The logic of unwitting collective agency

(Technical Report)

Marek Sergot

Department of Computing, Imperial College London
London SW7 2AZ, UK
mjs@doc.ic.ac.uk

May 2008
(Typos corrected June 2009; July 2010)

Abstract

The paper is about the logic of expressions of the form 'agent x brings it about that A is the case', or 'agent x is responsible for its being the case that A', or 'the actions of agent x are the cause of its being the case that A'. Agents could be deliberative (human or computer) agents, purely reactive agents, or simple computational devices. The 'brings it about' modalities are intended to express unintentional, perhaps even accidental, consequences of an agent's actions, as well as possibly intentional (intended) ones. Since we make no assumptions at all about the reasoning or perceptual capabilities of the agents we refer to this form of agency as 'unwitting'; unwitting can mean both inadvertent and unaware. The semantical framework is a form of labelled transition system extended with an extra component that picks out the actions of a particular agent in a transition, or its 'strand' as we call it. We define a modal language for talking about the actions of individual agents or groups of agents in transitions, including two defined modalities of the (unwitting) 'brings it about' kind. The novel feature is the switch of attention from talking about an agent's bringing it about that a certain state of affairs exists to talking about an agent's bringing it about that a transition has a certain property. The middle part of the paper presents axiomatisations of the logic, and comments on relationships to other work, in particular on resemblances to Pörn's (1977) logic of 'brings it about'. The last part is concerned with characterisations of (unwitting) collective agency, that is, the logic of expressions of the form 'the set G of agents, collectively though perhaps unwittingly, brings it about that A'.

1 Introduction

This paper is about the logic of expressions of the form 'agent x brings it about that A is the case', or 'agent x is responsible for its being the case that A', or 'the actions of agent x are the cause of its being the case that A'. The study of logics of this type has a very long tradition. They are sometimes referred to, particularly in the philosophical literature, as logics of action. We will refer to them as logics of agency here, to avoid confusion with other, quite different approaches to the formalisation of action such as those normally encountered in computer science and temporal logic. The best known examples of logics of agency are perhaps the 'stit' ('seeing to it that') family (see e.g. Belnap and Perloff, 1988; Horty and Belnap, 1995; Horty, 2001). 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. As Hilpinen observes: "The expression 'seeing to it that A' usually characterises deliberate, intentional action. 'Bringing it about that A' does not have such a connotation, and can be applied equally well to the unintentional as well as intentional (intended) consequences of one's actions, including highly improbable and accidental consequences." Our agency modalities are of this latter 'brings it about' kind. They are intended to express unintentional, perhaps even unwitting, consequences of an agent's actions, as well as possibly intentional (intended) ones.

Suppose, for example, that two agents a and b are positioned at either end of a table. On the table stands a vase. If one agent lifts its end of the table and the other does not, or if one lowers its end of the table and the other does not, then the table tilts. If the table tilts, the vase falls, and if it falls, it breaks. Suppose now that one agent lifts its end of the table and the other does not, and the vase falls and breaks. Which of the two agents, if either, 'brings about', or is responsible for, the breaking of the vase? It might be the one who lifted its end, or the one who failed to lift its end, or both of them collectively, or neither.

The agents in this example could be humans, or robots with perceptual and reasoning capabilities, or mechanical devices that merely follow a fixed set of rules that make them lift or lower their end of the table in response to certain stimuli. We make no distinction here. An agent can still meaningfully 'bring about' that the vase breaks even if it does so unintentionally, even if it has no way of predicting the other agent's actions, even if it is unaware of the other agent's existence, even if it has no way of detecting that there is a vase on the table. What we want to study here is *unwitting* agency—'unwitting' because that can mean both inadvertent and unaware, and both senses of the word are appropriate here.

This study was initially motivated by issues arising in the norm-governed regulation of multi-agent systems in computer science. An idea that has been gaining popularity in that field is that, in some cases, interactions among multiple, independently acting agents can best be regulated and managed by the use of norms. The term 'social laws' has also been used in this connection, usually with reference to 'artificial social systems'. A 'social law' has been described as a set of obligations and prohibitions on agents' actions, that, if respected, allow multiple, independently acting agents to co-exist in a shared environment.

As argued elsewhere, we want to be able to say that in a system transition representing many concurrent actions by multiple agents and possibly the environment, it is specifically one agent's actions rather than another's that are in compliance or non-compliance with norms governing its behaviour. This allows us in turn to identify and characterise several different categories of non-compliant behaviour. It allows us to distinguish between various forms of unavoidable or inadvertent non-compliance, behaviour where an agent does 'the best that it can' to comply with its individual norms but nevertheless fails to do so because of actions of other agents, and behaviour where an agent could have complied with its individual norms but did not. The aim, amongst other things, is to be able to investigate what kind of system properties emerge if we assume, for instance, that all agents of a certain class will do the best that they can to comply with their individual norms, or never act in such a way that they make non-compliance unavoidable for others.

Generally, the logic of norms and the logic of action/agency have often been studied together. Many authors have argued that an adequate theory of norms must be underpinned by a precise theory of action (by which is often meant agency). This is a feature of von Wright's seminal work (von Wright, 1963), for instance. A more recent example is the extended study by Horty (2001) which uses the framework of 'stit' logics to evaluate how various forms of utilitarianism account for norms governing individual agents and groups of agents. These remarks are for context. We will not address the representation of norms in this paper, except as an occasional source of motivating examples.

Specifically, we begin by presenting a two-sorted (modal) language for talking about properties of states and transitions in a labelled transition system. We then add a component that allows us to pick out the actions of a particular agent in a transition, or its 'strand' of the transition as we will call it. We then extend the language with modalities for talking about the actions of individual agents or groups of agents in transitions, including two defined modalities of the (unwitting) 'brings it about' kind. The novel feature is the switch of attention from talking about an agent's bringing it about that a certain state of affairs exists to talking about an agent's bringing it about that a transition has a certain property. Although the possibility of combining a logic of agency with a transition-based treatment of action has been mentioned from time to time, and elements exist in von Wright's early work and elsewhere, a detailed development has not been done before to our knowledge. The resulting logic resembles Ingmar Porn's (1977) logic of 'brings it about' action/agency, though there are also some very significant differences. The account generalises to talking about the collective actions of groups of agents and their consequences; the last part of the paper is concerned with characterisations of several different forms of (unwitting) collective agency.

