in the diagram are labelled *green*(a) and *red*(a) accordingly.

$$\begin{aligned} & s_1 \models \mathrm{O}_a(0{:}out \wedge 1{:}in) \\ & s_1 \models \underset{\rightarrow}{\Box}(green(a) \rightarrow [a]\,green(a)) \qquad \text{(`no moral luck')} \\ & s_1 \models Can_a\, green(a) \qquad (\underset{\rightarrow}{\Diamond}[a]\,green(a)) \end{aligned}$$

That is as expected. But note also that:

$$s_1 \models \underset{\rightarrow}{\Diamond}[b]\,red(a) \qquad (Can_b\, red(a))$$

The last says that in state s_1, b can act in such a way that a necessarily fails to fulfill its obligation. And that is also surely right: if b moves the vase from *out* to *over*, a could not simultaneously move it from *out* to *in*, which is a's obligation.

One can see from the diagram that all transitions where b moves the vase to *over* are $\mathit{red}(a)$. Thus:

$$\begin{aligned}
&s_1 \models \underset{\rightarrow}{\Box}(\,\mathrm{E}_b 1{:}\,\mathit{over} \rightarrow \mathit{red}(a)\,)\\
&s_1 \models \underset{\rightarrow}{\Box}(\,\mathit{red}(a) \rightarrow [a]\,\mathit{red}(a)\,) \qquad \text{(‘no moral luck’)}\\
&s_1 \models \underset{\rightarrow}{\Box}(\,\mathrm{E}_b 1{:}\,\mathit{over} \rightarrow [a]\,\mathit{red}(a)\,)
\end{aligned}$$

and moreover:

$$\begin{aligned}
&s_1 \models \underset{\rightarrow}{\Box}(\,\mathrm{E}_b 1{:}\,\mathit{over} \rightarrow \mathrm{E}_a\,\mathit{red}(a)\,)\\
&s_1 \models \underset{\rightarrow}{\Box}(\,\mathrm{E}_b 1{:}\,\mathit{over} \rightarrow \mathrm{E}_b\mathrm{E}_a\,\mathit{red}(a)\,)
\end{aligned}$$

The last two formulas in particular may seem counterintuitive on an informal reading. The first seems to say that if b brings about or is responsible for the vase's moving to *over* then a brings about or is responsible for violating its obligation; the second that b thereby brings about or is responsible for a's violating its obligation. The question of how these formulas may be read informally as *stit* statements does not arise because the example is not a *stit*-model. It does not satisfy the *stit*-independence condition.

5 Example: Vase, minor variation

The following minor variation of the example is intended to make some further observations about the representation of actions.

Let us suppose there are just two (mutually exclusive) locations *in* and *out* for the vase, and that agents a and b can both move the vase between them (but not simultaneously).

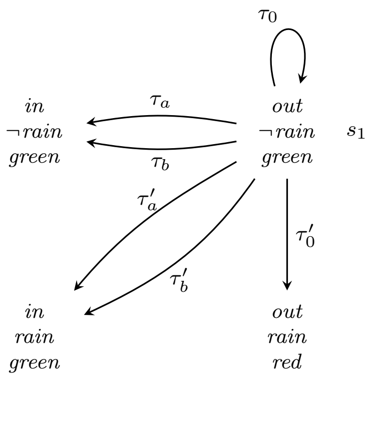


Figure 4: Transitions from state s_1

Informally, in Figure 4 τ_a and τ'_a are transitions where a moves the vase, and τ_b and τ'_b are transitions where b moves it.

Actions by a in state s_1:	$\{\,\{\tau_0, \tau'_0, \tau_b, \tau'_b\},\ \{\tau_a, \tau'_a\}\,\}$
Actions by b in state s_1:	$\{\,\{\tau_0, \tau'_0, \tau_a, \tau'_a\},\ \{\tau_b, \tau'_b\}\,\}$

There is no *stit*-independence in this model: a and b cannot both move the vase simultaneously.

$$\{\tau_a, \tau'_a\} \cap \{\tau_b, \tau'_b\} \;=\; \emptyset$$

Suppose for the sake of an example that a and b *both* have an obligation to move the vase *in* if it is *out*: the transitions τ_a and τ'_a are $green(a)$, τ_b and τ'_b are $green(b)$, and all other transitions from state s_1 are $red(a)$ and $red(b)$.

In this example there are different transitions between the same pairs of states and we cannot identify the actions of a and b by reference only to what holds in initial and final states.

$$\begin{aligned}\{\tau_a, \tau'_a\} &\;\neq\; \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}in\| \\ \{\tau_0, \tau'_0, \tau_b, \tau'_b\} &\;\neq\; \|0{:}\neg rain \wedge 0{:}out \wedge 1{:}\neg out\|\end{aligned}$$

(And likewise for b.)

It seems that in order to refer to a and b's actions we are forced to introduce some new (transition) atoms, which is something we were trying to avoid. But it happens that in this example the actions by a in state s_1 can be picked out as follows:

$$\begin{aligned}\{\tau_a, \tau'_a\} &= \|0{:}\neg rain \wedge \mathrm{E}_a(0{:}out \wedge 1{:}in)\| \\ &= \|0{:}\neg rain \wedge 0{:}out \wedge \mathrm{E}_a 1{:}in\|\end{aligned}$$

$$\begin{aligned}\{\tau_0, \tau'_0, \tau_b, \tau'_b\} &= \|0{:}\neg rain \wedge 0{:}out\| - \{\tau_a, \tau'_a\} \\ &= \|0{:}\neg rain \wedge 0{:}out \wedge \neg\mathrm{E}_a(0{:}out \wedge 1{:}in)\| \\ &= \|0{:}\neg rain \wedge 0{:}out \wedge \neg\mathrm{E}_a 1{:}in\|\end{aligned}$$

(And likewise for b.)