It is important to stress that we are making *no assumptions* here about the reasoning or perceptual capabilities of the agents. Agents could be deliberative (human or computer) agents, purely reactive agents, or simple computational devices. We make no distinction between them. This is for both methodological and practical reasons. From the methodological point of view, it is clear that genuine collective or joint action involves a very wide range of issues, including joint intention, communication between agents, awareness of another agent's capabilities and intentions, and many others. We want to factor out all such considerations, and investigate what can be said about individual or collective agency when all such considerations are ignored. The logic of unwitting collective agency might be extended and strengthened in due course by bringing in other factors such as (joint) intention one by one; we do not discuss any such possibilities here. From the practical point of view, there is a wide class of applications for multi-agent systems composed of agents with reasoning and deliberative capabilities. There is an even wider class of applications if we consider also simple 'lightweight' agents with no reasoning capabilities, or systems composed of simple computational units in interaction. We want to be able to consider this wider class of applications too.

Section 2 of the paper presents the two-sorted language used for talking about labelled transition systems. Section 3 introduces agent-stranded transition systems and the language used to talk about an individual agent's actions. Section 4 discusses two defined 'brings it about' modalities, and briefly how they relate to issues that have been discussed in the literature; Section 5 presents axiomatisations of the logic. Section 6 is concerned with the characterisation of (unwitting) collective agency of groups of agents.

2 Labelled transition systems

2.1 Preliminaries

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 $s_0\, \varepsilon_0\, s_1 \cdots s_{m-1}\, \varepsilon_{m-1}\, s_m$ $(m \geq 0)$ such that $(s_{i-1}, \varepsilon_{i-1}, s_i) \in R$ for $i \in 1..m$. 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.

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

A labelled transition system is a structure

$$\langle S, A, R, \mathit{prev}, \mathit{post}, \mathit{label}\rangle$$

where S, R, *prev*, and *post* are as above, and where *label* is a function from R to A. The special case of a LTS in which $R \subseteq S \times A \times S$ then corresponds to the case where $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\mathit{post}(\tau) = \mathit{post}(\tau')$ and $\mathit{label}(\tau) = \mathit{label}(\tau')$ implies $\tau = \tau'$, and in which $\mathit{prev}((s, \varepsilon, s')) = s$, $\mathit{post}((s, \varepsilon, s')) = s'$, and $\mathit{label}((s, \varepsilon, s')) = \varepsilon$. The more abstract account is of little practical significance but is helpful in that it allows a more concise statement of some of the things we want to say about transition systems. It is also more general: transitions are not identified by (s, ε, s') triples: there could be several transitions with the same initial and resulting states and the same label. Nothing in what follows turns on this. Henceforth, we will write $\langle S, A, R\rangle$ as shorthand for $\langle S, A, R, \mathit{prev}, \mathit{post}, \mathit{label}\rangle$ leaving the functions *prev*, *post*, and *label* implicit.

Interpreted transition systems Given a labelled transition system, it is usual to define a language of propositional 'fluents' or 'state variables' in order to express properties of states. Given an LTS $\langle S, A, R\rangle$ and a suitably chosen set of atomic propositions, a model is a structure $\mathcal{M} = \langle S, A, R, h^{\mathrm{f}}\rangle$ where h^{f} is a valuation function which specifies, for every atomic proposition p, the set of states in the LTS at which p is true.

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 events and transitions. Models are structures $\mathcal{M} = \langle S, A, R, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$ 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.

We then extend this two-sorted propositional language with (modal) operators for converting state formulas to transition formulas, and transition formulas to state formulas. Concretely, where φ is a transition formula, the state formula $[\varphi]F$ expresses that the state formula F is satisfied in every state following a transition of type φ. The transition formulas $0{:}F$ and $1{:}G$ are satisfied by a transition τ when the initial state of τ satisfies state formula F and the resulting state of τ satisfies state formula G, respectively. The details are summarised presently.

It is not clear whether evaluating formulas on transitions and states in this fashion is novel or not. Große and Khalil (1996) evaluate formulas on state-event pairs (s, ε) when the transition system is a set of triples (s, ε, s') but that is not the same as we have here. Venema (1999) uses a two-sorted language for expressing properties of points and lines in projective geometry, though naturally the choice of modal operators is different there.

In applications (see e.g. (Craven and Sergot, 2008; Sergot, 2008)) we find it convenient to add a little more structure to the underlying propositional language. It is not essential but makes the formulation of typical examples clearer and more concise. The following is adapted from (Giunchiglia et al., 2004). A *multi-valued propositional signature* σ is a set of symbols called *constants*. For each constant c in σ there is a finite non-empty set $dom(c)$ of values called the domain of c. An atom is an expression of the form $c{=}v$ where c is a constant in σ and $v \in dom(c)$. An interpretation is a function that maps every constant c in σ to some value v in $dom(c)$; an interpretation I satisfies an atom $c{=}v$ if $I(c) = v$. A Boolean constant is one whose domain is the set of truth values $\{\mathsf{t}, \mathsf{f}\}$. As observed in (Giunchiglia et al., 2004), a multi-valued signature of this type can always be translated to an equivalent Boolean signature. The use of a multi-valued signature makes the formulation of examples more concise but since we will not be looking at any in detail in this paper we will not use it. In this paper an expression of the form $c{=}v$ is a propositional atom whose internal structure can be ignored.

2.2 A language for states and transitions

The base propositional language is constructed from a set $\mathcal{P}_{\mathrm{f}}$ of *state atoms* (also known as 'fluents' or 'state variables') and a disjoint set $\mathcal{P}_{\mathrm{a}}$ of *event atoms*. In previous work we followed the terminology of (Giunchiglia et al., 2004) and called the atoms of $\mathcal{P}_{\mathrm{a}}$ 'action atoms'. This terminology is misleading however. Although event atoms *are* used to represent actions and attributes of actions, they are also used to express properties of an event or transition as a whole. Examples of event atoms might be $x{:}\mathit{move}{=}l$ and $x{:}\mathit{move}{=}r$ to represent that agent x moves in direction l and r, respectively. In applications we employ an (informal) convention that event atoms with a prefix 'x:' are intended to represent actions by an agent x. The event atom $a{:}\mathit{lifts}$ might be used to represent that a lifts its end of the table, for example. This is just an informal convention however. The agent prefix does not feature in the semantics. The event atom $\mathit{falls}(\mathit{vase})$ might be used to represent transitions in which the object *vase* falls. Here there is no prefix '*vase*:'—'falls' is not an action that is meaningfully performed by the object *vase*. Event atoms are also used to express properties of an event or a transition as whole: for instance, whether it is desirable or undesirable, timely or untimely, permitted or not permitted, and so on. For this reason we prefer the term 'event atom' for the elements of $\mathcal{P}_\mathrm{a}$, and we reserve the term 'action atom' for referring informally to those event atoms that are intended to represent actions by an agent. In general, an event (transition label) will represent multiple concurrent actions by agents and the environment, concurrent actions such as the falling or breaking of a vase that cannot be ascribed to any agent, and other properties of the event, such as whether it is permitted or not permitted, desirable or undesirable from a system designer's point of view, timely or untimely, and so on. So, for example, the formula