So, for convenience only, in this example we could define two new transition atoms $a{:}moves(out, in)$ and $b{:}moves(out, in)$ as follows:

$$\begin{aligned}a{:}moves(out, in) &=_{\mathrm{def}} \mathrm{E}_a(0{:}out \wedge 1{:}in) \\ b{:}moves(out, in) &=_{\mathrm{def}} \mathrm{E}_b(0{:}out \wedge 1{:}in)\end{aligned}$$

The possible actions by a in state s_1 are thus:

$$\begin{aligned}\{\tau_0, \tau'_0, \tau_b, \tau'_b\} &= \|0{:}\neg rain \wedge 0{:}out \wedge \neg a{:}moves(out, in)\| \\ \{\tau_a, \tau'_a\} &= \|0{:}\neg rain \wedge a{:}moves(out, in)\|\end{aligned}$$

(And likewise for b.)

I am not suggesting there is a general principle at work here. This is a very simple example where there are just two agents, and where each agent has just two possible actions in any state. In more complicated examples it is very far from obvious how to characterise possible actions by means of 'brings it about' formulas in this way. In bigger examples it very rarely works out so neatly.

It is perhaps worth reiterating that what seems natural in this framework is to say that the action performed by x in transition τ is not $[\tau]^{\sim_x}$ but $[\tau]^{=_x}$. Then a's possible actions in state s_1 would be simply $\|0{:}out \wedge \mathrm{E}_a 1{:}in\|$ and $\|0{:}out \wedge \neg\mathrm{E}_a 1{:}in\|$, i.e., $\|a{:}moves(out, in)\|$ and $\|0{:}out \wedge \neg a{:}moves(out, in)\|$.

In this example we have, among other things:

$$
\begin{aligned}
s_1 &\models \mathrm{O}_a\, a{:}\mathit{moves}(\mathit{out}, \mathit{in}) \land \mathrm{O}_b\, b{:}\mathit{moves}(\mathit{out}, \mathit{in})\\
s_1 &\models \mathrm{O}_a \mathrm{E}_a (0{:}\,\mathit{out} \land 1{:}\,\mathit{in}) \land \mathrm{O}_b \mathrm{E}_b (0{:}\,\mathit{out} \land 1{:}\,\mathit{in})\\
s_1 &\models \mathit{Can}_a\, a{:}\mathit{moves}(\mathit{out}, \mathit{in}) \land \mathit{Can}_b\, b{:}\mathit{moves}(\mathit{out}, \mathit{in})\\
s_1 &\models \neg \underset{\rightarrow}{\Diamond} (a{:}\mathit{moves}(\mathit{out}, \mathit{in}) \land b{:}\mathit{moves}(\mathit{out}, \mathit{in}))\\
s_1 &\models \underset{\rightarrow}{\Box} (\, \mathit{green}(a) \leftrightarrow \mathit{red}(b)\,)\\
s_1 &\models \underset{\rightarrow}{\Box} (\, a{:}\mathit{moves}(\mathit{out}, \mathit{in}) \rightarrow \mathrm{E}_b\, \mathit{red}(b)\,)\\
s_1 &\models \mathit{Can}_a \mathrm{E}_b\, \mathit{red}(b)\\
s_1 &\models \underset{\rightarrow}{\Box} (\, a{:}\mathit{moves}(\mathit{out}, \mathit{in}) \rightarrow \mathrm{E}_a \mathrm{E}_b\, \mathit{red}(b)\,)
\end{aligned}
$$

6 Example: Table

This example is intended to raise some questions about the treatment of agency, and in particular about the 'necessity' condition.

Suppose there is an agent a who can lift or lower its end of a table, or do neither. On the table stands a vase. If the table tilts, the vase might fall or it might not. If the vase falls, it might break or it might not. If the table does not tilt then the vase does not fall; if it does not fall, it does not break.

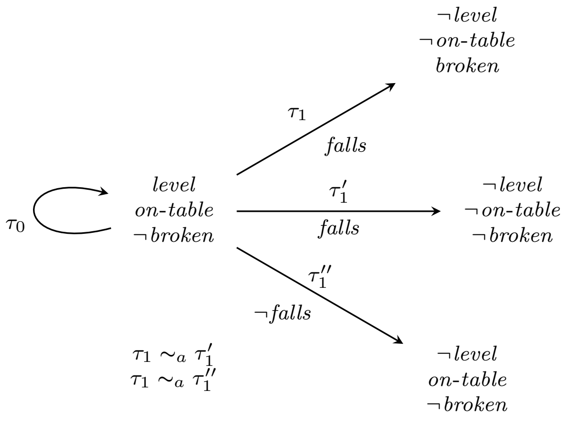


Figure 5: Transitions from the state in which the table is level and the vase stands on it

Figure 5 shows transitions from the state in which the table is level and the vase stands on it. State atoms *level*, *on-table* and *broken* have the obvious intended readings. There are other transitions not shown in the diagram and two more states, those in which the table is level (*level*) but the vase is not on it ($\neg$ *on-table*); in one of these the vase is broken, in the other it is not.