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

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

Propositional formulas of $\mathcal{P}_\mathrm{a}$ are evaluated on transition labels/events. When an event satisfies a propositional formula φ of $\mathcal{P}_\mathrm{a}$ we say that the event is an event of type φ. So, for example, all events of type $a{:}\mathit{lifts} \wedge \neg b{:}\mathit{lifts}$ are also events of type $a{:}\mathit{lifts}$, and events of type $\neg\, b{:}\mathit{lifts}$. By extension, we also say that a transition is of type φ when its label (event) is of type φ. However, there are things we want to say about transitions that are not properties of their events (labels), in particular, whenever we want to refer to what holds in the initial state or final state of the transition. Transition formulas subsume event formulas but are more general. Although evaluating formulas on transitions seems to be unusual, representing events by Boolean compounds of propositional atoms is not so unusual. It is a feature of the action language $\mathcal{C}+$ (Giunchiglia et al., 2004), for example, and has also been used recently in (Sauro et al., 2006) in discussions of agent 'ability'.

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

$$F \ ::= \ \top \mid \bot \mid \text{any atom } p \text{ of } \mathcal{P}_\mathrm{f} \mid \neg F \mid F \wedge F \mid [\varphi]F$$

Transition formulas:

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

where F is any propositional state formula (i.e., a propositional formula of $\mathcal{P}_\mathrm{f}$). We refer to the propositional formulas of $\mathcal{P}_\mathrm{a}$ as *event formulas*.

$\top$ and $\bot$ are 0-ary connectives with the usual interpretation. The other truth-functional connectives (disjunction $\vee$, material implication $\rightarrow$, and bi-implication $\leftrightarrow$) are introduced as abbreviations in the standard manner.

Models Models are structures

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

where h^{f} and h^{a} are the valuation functions for state atoms and event atoms respectively:

$$h^{\mathrm{f}} \colon \mathcal{P}_{\mathrm{f}} \to \wp(S) \qquad \text{and} \qquad h^{\mathrm{a}} \colon \mathcal{P}_{\mathrm{a}} \to \wp(A)$$

For every state s in S and event/label ε in A we have:

$$\begin{array}{lll} \mathcal{M}, s \models p & \text{iff} & s \in h^{\mathrm{f}}(p) \\ \mathcal{M}, \varepsilon \models \alpha & \text{iff} & \varepsilon \in h^{\mathrm{a}}(\alpha) \end{array}$$

and for every transition τ in R and event atom α in $\mathcal{P}_{\mathrm{a}}$:

$$\mathcal{M}, \tau \models \alpha \quad \text{iff} \quad \mathcal{M}, \mathit{label}(\tau) \models \alpha$$

It would be possible to introduce a third sort $\mathcal{P}_R$ of propositional atoms for expressing properties of transitions, different from $\mathcal{P}_{\mathrm{a}}$ though not disjoint. A model would then include a third valuation function $h^R \colon \mathcal{P}_R \to \wp(R)$ with

$$\mathcal{M}, \tau \models \alpha \quad \text{iff} \quad \tau \in h^R(\alpha)$$

We will not bother with that extension here. Atoms in $\mathcal{P}_{\mathrm{a}}$ are evaluated on both event/transition labels and transitions in the present set up. The difference is that event formulas are only the propositional formulas of $\mathcal{P}_{\mathrm{a}}$ whereas transition formulas are more general (as defined above). Transition formulas will be extended with some additional constructs in Section 3.

When φ is a formula of $\mathcal{P}_{\mathrm{a}}$ and τ is a transition in R we say that τ is a transition of type φ when τ satisfies φ, i.e., when $\mathcal{M}, \tau \models \varphi$, and sometimes that φ is true at, or true in, the transition τ. A state s satisfies a formula F when $\mathcal{M}, s \models F$. We sometimes say a formula F 'holds in' state s or 'is true in' state s as alternative ways of saying that s satisfies F.

Semantics Let $\mathcal{M} = \langle S, A, R, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ and let s and τ be a state and transition of $\mathcal{M}$ respectively. The satisfaction definitions for atomic propositions are described above. For negations, conjunctions, and all other truth functional connectives, we take the usual definitions. The satisfaction definitions for the other operators are as follows, for any state formula F and any transition formula φ.

State formulas:

$$\begin{array}{lll} \mathcal{M}, s \models [\varphi]F & \text{iff} & \mathcal{M}, \mathit{post}(\tau) \models F \text{ for every } \tau \in R \text{ such that} \\ & & \mathit{prev}(\tau) = s \text{ and } \mathcal{M}, \tau \models \varphi. \end{array}$$

$\langle\varphi\rangle$ is the dual of $[\varphi]$: $\langle\varphi\rangle F =_{\mathrm{def}} \neg[\varphi]\neg F$.

Transition formulas:

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

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

As usual, we say that 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 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 $\mathcal{M}$ (written $\models F$ and $\models \varphi$, respectively).

Let us discuss the transition formulas first. They are the main focus of attention in this paper.

A transition is of type $0{:}F$ when its initial state satisfies the state formula F, and of type $1{:}G$ when its resulting state satisfies G. The following transition formula represents a transition from a state where (state atom) p holds to a state where it does not:

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

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 as the transition formula:

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

Our notation is more general. We will make some further comments in Section 4.4.

For example, let the state atom $\textit{on-table}(\textit{vase})$ represent that the vase is on the table. A transition of type $0{:}\textit{on-table}(\textit{vase}) \wedge 1{:}\neg\,\textit{on-table}(\textit{vase})$, equivalently, of type $0{:}\textit{on-table}(\textit{vase}) \wedge \neg 1{:}\textit{on-table}(\textit{vase})$ is one from a state in which the vase is on the table to one in which it is not on the table. Suppose that the event atom $\textit{falls}(\textit{vase})$ represents the falling of the vase from the table. A vase-falling transition is also a transition from a state in which the vase is on the table to a state in which the vase is not on the table, and so any LTS model $\mathcal{M}$ modelling this system will have the validity

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

There may be other ways that the vase can get from the table to the ground. One of a or b, or some other agent c, might move the vase from the table to the ground, for example. That would also be a transition of type $0{:}\textit{on-table}(\textit{vase}) \wedge 1{:}\neg\,\textit{on-table}(\textit{vase})$ but not a transition of type $\textit{falls}(\textit{vase})$.

The operators 0: and 1: are both normal. Since *prev* and *post* are (total) functions on R, we have

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

(which also means that 0: and 1: distribute over *all* truth-functional connectives).

Now some brief comments about state formulas. When φ is a transition formula, then $[\varphi]F$ is true at a state s when every transition of type φ results in a state where F is true. So, for example, when α is an event formula, that is, a propositional formula of $\mathcal{P}_\mathrm{a}$, then

$$\langle\alpha\rangle\top \qquad \text{i.e.} \quad \neg[\alpha]\bot$$

represents that an event of type α is executable in the current state. More generally, $\langle\varphi\rangle\top$, i.e., $\neg[\varphi]\bot$, says that there is a transition of type φ from the current state. For example,

$$\langle 1{:}F\rangle\top$$

expresses that there is a transition from the current state to a state where F is true.

$$\langle\varphi \wedge 1{:}F\rangle\top$$

says that there is a transition of type φ from the current state to a state where F is true.

It is important not to confuse the state formula $[\varphi]F$ with the notation $[\varepsilon]F$ used in Propositional Dynamic Logic (PDL). In PDL, the term ε in an expression $[\varepsilon]F$ is a transition label/event ε of A, not a transition *formula* as here. For example, $[0{:}F \wedge \varphi]G$ and $\langle 0{:}F \wedge \varphi \wedge 1{:}G\rangle\top$ are both state formulas. The first is equivalent to $F \to [\varphi]G$ and the second to $F \wedge \langle\varphi\rangle G$.

The logic of each $[\varphi]$ is normal. Moreover:

$$\text{if } \mathcal{M} \models\ \varphi \to \varphi' \quad \text{then} \quad \mathcal{M} \models\ \langle\varphi\rangle F \to \langle\varphi'\rangle F$$

as is easily confirmed, and hence equivalently

$$\text{if } \mathcal{M} \models\ \varphi \to \varphi' \quad \text{then} \quad \mathcal{M} \models\ [\varphi']F \to [\varphi]F$$

We also have validity of:

$$([\varphi]F \wedge [\varphi']F) \to [\varphi \vee \varphi']F$$

and of

$$[\bot]\bot$$

(Sauro et al., 2006) have recently employed a similar device in a logic of agent 'ability' though in a more restricted form than we allow. (Their notation is slightly different.) They give a sound and complete axiomatisation for the logic of expressions $[\alpha]F$ where (in our terms) F is a propositional formula of $\mathcal{P}_\mathrm{f}$ and α is an event formula, that is, a propositional formula of $\mathcal{P}_\mathrm{a}$. We will not attempt to give a complete axiomatisation of our more general language here. It is not essential for the purposes of this paper. We note only that an axiomatisation is more complicated for the more general expressions $[\varphi]F$ because there are some further relationships between state formulas and transition formulas that need to be taken into account. For example, all instances of the following state formulas are obviously valid

$$[1{:}F]F$$

as are all instances of

$$(F \to [\varphi]G) \ \leftrightarrow\ [0{:}F \wedge \varphi]G$$

We will not develop that here. For the purposes of this paper it is transition formulas that are of primary interest.

Generally speaking, we find that properties of labelled transition systems are more easily and clearly expressed as transition formulas rather than state formulas. For example, although we cannot say using a transition formula that in a particular state of $\mathcal{M}$, every transition of type φ leads to a state which satisfies G, we can say (as we often want to) that whenever a state of $\mathcal{M}$ satisfies F, every transition of type φ from that state leads to a state which satisfies G. That is:

$$\mathcal{M} \models (0{:}F \land \varphi) \to 1{:}G$$

Properties of models can often be expressed equivalently as validities of state formulas or of transition formulas. We have

$$\mathcal{M} \models F \to [\varphi]G \quad \text{iff} \quad \mathcal{M} \models (0{:}F \land \varphi) \to 1{:}G$$

Suppose, for example, that the state atoms *light*=*on* and *light*=*off* represent the status of a particular light, and *loc*(x)=p that agent x is at location p. And suppose that the event atoms *toggle* and x:*move* represent that the light switch is toggled and that agent x moves. We would expect the following properties of any model $\mathcal{M}$ modelling this domain:

- state formulas

$$\mathcal{M} \models \mathit{light}{=}\mathit{on} \to [\mathit{toggle}]\, \mathit{light}{=}\mathit{off}$$
$$\mathcal{M} \models \mathit{loc}(x){=}p \to [\neg x{:}\mathit{move}]\, \mathit{loc}(x){=}p$$

- transition formulas

$$\mathcal{M} \models (0{:}\mathit{light}{=}\mathit{on} \land \mathit{toggle}) \to 1{:}\mathit{light}{=}\mathit{off}$$
$$\mathcal{M} \models (0{:}\mathit{loc}(x){=}p \land \neg x{:}\mathit{move}) \to 1{:}\mathit{loc}(x){=}p$$

We find transition formulas are generally more useful and clearer.

Example: Norms and coloured transition systems

A simple way of representing norms is to partition the states (and here transitions) of a transition system $\langle S, A, R\rangle$ into two categories:

- $S_g \subseteq S$, the set of 'permitted' ('acceptable', 'ideal', 'legal') states—we call S_g the 'green' states of the system;
- $R_g \subseteq R$, the set of 'permitted' ('acceptable', 'ideal', 'legal') transitions—we call R_g the 'green' transitions of the system.

We refer to the complements $S_{red} = S \setminus S_g$ and $R_{red} = R \setminus R_g$ as the 'red states' and 'red transitions', respectively. Semantical devices which partition states (and here, transitions) into two categories are familiar in the field of deontic logic. One can find many examples and variations in the literature. In (Sergot and Craven, 2006) we presented a refinement in which the states of a transition systems were ordered depending on how well each complied with a set of explicitly stated norms. We will stick to a simple binary classification in this paper.

We require that a coloured transition system of this type must further satisfy the following constraint, for all states s and s' in S and all transitions τ in R:

$$\text{if } \tau \in R_g \text{ and } \mathit{prev}(\tau) \in S_g \text{ then } \mathit{post}(\tau) \in S_g \tag{1}$$

We refer to this as the *green-green-green* constraint, or *ggg* for short. (It is difficult to find a suitable mnemonic.)