For convenience, let the transition atom *falls* be defined as follows:

$$\mathit{falls} \;=_{\mathrm{def}}\; 0{:}\,\mathit{on\text{-}table} \land 1{:}\neg\, \mathit{on\text{-}table}$$

The possible actions by a in this state are $\{\,\{\tau_0\},\ \{\tau_1, \tau_1', \tau_1''\}\,\}$. By reference to previous examples, there are various ways we can describe them, e.g.

$$\begin{aligned}
\{\tau_1, \tau_1', \tau''\} &= \|0{:}\,\textit{on-table} \wedge \mathrm{E}_a(0{:}\,\textit{level} \wedge 1{:}\neg\,\textit{level})\| \\
&= \|0{:}\,\textit{on-table} \wedge 0{:}\,\textit{level} \wedge \mathrm{E}_a 1{:}\neg\,\textit{level}\| \\[1ex]
\{\tau_0\} &= \|0{:}\,\textit{on-table} \wedge \neg\mathrm{E}_a(0{:}\,\textit{level} \wedge 1{:}\neg\,\textit{level})\| \\
&= \|0{:}\,\textit{on-table} \wedge 0{:}\,\textit{level} \wedge \neg\mathrm{E}_a 1{:}\neg\,\textit{level}\|
\end{aligned}$$

(Here, there is just a single agent a in the example and so the operator E_a could be omitted from all of the above.) The simpler expressions $\|\mathrm{E}_a(0{:}\,\textit{level} \wedge 1{:}\neg\,\textit{level})\|$ and $\|\neg\mathrm{E}_a(0{:}\,\textit{level} \wedge 1{:}\neg\,\textit{level})\|$ are not sufficient to pick out a's actions: there are other transitions not shown in the diagram where a lifts or lowers its end of the table when the vase is not on it. On the other hand, as observed earlier, we might well want to say that the actions of a's lifting its end of the table or not lifting are the same whether the vase stands on it or not. That would identify actions with equivalence classes of $=_a$ rather than $\sim_a$.

But here is the main point. Suppose that a tilts the table and the vase falls and breaks:

$$\tau_1 \models \textit{falls} \wedge 0{:}\neg\,\textit{broken} \wedge 1{:}\,\textit{broken}$$

Had a not tilted the table the vase would not have fallen. But a is not responsible for, does not bring about, the breaking of the vase:

$$\begin{array}{ll}
\tau_1 \not\models \mathrm{E}_a \textit{falls} & \quad\text{(because } \tau_1 \not\models [a]\textit{falls}) \\
\tau_1 \not\models \mathrm{E}_a 1{:}\,\textit{broken} & \quad\text{(because } \tau_1 \not\models [a]1{:}\,\textit{broken})
\end{array}$$

It is not necessary for what a does in τ_1 that the vase falls, and it is not necessary for what a does in τ_1 that the vase breaks.

Examples such as this, and many others, suggest that there is a weaker sense in which a 'brings about' or is responsible for the falling and breaking of the vase when a tilts the table. What is this weaker form?

There are two obvious candidates:

$$\begin{array}{ll}
(1) & \varphi \wedge [\overline{x}]\neg\varphi \\
(2) & \varphi \wedge \langle\overline{x}\rangle\neg\varphi
\end{array}$$

The first says that x is responsible for φ because φ is true and had x acted differently, φ would not have been true. (1) is too strong (demands too much). The second is more plausible and is mentioned briefly in (Pörn, 1977): x is responsible for φ because φ is true, and had x acted differently, φ might not have been true. But (2) is too weak.

In this particular example, both are plausible at first sight: in transition τ_1, the vase fell and broke but had a acted differently and not tilted the table, the vase would not have fallen and would not have broken.

(1) $\varphi \wedge [\overline{x}]\neg\varphi$ is too strong (demands too much). For suppose there were another way in which agent a could cause the vase to fall: suppose that a could dislodge the vase by jolting the table (say). Now, suppose that a lifts its end of the table and the vase falls. That would be a transition of type *falls*; but $\textit{falls} \wedge [\overline{a}]\neg\textit{falls}$ is false in that transition since there is another transition, where

a jolts instead of lifting, which also has *falls* true. So on that reading, a is not responsible for the vase's falling.

The candidate form (2) $\varphi \wedge \langle \overline{x} \rangle \neg \varphi$ is more plausible but is too weak. Consider a version of the earlier vase example in which agent a moves the vase between *in* and *out*. Consider a transition in which a moves the vase to *out* and it rains, that is, a transition of type $1{:}(\mathit{out} \wedge \mathit{rain})$. It is a who moves the vase, no-one else. In that transition, $\mathrm{E}_a 1{:}(\mathit{out} \wedge \mathit{rain})$ is false because it is not necessary for what a does that $1{:}(\mathit{out} \wedge \mathit{rain})$: $[a]1{:}(\mathit{out} \wedge \mathit{rain})$ is false because it might not have rained. However, had a acted differently (by not moving the vase) it might have been otherwise: $1{:}(\mathit{out} \wedge \mathit{rain}) \wedge \langle \overline{a} \rangle \neg 1{:}(\mathit{out} \wedge \mathit{rain})$ is true.

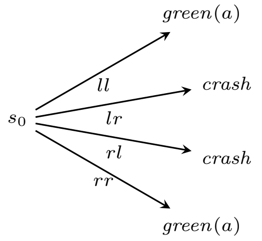
But that is too weak. By exactly the same argument, $1{:}\mathit{rain} \wedge \langle \overline{a} \rangle \neg 1{:}\mathit{rain}$ is also true in that transition: it rains, and had x acted differently, it might not have rained. Yet we would not want to say that agent a is reponsible for, or the cause of, or the one who brings about that it is raining.

It is far from clear whether this weaker sense of 'brings it about' or 'responsible for' can be articulated using the available resources. The problem is that nearly everything we want to say about agency in practice is of this weaker form. If a man walks into a room, puts a loaded revolver in his mouth and blows his brains out, we would surely want to say that he killed himself, that he was responsible for his death, that it was his actions that caused it. Yet he did not see to it or bring it about: it was not necessary for what he did that he died. The gun might have jammed, the bullet might have hit a thick part of the skull, the resulting injury might not have been fatal for any number of reasons. And this has nothing to do with probabilities. If a man walks in a room, picks a bullet at random from a barrel containing live and blank ammunition, loads his revolver, spins the chamber, then pulls the trigger and blows his brains out, we would say that he killed himself, even though the likelihood that those actions result in death is very small.

7 Example: Avoidance (fixed)

The next series of examples illustrates some common patterns in which the actions of one agent constrain, or possibly even determine, the actions of another.