The *ggg* constraint (1) expresses a kind of *well-formedness* principle: a green (permitted, acceptable, legal) transition in a green (permitted, acceptable, legal) state always leads to a green (acceptable, legal, permitted) state. It may be written equivalently as:

$$\text{if } \mathit{prev}(\tau) \in S_g \text{ and } \mathit{post}(\tau) \in S_{red} \text{ then } \tau \in R_{red} \tag{2}$$

Any transition from a green (acceptable, permitted) state to a red (unacceptable, non-permitted) state must itself be undesirable (unacceptable, non-permitted), i.e., 'red', in a well-formed system specification.

Instead of introducing a special category of coloured transition systems with extra components S_g and R_g as in (Sergot and Craven, 2006; Craven and Sergot, 2008), we now prefer to speak of labelled transition systems generally and introduce colourings for states and transitions by means of suitably chosen propositional atoms. Let the state atom *status=green* represent that a state is coloured green, and the event atom *trans=green* that a transition is coloured green. Let *status=red* and *trans=red* be abbreviations for $\neg\,$*status=green* and $\neg\,$*trans=green*, respectively. $\|\mathit{status}{=}\mathit{green}\|^{\mathcal{M}}$ then denotes the 'green states' of a model $\mathcal{M}$ and $\|\mathit{status}{=}\mathit{red}\|^{\mathcal{M}} = S \setminus \|\mathit{status}{=}\mathit{green}\|^{\mathcal{M}}$ its 'red states'; $\|\mathit{trans}{=}\mathit{green}\|^{\mathcal{M}}$ denotes the 'green transitions' and $\|\mathit{trans}{=}\mathit{red}\|^{\mathcal{M}} = R \setminus \|\mathit{trans}{=}\mathit{green}\|^{\mathcal{M}}$ the 'red transitions'.

The *ggg* constraint (1) can then be expressed as validity in the model $\mathcal{M}$ of the state formula

$$\mathit{status}{=}\mathit{green} \rightarrow [\mathit{trans}{=}\mathit{green}]\,\mathit{status}{=}\mathit{green}$$

or, equivalently, of the transition formula

$$(0{:}\mathit{status}{=}\mathit{green} \land \mathit{trans}{=}\mathit{green}) \rightarrow 1{:}\mathit{status}{=}\mathit{green}$$

One can consider a range of other properties that we might require of a coloured transition system. For example: that the transition relation must be serial, that there must be at least one green state, that from every green state there must be at least one green transition, that from every green state reachable from some specified initial state(s) there must be at least one green transition, and so on. With the exception of the last one, all these properties are easily expressed in this language, and can be checked on (a symbolic representation of) a transition system. Reachability properties of a model, such as the last example given above, can be checked but are not expressible as formulas of the language. That could be fixed by extending the language but we will not consider that here.

Notice that we would get much more precision by colouring *paths/runs* of the transition system instead of just its states and transitions. One could then extend the logics presented in this paper with features from a temporal logic such as CTL. The details seem straightforward but we leave them for future investigation.

Example: Agent-specific norms

In the context of using norms or 'social laws' to regulate the interactions of multiple, independently acting agents in a multi-agent computer system, the colourings of states and transitions as 'green' or 'red' represent what in (Craven and Sergot, 2008; Sergot, 2008) we call *system norms*. They express a system designer's point of view of what system states and transitions are legal, permitted, desirable, and so on. There is a separate category of individual *agent-specific* norms that are intended to guide an individual agent's behaviours and are supposed to be taken into account in the agent's implementation, or reasoning processes, in one way or another. These have a different character. In order to be effective, or even meaningful, they must be formulated in terms of what an agent can actually sense or perceive and the actions that it can actually perform. So, in the table-vase example an agent-specific norm could not meaningfully prohibit an agent a from acting in such a way that the vase falls off the table. Agent a may not be able to perceive if the vase is on the table or if the table is tilted. More to the point, agent a cannot predict (we are supposing) how agent b will act, and may even be unaware of agent b's existence. These are aspects of agent-specific norms, and of agent capabilities generally, that are not modelled at the level of detail we are considering in this paper.

That aside, given a transition system $\mathcal{M}$ modelling the possible system behaviours, and some (finite) set Ag of agent names, we specify for every agent x in Ag the norms specific to x that govern x's individual actions: some subset of the transitions will be designated as $green(x)$ and the others as $red(x)$. A transition is designated as $green(x)$ when x's actions in that transition comply with the agent-specific norms for x.

As in the case of coloured transition systems, we prefer to speak of transition systems in general, and use suitably chosen propositional atoms to represent the properties of interest. So, let $\mathcal{P}_{\mathrm{a}}$ contain event atoms $green(x)$ for every agent $x \in Ag$, and let $red(x)$ be an abbreviation for $\neg green(x)$. A transition τ in $\mathcal{M}$ is, or is coloured, $green(x)$, respectively $red(x)$, in a model $\mathcal{M}$ when $\mathcal{M}, \tau \models green(x)$, or $\mathcal{M}, \tau \models red(x)$, respectively. The $green(x)$ transitions in a model $\mathcal{M}$ are $\|green(x)\|^{\mathcal{M}}$; the $red(x)$ transitions are $\|red(x)\|^{\mathcal{M}} = R \setminus \|green(x)\|^{\mathcal{M}}$.

We retain the *ggg* constraint for the colouring of states and transitions as (globally) green or red as determined by the system norms. There is no analogue of the *ggg* constraint for the colourings representing agent-specific norms. However, it is natural to consider an optional *coherence constraint* relating the agent-specific colourings of a transition to its global (system norm) colouring. The colouring of a transition as (globally) red represents that the system as a whole fails to satisfy the required standard of acceptability, legality, desirability represented by the global green/red colouring. In many settings it is then natural to say that if any one of the system components (agents) fails to satisfy its standards of acceptability, legality, desirability, then so does the system as a whole: if a transition is $red(x)$ for some agent x then it is also (globally) red. Formally, given a finite set Ag of agent names, the model $\mathcal{M} = \langle S, A, R, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ satisfies the local-global coherence constraint whenever, for all agents $x \in Ag$, $red(x) \subseteq R_{red}$, that is to say, when

$$\mathcal{M} \models red(x) \rightarrow trans{=}red \tag{3}$$

The coherence constraint (3) is optional and not appropriate in all settings.

Notice though, that even if the coherence constraint is adopted, it is possible that a transition can be coloured $green(x)$ for every agent x and still itself be coloured globally red ($trans{=}red$). (Craven and Sergot, 2008; Sergot, 2008) present examples showing that this is common, and indeed often desirable.

There are other, more fundamental constraints that we must place on agent-specific colourings. We defer discussion of those, and various different categories of norm compliance and non-compliance, until Section 3.