Two agents a and b (cyclists, say) approach each other on a path. If both swerve left or both swerve right they avoid the crash; otherwise they crash. There is an obligation on a that there is no crash.



$$\begin{array}{rcl}
\mathit{ll} & \models & \mathit{green}(a) \wedge \neg [a]\, \mathit{green}(a) \\
\mathit{lr} & \models & \mathit{red}(a) \wedge \neg [a]\, \mathit{red}(a) \\
\mathit{rl} & \models & \mathit{red}(a) \wedge \neg [a]\, \mathit{red}(a) \\
\mathit{rr} & \models & \mathit{green}(a) \wedge \neg [a]\, \mathit{green}(a)
\end{array}$$

Figure 6: a and b can both swerve to left or right

Figure 6 shows the possible transitions as the agents approach each other.

The labels ll, lr, ... on transitions are just mnemonics: ll indicates that a and b both swerve left, lr that a swerves left and b swerves right, and so on. *crash* is a transition atom with the obvious intended reading. The transition atom $\mathit{green}(a)$ represents transitions in which a complies with its obligation. $\mathit{red}(a)$ is shorthand for $\neg\,\mathit{green}(a)$. In this model, $\mathit{green}(a) \leftrightarrow \neg\,\mathit{crash}$ is valid, or at least true in all transitions from the state s_0 depicted in the diagram.

One can see that in the case of a crash, agents a and b collectively bring it about that there is a crash, though neither individually does so. And similarly in the case where both swerve and there is no crash. We will not discuss possible forms of collective agency in this paper.

a has an obligation to avoid the crash but cannot guarantee that its actions will comply: 'Ought implies Can' (2) fails for this obligation:

$$s_0 \not\models \ \mathit{Can}_a\,\mathit{green}(a) \qquad\qquad (s_0 \not\models \ \underset{\rightarrow}{\Diamond}[a]\,\mathit{green}(a))$$

And there is no 'absence of moral luck' (which follows from the above)

$$s_0 \not\models \ \underset{\rightarrow}{\Box}(\mathit{green}(a) \to [a]\,\mathit{green}(a))$$

Agent as automaton

Consider the same example but suppose now that b has a fixed behaviour in this situation—a reflex or a deliberative decision procedure of some kind that always chooses the same action by b in these circumstances—b always swerves left (say). The obligation is still on a that there is no crash ($\mathit{green}(a) \leftrightarrow \neg\,\mathit{crash}$).

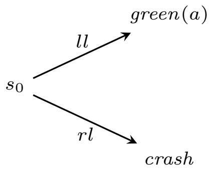


Figure 7: b always swerves left

At one level of detail, the possible transitions are as shown in Figure 7. Note first that there is 'absence of moral luck': $s_0 \models \underset{\rightarrow}{\Box}(\mathit{green}(a) \to [a]\,\mathit{green}(a))$. Moreover:

$$s_0 \models \ \underset{\rightarrow}{\Diamond}[a]\neg\,\mathit{crash} \qquad (\mathit{Can}_a \neg\,\mathit{crash})$$

(though a might not know this, or know how to avoid the crash).

But who is responsible in the case of a crash?

$$\begin{aligned} ll &\models \ [a]\neg\,\mathit{crash} \wedge \neg[b]\neg\,\mathit{crash} \wedge \mathrm{E}_a \neg\,\mathit{crash} \\ rl &\models \ [a]\,\mathit{crash} \wedge \neg[b]\,\mathit{crash} \wedge \mathrm{E}_a\,\mathit{crash} \end{aligned}$$

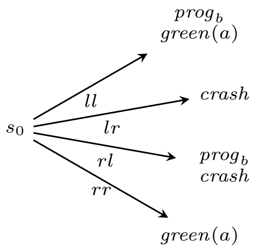
Because b's actions are fixed, b never brings about crash or no-crash: a is always solely responsible.

$$s_0 \models \ \underset{\rightarrow}{\Box}(\mathit{crash} \leftrightarrow \mathrm{E}_a\,\mathit{crash}) \ \wedge \ \underset{\rightarrow}{\Box}(\neg\,\mathit{crash} \leftrightarrow \mathrm{E}_a \neg\,\mathit{crash})$$

Perhaps this seems odd. Perhaps not—after all, this transition system models how b will *actually* behave. b's behaviour is treated here as if it were just part of the environment in which a operates, like a gate operated by a sensor or a traffic light. This seems perfectly reasonable if b is an automaton or a mechanical device of some kind. But what if b is not an automaton? What if b makes deliberate decisions about other actions but reacts automatically when faced by an oncoming a as here? b behaves like an automaton *in this respect* but not in every other.

Here is an alternative way of modelling this scenario. Let transition atom $prog_b$ represent that b acts in accordance with its protocol/decision procedure. (Here, to swerve left whatever a does.) We can assume $\mathcal{M} \models prog_b \leftrightarrow [b]prog_b$.

We need some way of referring to b's actions. Unlike in previous examples, there seems to be no recourse but to introduce a transition atom for this purpose. Let transition atom $b{:}l$ represent that b swerves left. b's protocol requires that $s_0 \models \underset{\rightarrow}{\Box}(\, prog_b \leftrightarrow b{:}l\,)$.



$$\begin{aligned} ll &\models \neg crash \wedge prog_b \\ lr &\models crash \wedge \neg prog_b \\ rl &\models crash \wedge prog_b \\ rr &\models \neg crash \wedge \neg prog_b \end{aligned}$$

Figure 8: b swerves left (explicit protocol)

Figure 8 depicts the model. In this version:

$$\begin{aligned} &s_0 \models \underset{\rightarrow}{\Box}(\, prog_b \rightarrow \mathrm{E}_b prog_b\,) \wedge \underset{\rightarrow}{\Box}(\, b{:}l \rightarrow \mathrm{E}_b b{:}l\,) \\ &s_0 \not\models Can_b \neg crash \\ &s_0 \models Can_a(\, prog_b \rightarrow \neg crash\,) \\ &s_0 \not\models \underset{\rightarrow}{\Box}(\, crash \rightarrow \mathrm{E}_a(prog_b \rightarrow crash)\,) \end{aligned}$$

The last is because $\mathrm{E}_a(prog_b \rightarrow crash)$ is false in the transition lr. Moreover:

$$s_0 \not\models \underset{\rightarrow}{\Box}(\, prog_b \rightarrow (crash \rightarrow \mathrm{E}_a\, crash)\,)$$

Of course it is a matter of *choice* how we model the example. It is not that one is right and the other is wrong. They model different things. Let us call the models in Figures 7 and 8 *actual* and *explicit protocol*, respectively.

In both models, b cannot avoid the crash, in the sense that:

$$s_0 \not\models \ Can_b \neg crash$$

And in both models a can avoid the crash (though a might not know this, nor know how). In the 'actual' model (Figure 7):

$$s_0 \models \ Can_a \neg crash$$

In the ‘explicit protocol’ model (Figure 8):

$$s_0 \models \; \mathit{Can}_a(\,\mathit{prog}_b \to \neg\,\mathit{crash}\,)$$

What differs is who is responsible in the case of a crash. In the ‘actual’ model (Figure 7) it is a:

$$s_0 \models \; \underset{\to}{\Box}(\,\mathit{crash} \to \mathrm{E}_a\,\mathit{crash}\,)$$

However in the ‘explicit protocol’ model (Figure 8):

$$s_0 \not\models \; \underset{\to}{\Box}(\,\mathit{crash} \to \mathrm{E}_a\,(\mathit{prog}_b \to \mathit{crash})\,)$$
$$s_0 \not\models \; \underset{\to}{\Box}(\,\mathit{prog}_b \to (\mathit{crash} \to \mathrm{E}_a\,\mathit{crash})\,)$$

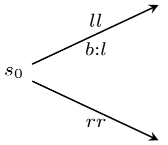
I find this slightly disturbing. I cannot see any general principles for choosing one of these models over the other. Both seem reasonable formalisations of the example, in their own way. And if one model has it that a is responsible for the crash, then it seems the other should have something comparable. But what? The two obvious candidates (the last two formulas above) do not work. It is not immediately obvious whether a sense of responsibility for crashing in the second model could be expressed and related neatly to the first.

8 Example: Avoidance (reaction)

Suppose now that b’s fixed reflex, program, deliberative procedure is to *react* to a—if a goes left, so does b; if a goes right, so does b. (The obligation on a that there is no crash will play no role in this example.)

As a first shot, let us ignore the temporal structure implicit in the term ‘reacts to’ and represent the possible behaviours in the example as atomic transitions.

We begin with ‘actual’ behaviour, as depicted in Figure 9.



$$\begin{aligned} \mathit{ll} &\models [a]\neg\,\mathit{crash} \wedge [b]\neg\,\mathit{crash} \\ \mathit{rr} &\models [a]\neg\,\mathit{crash} \wedge [b]\neg\,\mathit{crash} \end{aligned}$$

Figure 9: b reacts to a (atemporal, ‘actual’)

From the diagram:

$$s_0 \models \; \mathit{Can}_b\neg\,\mathit{crash} \qquad \text{(trivially, since } \underset{\to}{\Box}\neg\,\mathit{crash}\text{)}$$
$$s_0 \models \; \mathit{Can}_a\neg\,\mathit{crash}$$
$$s_0 \models \; \underset{\to}{\Box}(\,b{:}l \to \mathrm{E}_b\,b{:}l\,)$$

But also:

$$s_0 \models \; \underset{\to}{\Box}(\,b{:}l \to \mathrm{E}_a\,b{:}l\,)$$

For consider:

$$\begin{aligned} &ll \models [a]b{:}l \wedge \neg\Box b{:}l, \quad \text{and hence } ll \models \mathrm{E}_a b{:}l \\ &ll \models \mathrm{E}_b b{:}l \qquad \text{(similarly)} \end{aligned}$$

Furthermore:

$$\begin{aligned} &ll \models [a]\mathrm{E}_b b{:}l \wedge \neg\Box \mathrm{E}_b b{:}l \\ &ll \models \mathrm{E}_a \mathrm{E}_b b{:}l \\ &ll \models \mathrm{E}_b \mathrm{E}_a b{:}l \qquad \text{(similarly)} \end{aligned}$$

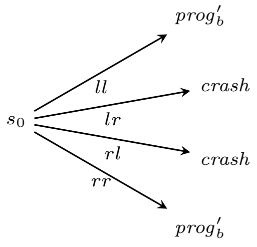
So then:

$$\begin{aligned} s_0 &\models \underset{\to}{\Box}(\, b{:}l \to \mathrm{E}_a b{:}l \,) \\ s_0 &\models \underset{\to}{\Box}(\, b{:}l \to \mathrm{E}_a \mathrm{E}_b b{:}l \,) \\ s_0 &\models \underset{\to}{\Box}(\, b{:}l \to \mathrm{E}_b \mathrm{E}_a \mathrm{E}_b b{:}l \,) \end{aligned}$$

There is obviously no *stit*-independence in this model. If there were then $\mathrm{E}_a \mathrm{E}_b \varphi$ would be false for every formula φ. That is a property validated by *stit*-independence.

Perhaps some of these formulas seem counterintuitive? What if we represent the temporal structure implicit in 'reacts to'? We will turn to that in a moment. Before that, for the sake of completeness, let us consider the 'explicit protocol' formulation of the atemporal model.