3 Agent-stranded transition systems

The transition systems as they currently stand do not have the capacity to represent that it is specifically one agent's actions rather than another's which must be marked as 'red'. There is no way to extract from, or represent in, the transition system that a particular agent's actions in the transition are illegal, sub-ideal, undesirable, and so on, or more generally, that it is specifically one agent's actions that are responsible for, or the cause of, a transition's having a certain property φ. There is no explicit concept of an individual agent in the semantics at all.

Let Ag be a (finite) set of agent names. An 'agent' in Ag could be a deliberative (human or computer) agent, or it could be a purely reactive component such as a simple computational unit or some other device.

An *agent-stranded LTS* is a structure

$$\langle S, A, R, Ag, strand \rangle$$

where $\langle S, A, R \rangle$ is an LTS. Models are structures $\mathcal{M} = \langle S, A, R, Ag, strand, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ where h^{f} and h^{a} are the valuation functions for the propositional atoms $\mathcal{P}_{\mathrm{f}}$ and $\mathcal{P}_{\mathrm{a}}$, as before.

The new component is $strand$, which is a function on $Ag \times A$. $strand(x, \varepsilon)$ picks out from a transition label/event ε the component or 'strand' that corresponds to agent x's contribution to the event ε. We will write ε_x for $strand(x, \varepsilon)$. For example, where $Ag = \{1, \ldots, n\}$, the transition labels A may, but need not, be tuples

$$A \subseteq A_1 \times \cdots A_i \times \cdots \times A_n \times A_{\mathrm{env}}$$

where each A_i represents the possible actions of the agent i and A_{env} represents possible actions in the environment. Transition labels (events) with this structure are often used in the literature on multi-agent systems and distributed computer systems. In that case, $strand$ would be defined so that

$$(a_1, \ldots, a_i, \ldots, a_n, a_{\mathrm{env}})_i = a_i$$

However, it is not necessary to restrict attention to transition labels/events A of that particular form. All we require is that there is a function $strand$ defined on $Ag \times A$ which picks out unambiguously an agent x's contribution to an event/transition label ε of A. As usual, ε_x may represent several concurrent actions by x, or actions with non-deterministic effects (by which we mean that there could be transitions τ and τ' with $prev(\tau) = prev(\tau')$, $\varepsilon_x = \varepsilon'_x$ where ε and ε' are the labels of τ and τ' respectively, and $post(\tau) \neq post(\tau')$).

Similarly, given a transition τ in R and an agent x in Ag, we can speak of x's strand, τ_x, of the transition τ. Agent x's strand of a transition τ is that of the transition label/event of τ:

$$\tau_x =_{\mathrm{def}} \mathit{strand}(x, \mathit{label}(\tau)\,)$$

τ_x may be thought of as the actions of agent x in the transition τ, where this does *not* imply that τ_x necessarily represents deliberate action, or action which has been freely chosen by x.

Example: agent-specific norms

We assume as before that there are state atoms $\mathit{status}{=}\mathit{green}$ and $\mathit{status}{=}\mathit{red}$ in $\mathcal{P}_{\mathrm{f}}$ for colouring states (globally) green or red, with $\mathit{status}{=}\mathit{red}$ as an abbreviation for $\neg\mathit{status}{=}\mathit{green}$, event atoms $\mathit{trans}{=}\mathit{green}$ and $\mathit{trans}{=}\mathit{red}$ in $\mathcal{P}_{\mathrm{a}}$ for colouring transitions (globally) green or red, with $\mathit{trans}{=}\mathit{red}$ as an abbreviation for $\neg\mathit{trans}{=}\mathit{green}$, and event atoms $\mathit{green}(x)$ and $\mathit{red}(x)$ in $\mathcal{P}_{\mathrm{a}}$ for each agent x in Ag, with $\mathit{red}(x)$ as an abbreviation for $\neg\mathit{green}(x)$.

We impose the *ggg* constraint for the global colourings representing system norms, but not for the colourings representing agent-specific norms. The local-global coherence constraint $\mathcal{M} \models \mathit{red}(x) \to \mathit{red}$ is optional. However, we do impose the following constraint on agent-specific colourings: if τ is a $\mathit{green}(x)$ (resp., $\mathit{red}(x)$) transition from a state s in model $\mathcal{M}$, then every transition τ' from state s in which agent x behaves in the same way as it does in τ must also be $\mathit{green}(x)$ (resp., $\mathit{red}(x)$). In other words, for all transitions τ and τ' in a model $\mathcal{M}$, and all agents $x \in Ag$:

$$\text{if } \mathit{prev}(\tau) = \mathit{prev}(\tau') \text{ and } \tau_x = \tau'_x \text{ then } \mathcal{M}, \tau \models \mathit{green}(x) \text{ iff } \mathcal{M}, \tau' \models \mathit{green}(x) \tag{4}$$

(And hence also $\mathcal{M}, \tau \models \mathit{red}(x)$ iff $\mathcal{M}, \tau' \models \mathit{red}(x)$ whenever $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\tau_x = \tau'_x$.) This reflects the idea that whether actions of agent x are in accordance with the agent-specific norms for x depends only on x's actions, not on the actions of other agents, nor actions in the environment, nor other extraneous factors: we might, with appropriate philosophical caution, think of this constraint as an insistence on the absence of 'moral luck'.

Notice that the constraint (4) covers the case where $\mathit{label}(\tau) = \mathit{label}(\tau')$, that is to say, the case where there are transitions τ and τ' with $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\mathit{label}(\tau) = \mathit{label}(\tau')$ but different resulting states $\mathit{post}(\tau) \neq \mathit{post}(\tau')$: the event $\varepsilon = \mathit{label}(\tau)$ is non-deterministic in the state $s = \mathit{prev}(\tau)$. Constraint (4) requires that, for every agent x, both of these transitions are coloured the same way by agent-specific norms for x. We are not putting this forward as a general principle of morality or ethics. It is a practical matter. The intention is that, in the setting of a multi-agent system of independently acting agents, the agent-specific norms for x are effective in guiding x's actions only if they are formulated in terms of what agent x can actually perceive/sense and the actions it can itself perform. At the level of detail treated here we are not modelling perceptual/sensing capabilities or actions performable by an agent. These features can be added but raise more questions than we have space for here. For now, we insist on the 'absence of moral luck' constraint (4) as a minimal requirement for agent-specific norms.

Example: Some categories of non-compliant behaviour

Craven and Sergot (2008) identify several different categories of non-compliant behaviour that can usefully be distinguished. A $\mathit{red}(x)$ transition τ from a state s is $\mathit{unavoidably\text{-}red}(x)$ if there is no $\mathit{green}(x)$ transition from state s if every agent other than x acts in the same way as it does in τ. In an $\mathit{unavoidably\text{-}red}(x)$ transition the agent x fails to comply with its individual norms but this is because the collective actions of other agents make compliance impossible for x.

We also want to be able to say that in certain cases an agent could have complied with its individual norms but did not. A transition τ from a state s is $\mathit{sub\text{-}standard}(x)$ if the transition is $\mathit{red}(x)$ and, had x acted differently in state s the transition from state s could have been $\mathit{green}(x)$, even if all other agents besides x acted in the same way as they did in τ: x could have acted differently in state s and complied with its individual norms, irrespectively of the actions of other agents.

We might also be interested in behaviours where the actions of one agent make it unavoidable that other agents fail to comply with their agent-specific norms. We now extend the language so that these distinctions, and other properties of transition systems, can be expressed as formulas.

3.1 A modal language for agent strands

Now we introduce a modal language for talking about the agent-specific components of transitions (their 'strands'). We extend the transition formulas of Section 2 with a (unary) operator $[\mathrm{alt}]$, and (unary) operators $[x]$ and $[\backslash x]$ for every agent $x \in \mathit{Ag}$.

Let $\mathcal{M} = \langle S, A, R, \mathit{Ag}, \mathit{strand}, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ be an agent-stranded LTS model.

$$\mathcal{M}, \tau \models [\mathrm{alt}]\varphi \quad \text{iff} \quad \begin{array}{l} \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in R \text{ such that} \\ \mathit{prev}(\tau) = \mathit{prev}(\tau'). \end{array}$$

$\langle \mathrm{alt} \rangle$ is the dual of $[\mathrm{alt}]$.

$[\mathrm{alt}]\varphi$ is satisfied by, or 'true at', a transition τ when all alternative transitions *from the same initial state* as τ satisfy φ. It is for this reason that we choose the notation $[\mathrm{alt}]$ rather than something simpler such as $\Box$. Use of $\Box$ might suggest that we are talking about all transitions in an LTS, and we are not.

$[\mathrm{alt}]$ is a normal modality of type S5. In particular, we have validity (in every agent-stranded LTS) of the following schemas:

$$[\mathrm{alt}]\varphi \rightarrow \varphi$$

$$[\mathrm{alt}]\varphi \rightarrow [\mathrm{alt}][\mathrm{alt}]\varphi$$

$$\neg[\mathrm{alt}]\varphi \rightarrow [\mathrm{alt}]\neg[\mathrm{alt}]\varphi$$

Clearly the following is valid

$$0{:}F \rightarrow [\mathrm{alt}]\, 0{:}F$$

From this follows also:

$$\models \ (0{:}F \rightarrow [\mathrm{alt}]\varphi) \ \leftrightarrow \ [\mathrm{alt}](0{:}F \rightarrow \varphi)$$

and

$$\models \ \langle \mathrm{alt} \rangle (0{:}F \rightarrow \varphi) \ \leftrightarrow \ (0{:}F \rightarrow \langle \mathrm{alt} \rangle \varphi)$$

These are both easily confirmed. It also follows, for example, that

$$0{:}F \to [\text{alt}](\varphi \to 1{:}G)$$

is equivalent to

$$[\text{alt}](0{:}F \land \varphi \to 1{:}G)$$

This is often useful when expressing properties of a model.

Now we add the (unary) operators $[x]$ and $[\backslash x]$ for every agent $x \in Ag$.

$$\begin{array}{lll} \mathcal{M}, \tau \models [x]\varphi & \text{iff} & \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in R \text{ such that } \mathit{prev}(\tau) = \mathit{prev}(\tau') \\ & & \text{and } \tau_x = \tau'_x; \\ \mathcal{M}, \tau \models [\backslash x]\varphi & \text{iff} & \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in R \text{ such that } \mathit{prev}(\tau) = \mathit{prev}(\tau') \\ & & \text{and } \tau_y = \tau'_y \text{ for every } y \in Ag \setminus \{x\}. \end{array}$$

$\langle x \rangle$ and $\langle \backslash x \rangle$ are the respective duals.

As in the case of $[\text{alt}]$, $[x]$ and $[\backslash x]$ are used to talk about properties of alternative transitions from the same initial state: those, respectively, in which x and $Ag \setminus \{x\}$ behave in the same way. We thus have validity of the following:

$$[\text{alt}]\varphi \to [x]\varphi \qquad [\text{alt}]\varphi \to [\backslash x]\varphi$$

We will say, for short, that when $[x]\varphi$ is true at a transition τ, φ is necessary for how x acts in τ; and when $[\backslash x]\varphi$ is true at τ, that φ is necessary for how the agents $Ag \setminus \{x\}$ 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.)

$[x]$ and $[\backslash x]$ are also normal modalities of type S5, so we have validity (in every agent-stranded LTS) of the following schemas:

$$\begin{array}{cc} [x]\varphi \to \varphi & [\backslash x]\varphi \to \varphi \\ [x]\varphi \to [x][x]\varphi & [\backslash x]\varphi \to [\backslash x][\backslash x]\varphi \\ \neg[x]\varphi \to [x]\neg[x]\varphi & \neg[\backslash x]\varphi \to [\backslash x]\neg[\backslash x]\varphi \end{array}$$

It also follows immediately from the satisfaction definitions that the following schema is valid for all pairs of distinct agents $x \neq y$ in Ag:

$$[y]\varphi \to [\backslash x]\varphi \qquad (x \neq y)$$

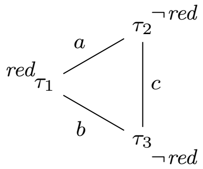
equivalently, as long as Ag is not a singleton, $Ag \neq \{x\}$:

$$\textstyle\bigvee_{y \in Ag \setminus \{x\}} [y]\varphi \to [\backslash x]\varphi \qquad\qquad (Ag \neq \{x\})$$

The other direction is *not* valid:

$$\not\models [\backslash x]\varphi \to \textstyle\bigvee_{y \in Ag \setminus \{x\}} [y]\varphi$$

This is important. Here is a simple example in case it is not obvious.



There are three transitions. a acts the same way in τ_1 and τ_2, b acts the same way in τ_1 and τ_3, and c acts the same way in τ_2 and τ_3. *red* is a propositional atom representing some transition property of interest. It does not matter for present purposes what it represents. Now, $\tau_1 \models [\backslash c]\,\mathit{red}$ but $\tau_1 \not\models [a]\,\mathit{red}$ and $\tau_1 \not\models [b]\,\mathit{red}$.

For the special case of a singleton set of agents $Ag = \{x\}$ we have validity of

$$[\backslash x]\varphi \leftrightarrow [\text{alt}]\varphi \qquad (Ag = \{x\})$$

and hence also of $[\backslash x]\varphi \to [x]\varphi$.