Let the transition atom prog'_b represent that b acts in accordance with its reaction procedure. We can assume $\mathcal{M} \models \mathit{prog}'_b \to [b]\mathit{prog}'_b$.



$$\begin{aligned} ll &\models \neg\,\mathit{crash} \wedge \mathit{prog}'_b \\ lr &\models \mathit{crash} \wedge \neg\mathit{prog}'_b \\ rl &\models \mathit{crash} \wedge \neg\mathit{prog}'_b \\ rr &\models \neg\,\mathit{crash} \wedge \mathit{prog}'_b \end{aligned}$$

Figure 10: b reacts to a (atemporal, 'explicit protocol')

Obviously in this example: $s_0 \models \underset{\to}{\Box}(\, \mathit{crash} \leftrightarrow \neg\mathit{prog}'_b \,)$. But suppose that b fails to react correctly, that is, that prog'_b is false. Is b then responsible for the crash? No:

$$\begin{aligned} s_0 &\not\models \underset{\to}{\Box}(\, \neg\mathit{prog}'_b \to [b]\,\mathit{crash} \,) \\ s_0 &\not\models \underset{\to}{\Box}(\, \neg\mathit{prog}'_b \to \mathrm{E}_b\, \mathit{crash} \,) \end{aligned}$$

b's protocol is to react to a: if b goes left and by doing so abides by its protocol, does it follow that a brings this about? No:

$$s_0 \not\models \underset{\to}{\Box}(\, b{:}l \to (\mathit{prog}'_b \to \mathrm{E}_a b{:}l)\,)$$

Though a does bring it about in the following sense:

$$s_0 \models \underset{\rightarrow}{\Box}(\, b{:}l \rightarrow \mathrm{E}_a(\mathit{prog}'_b \rightarrow b{:}l)\,)$$

And if transition atom $a{:}l$ represents that a swerves left, then:

$$s_0 \models \underset{\rightarrow}{\Box}(\, a{:}l \rightarrow \mathrm{E}_a(\mathit{prog}'_b \rightarrow b{:}l)\,)$$

Furthermore:

$$\begin{array}{l}
s_0 \not\models \mathit{Can}_b \neg\,\mathit{crash}, \quad \text{but } s_0 \models \ \mathit{Can}_b(\mathit{prog}'_b \rightarrow \neg\,\mathit{crash})) \\
s_0 \not\models \mathit{Can}_a \neg\,\mathit{crash}, \quad \text{but } s_0 \models \ \mathit{Can}_a(\mathit{prog}'_b \rightarrow \neg\,\mathit{crash})) \\
s_0 \not\models \underset{\rightarrow}{\Box}(\,\neg\,\mathit{crash} \rightarrow \mathrm{E}_a(\mathit{prog}'_b \rightarrow \neg\,\mathit{crash})\,) \\
s_0 \not\models \underset{\rightarrow}{\Box}(\,\neg\,\mathit{crash} \rightarrow \mathrm{E}_b(\mathit{prog}'_b \rightarrow \neg\,\mathit{crash})\,)
\end{array}$$

Notice that in this example we have had to rely on transition atoms to refer to the actions of a and b. I cannot see how we could do without them.

Temporal structure

Let us now compare a model at a finer level of detail, by making explicit the temporal structure implicit in the term 'reacts to'. We will consider the 'actual behaviour' model. The 'explicit protocol' version can be constructed in similar fashion but adds little new so we leave it out.

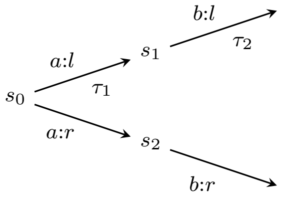


Figure 11: b reacts to a (temporal structure, 'actual behaviour')

In transition τ_2 of Figure 11 b reacts by swerving left after a swerves left in transition τ_1. We have

$$\tau_2 \models \ [b]\,b{:}l \wedge \neg \mathrm{E}_b\, b{:}l$$

and hence at τ_1

$$\tau_1 \models \ \mathrm{E}_a\, 1{:}\underset{\rightarrow}{\Box}\, b{:}l$$

So:

$$\begin{array}{l}
s_0 \models \underset{\rightarrow}{\Box}(\, a{:}l \rightarrow \mathrm{E}_a\, 1{:}\underset{\rightarrow}{\Box}\, b{:}l\,) \\
s_0 \not\models \underset{\rightarrow}{\Box}(\, a{:}l \rightarrow \mathrm{E}_a\, 1{:}\underset{\rightarrow}{\Box}\, \mathrm{E}_b\, b{:}l\,)
\end{array}$$

Indeed, in general the following are validities:

$$\begin{aligned} &\models \neg \mathrm{E}_x 1{:}\underset{\rightarrow}{\Box}\,\mathrm{E}_y\varphi \qquad \text{(any x, y, including $x = y$)} \\ &\models \neg [x] 1{:}\underset{\rightarrow}{\Box}\,\mathrm{E}_y\varphi \end{aligned}$$

This is because:

$$\models \neg\underset{\rightarrow}{\Box}\,\mathrm{E}_x\varphi$$

It is straightforward to derive this in the logic, or one can argue informally as follows. Suppose $s \models \underset{\rightarrow}{\Box}\,\mathrm{E}_x\varphi$. Then all transitions from s must have $\mathrm{E}_x\varphi$ true and hence also φ true and $\Diamond\neg\varphi$ true (by definition of $\mathrm{E}_x\varphi$). But if any transition from s has $\Diamond\neg\varphi$ true then it cannot be that all transitions from s have φ true, which contradicts the assumption.

So to recap: in the atemporal representation where the behaviours of a then b are modelled as atomic transitions

$$\begin{aligned} s_0 &\models \underset{\rightarrow}{\Box}(\, a{:}l \to \mathrm{E}_a b{:}l\,) \\ s_0 &\models \underset{\rightarrow}{\Box}(\, a{:}l \to \mathrm{E}_a \mathrm{E}_b b{:}l\,) \end{aligned}$$

At this level of detail a brings it about that b brings it about that b turns left. But at a finer level of detail where we make the temporal structure explicit

$$s_0 \not\models \underset{\rightarrow}{\Box}(\, a{:}l \to \mathrm{E}_a 1{:}\underset{\rightarrow}{\Box}\,\mathrm{E}_b b{:}l\,)$$

Instead a's actions force b's reaction, in the following sense:

$$s_0 \models \underset{\rightarrow}{\Box}(\, a{:}l \to \mathrm{E}_a 1{:}\underset{\rightarrow}{\Box}\, b{:}l\,)$$

In the temporal model then, $b{:}l \leftrightarrow \mathrm{E}_b b{:}l$ is not valid. On a casual reading one might think it should be.

My point is that I can see no general principle why we should always insist on picking the most detailed model. Indeed, why should we think that there is a most detailed model? What looks like an atomic transition at one level of detail can always be decomposed into something with finer structure if we look closely enough.

9 Example: Granularity

This last example is to illustrate that granularity of a model does not always depend on temporal structure.

Suppose there are two agents a and b. Both can be in one of two rooms, left and right, separated by a doorway. The agents can stay where they are or pass from one room to the other, but not simultaneously (the doorway is too narrow).

The diagram on the left of Figure 12 shows the possible transitions from the state where both agents are in the room on the left.

The possible actions by a and b in this state are as follows:

$$\begin{aligned} \text{Actions by } a\text{:} &\quad \{\,\{\tau_0,\tau_2\},\ \{\tau_1\}\,\} \\ \text{Actions by } b\text{:} &\quad \{\,\{\tau_0,\tau_1\},\ \{\tau_2\}\,\} \end{aligned}$$

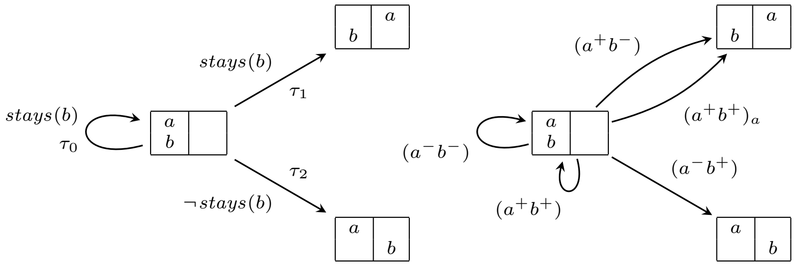


Figure 12: The same example at two different levels of detail

There is no *stit*-independence in this model (a and b cannot both pass through the doorway at the same time):

$$\{\tau_1\} \cap \{\tau_2\} = \emptyset$$

Let transition atom $stays(b)$ be true in transitions where b remains in the room on the left, as shown in the diagram.

Consider the transition τ_1 where a moves from left to right:

$$\tau_1 \models \mathrm{E}_b\, stays(b)$$

But also:

$$\tau_1 \models \mathrm{E}_a\, stays(b)$$
$$\tau_1 \models \mathrm{E}_a \mathrm{E}_b\, stays(b)$$

Indeed, if transition atom $moves(a)$ represents transitions where the location of a changes from left to right, then (amongst other things):

$$\mathcal{M} \models \; moves(a) \rightarrow \mathrm{E}_a \mathrm{E}_b\, stays(b)$$

Let us now consider the same example, but at a greater level of precision. a and b cannot both pass through the doorway at the same time. Why? For the sake of an example, suppose that if both try then one of two things can happen: either both fail and stay in the room on the left, or just a succeeds in moving through, because a is a little stronger or faster than b, say. b can never get through ahead of a. (Many other versions of the example are possible.)

The diagram on the right of Figure 12 depicts a model at this finer level of detail. The labels on the transitions are just mnemonics. In (a^+b^-), a tries to move to the room on the right (and succeeds) while b does not try to move. In (a^+b^+) both a and b try to get through the door but neither succeeds. In $(a^+b^+)_a$ both try to get through the door; a succeeds but b does not. In (a^-b^-) neither try. The possible actions of a and b in this state are therefore:

$$\text{Actions by } a\text{:} \quad \{\, \{(a^-b^-), (a^-b^+)\},\ \{(a^+b^-), (a^+b^+), (a^+b^+)_a\}\,\}$$
$$\text{Actions by } b\text{:} \quad \{\, \{(a^-b^-), (a^+b^-)\},\ \{(a^-b^+), (a^+b^+), (a^+b^+)_a\}\,\}$$

At this level of detail there is *stit*-independence in the model. (That does not always happen. Adding detail does not always produce *stit*-independence. It happens in this example.)

Let the transition atom *stays*(b) again represent those transitions where b stays on the left. *stit*-independence validates $\neg \mathrm{E}_a \mathrm{E}_b \varphi$ for all transition formulas φ. In particular, in this more detailed model of the example

$$\mathcal{M} \not\models \ \mathit{moves}(a) \rightarrow \mathrm{E}_a \mathrm{E}_b \mathit{stays}(b)$$

We still have:

$$\mathcal{M} \models \ \mathit{moves}(a) \rightarrow \mathrm{E}_a \mathit{stays}(b)$$

Evidently in this more detailed version of the model

$$\mathcal{M} \not\models \ \mathit{stays}(b) \leftrightarrow \mathrm{E}_b \mathit{stays}(b)$$

My point is that again important properties of the example change as detail is added. And it is not as though there is some most detailed model for which we should always aim. In the present example, a can sometimes get through the doorway ahead of b but not the other way round. We could also build a more detailed representation that models how that happens. So again: the models are different in some essential respects. We look to see which agent is responsible for, say, bringing it about that b stays where it is. At one level of detail, it is both a and b; at another level of detail it is only a. Indeed, it could be that at this level of detail, a brings it about that b does not bring it about that b stays where it is.

10 Conclusion

The purpose of the paper was to explore how easy or difficult it would be to formulate some simple examples in a *stit*-like framework. I deliberately picked examples with a simple temporal structure. An element of indeterminism is present, either because of the uncertainty of the environment or because of the actions of other agents (for simplicity in these examples, at most one other). Here is a brief summary of the main points.

1) An essential feature of the *stit* framework is that it does not refer explicitly to the actions performed by an agent but only to the way an agent's choices (intentional, deliberative but also possibly automatic or unwitting) shape the course of future histories. The result is a very elegant and appealing abstraction which gives a natural denotation for actions whilst doing away with the need to identify and name them. The examples were intended in part to explore how easy it would be to exploit this abstract treatment. In the first few it worked out quite neatly. Here it was possible to identify and describe an agent's actions in terms of transitions of certain kinds between observable states, such as the location of a vase or whether a certain table was level or not. In other examples that does not work out so well. Often it is necessary to refer to the occurrence of a specific kind of action—jolting a table, swerving to the left, kicking an opponent—where the action cannot be picked out by reference to its effects on states. Perhaps dislodging a vase by lifting one end of a table is forbidden but causing it to fall by jolting the table is not. In these cases there seems to be no alternative but to introduce propositional atoms to name specific actions.

2) We very often want to say that the actions of a particular agent are responsible for or the cause of such-and-such in a much weaker sense than is captured by typical *stit* or 'brings it about' constructions. Here it is the 'necessity' condition that is too strong. When an agent lifts a table and a vase standing on it falls and breaks, we want to say that the agent 'broke the vase': it was his actions that were responsible for the falling and the breaking, even though the vase might not have fallen when he lifted the table, and might not have broken when it fell. I looked briefly at two natural candidates for expressing a weaker sense of 'brings it about', which refer to what would, or might, have happened had the agent acted differently. One of these candidates is clearly much too strong (too demanding); the other is much too weak. It is far from clear that there is a way of expressing the required causal relationships using the available resources. I believe this is an important and urgent question because in practice it is precisely these weaker senses of responsibility and 'brings it about' that dominate.

3) Sometimes an agent (human or artificial) behaves in some respects like an automaton, in that in some circumstances it follows a fixed protocol or decision procedure to select its course of action. It might do this unwittingly, as in the case of a reflex, or as a result of a long process of deliberation. Either way it seems very unsatisfactory to model this form of behaviour as if it were a fixed part of the environment in which other agents act. I suggested a simple device for distinguishing between modelling what I called 'actual' and 'explicit protocol' behaviours. I am sure there is much more that can be said about these matters, and about the formal relationships between models of these respective kinds.

4) Finally, some of the key properties of the examples seem to depend critically on the level of detail that is being modelled. For some purposes it is perfectly reasonable to model, say, the moving of a vase from one place to another as an atomic transition with no further structure. For other purposes we might want to look more closely, and model in more detail how the vase is picked up, transported, and set down. For some purposes, we choose to model the movements of agents, physical robotic devices, say, as atomic transitions where unarticulated spatio-temporal constraints make certain combinations of movings impossible. At another level of detail, we model something of what these spatio-temporal constraints are: a doorway is too narrow to allow two agents to pass through simultaneously, there is a single power source which some of the agents have to share, some of the agents are connected by inextensible physical wires, and so on. What we find is that at one level of detail, agent b sees to it that φ, and agent a sees to it that b sees to it that φ. When more detail is added, the same example says that b does not see to it that φ, and perhaps even that a sees to it that b does not see to it that φ. This is disturbing because these are precisely the kinds of properties that we want to examine. Of course there is nothing surprising about the fact that models at different levels of detail have different properties. Some properties are preserved as detail is added and some are not. There is a great deal of work on these matters, for example in the current literature on abstraction methods in model checking. However, the 'stit' and 'brings it about' patterns seem unusually sensitive to choice of detail. I would like to understand better how different models of the same example at different levels of detail relate to one another.

References

Philippe Balbiani, Andreas Herzig, and Nicolas Troquard. Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37(4): 387–406, 2008.

N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. *Theoria*, 54:175–199, 1988. Corrected version in (Belnap and Perloff, 1990).

N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. In H. E. Kyburg, Jr., R. P. Loui, and G. N. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, volume 5 of *Studies in Cognitive Systems*, pages 167–190. Kluwer, Dordrecht, Boston, London, 1990.

Nuel Belnap and Michael Perloff. In the realm of agents. *Annals of Mathematics and Artificial Intelligence*, 9(1–2):25–48, 1993.

Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the future: Agents and choices in our indeterminist world.* Oxford University Press, 2001.

Mark A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17:1–26, 1988.

Robert Craven and Marek Sergot. Agent strands in the action language nC+. *Journal of Applied Logic*, 6(2):172–191, June 2008.

R. Hilpinen. On action and agency. In E. Ejerhed and S. Lindström, editors, *Logic, Action and Cognition—Essays in Philosophical Logic*, volume 2 of *Trends in Logic, Studia Logica Library*, pages 3–27. Kluwer Academic Publishers, Dordrecht, 1997.

J. F. Horty. *Agency and Deontic Logic.* Oxford University Press, 2001.

J. F. Horty and N. Belnap. The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24(6):583–644, 1995.

Ingmar Pörn. *Action Theory and Social Science: Some Formal Models.* Number 120 in Synthese Library. D. Reidel, Dordrecht, 1977.

K. Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51 (3–4):347–378, 1992.

Marek Sergot. Action and agency in norm-governed multi-agent systems. In A. Artikis, G.M.P. O'Hare, K. Stathis, and G. Vouros, editors, *Engineering Societies in the Agents World VIII. 8th Annual International Workshop, ESAW 2007, Athens, October 2007, Revised Selected Papers*, LNCS 4995, pages 1–54. Springer, 2008a.

Marek Sergot. The logic of unwitting collective agency. Technical Report 2008/6, Department of Computing, Imperial College London, 2008b.

Georg Henrik von Wright. *Norm and Action—A Logical Enquiry.* Routledge and Kegan Paul, London, 1963.