Section 5 is concerned with axiomatisations of this language. It actually simplifies that presentation if we generalise the language to have operators $[G]$ for any subset $G \subseteq Ag$.

Definition. For any model $\mathcal{M}$, any transition τ in $\mathcal{M}$ and any $G \subseteq Ag$:

$$\mathcal{M}, \tau \models [G]\varphi \quad \text{iff} \quad \mathcal{M}, \tau' \models \varphi \text{ for every } \tau' \in R \text{ such that } \mathit{prev}(\tau) = \mathit{prev}(\tau') \text{ and } \tau_y = \tau'_y \text{ for every } y \in G.$$

$\langle G \rangle$ are the respective duals.

$[x]\varphi$ and $[\backslash x]\varphi$ are thus abbreviations for $[\{x\}]\varphi$ and $[Ag\backslash\{x\}]\varphi$. $[\text{alt}]\varphi$ is an abbreviation for $[\emptyset]\varphi$. For the time being we want to illustrate some uses of the language, and in particular the forms of agency that it can express. For that purpose we will restrict attention to the special cases $[x]\varphi$ and $[\backslash x]\varphi$. The more general forms $[G]\varphi$ will be discussed later.

Example

The 'absence of moral luck' property (4) for an agent x with respect to its agent-specific colouring *red*(x) in a model $\mathcal{M}$ is expressed as the validity:

$$\mathcal{M} \models \ \mathit{red}(x) \to [x]\,\mathit{red}(x)$$

Equivalently: $\mathcal{M} \models \ \mathit{green}(x) \to [x]\,\mathit{green}(x)$.

A transition τ in a model $\mathcal{M}$ is *unavoidably-red*(x) when

$$\mathcal{M}, \tau \models [\backslash x]\,\mathit{red}(x)$$

There is one special sub-category of *unavoidably-red*(x) in which *every* transition from a particular state is *red*(x). A transition τ in a model $\mathcal{M}$ is *degenerately-red*(x) when

$$\mathcal{M}, \tau \models [\text{alt}]\,\mathit{red}(x)$$

No well designed set of agent-specific norms should have transitions that are *degenerately-red*(x). A transition τ in a model $\mathcal{M}$ is thus *unavoidably-red*(x) but not *degenerately-red*(x) when

$$\mathcal{M}, \tau \models [\backslash x]\,\mathit{red}(x) \wedge \neg[\text{alt}]\,\mathit{red}(x)$$

What about behaviours where the actions of one agent x make it unavoidable that the other agents fail to comply with their individual norms? That would be $[x][y]\varphi$ or $[x](\bigvee_{y \in Ag\backslash\{x\}} \varphi)$ or $[x][Ag\backslash\{x\}]\varphi$ (that is $[x][\backslash x]\varphi$), depending on which of these we meant, where φ represents some system property of interest. It could be $[x]\,[y]\,\mathit{red}(y)$ or $[x]\,(\bigvee_{y\in Ag\setminus\{x\}} \mathit{red}(y))$ or $[x]\,[Ag\setminus\{x\}]\,(\bigwedge_{y\in Ag\setminus\{x\}} \mathit{red}(y))$ or $[x]\,[Ag\setminus\{x\}]\,(\bigvee_{y\in Ag\setminus\{x\}} \mathit{red}(y))$, for example.

What about that category of non-compliance where an agent x could have complied with its agent-specific norms but did not, or what we called *sub-standard*(x) behaviour earlier?

As defined earlier, a transition τ in a model $\mathcal{M}$ is *sub-standard*(x) when:

(1) the transition is $\mathit{red}(x)$, that is, $\mathcal{M}, \tau \models \mathit{red}(x)$, and

(2) there is a transition τ' in $\mathcal{M}$ such that $\mathit{prev}(\tau') = \mathit{prev}(\tau)$ and $\tau_x \neq \tau'_x$ and $\mathcal{M}, \tau' \models \neg\mathit{red}(x)$, and $\tau_y = \tau'_y$ for all other agents $y \in Ag \setminus \{x\}$.

The form of the second condition (2) suggests that we need to introduce another set of operators for talking about alternative transitions in which an agent x acts differently. We will introduce operators of this type later but we do not need them yet.

For suppose that we have the 'absence of moral luck' property in a model $\mathcal{M}$, that is, the validity $\mathcal{M} \models \mathit{red}(x) \to [x]\,\mathit{red}(x)$. Agent-specific colourings must have this property as the minimal requirement for agent-specific norms of the type we are discussing—*sub-standard*(x) is not meaningful without it. We then have $\mathcal{M} \models \mathit{red}(x) \leftrightarrow [x]\,\mathit{red}(x)$, and this means that a transition τ in a model $\mathcal{M}$ is *sub-standard*(x) when:

(1′) $\mathcal{M}, \tau \models [x]\,\mathit{red}(x)$, and

(2) there is a transition τ' in $\mathcal{M}$ such that $\mathit{prev}(\tau') = \mathit{prev}(\tau)$ and $\tau_x \neq \tau'_x$ and $\mathcal{M}, \tau' \models \neg\mathit{red}(x)$, and $\tau_y = \tau'_y$ for all other agents $y \in Ag \setminus \{x\}$.

Now, condition (1′) allows condition (2) to be simplified: if there is a transition τ' in $\mathcal{M}$ such that $\mathit{prev}(\tau') = \mathit{prev}(\tau)$ and $\mathcal{M}, \tau' \models \neg\mathit{red}(x)$ and $\tau_x = \tau'_x$, then condition (1′) does not hold: $\mathcal{M}, \tau \not\models [x]\,\mathit{red}(x)$. So in condition (2) we can drop the constraint that $\tau_x = \tau'_x$ and state the condition more simply as:

(2′) there is a transition τ' in $\mathcal{M}$ such that $\mathit{prev}(\tau') = \mathit{prev}(\tau)$ and $\mathcal{M}, \tau' \models \neg\mathit{red}(x)$, and $\tau_y = \tau'_y$ for all other agents $y \in Ag \setminus \{x\}$.

Condition (2′) is just $\mathcal{M}, \tau \models \langle\backslash x\rangle\neg\mathit{red}(x)$, or equivalently, $\mathcal{M}, \tau \models \neg[\backslash x]\,\mathit{red}(x)$.

So, a transition τ in a model $\mathcal{M}$ is *sub-standard*(x) when

$$\mathcal{M}, \tau \models [x]\,\mathit{red}(x) \wedge \neg[\backslash x]\,\mathit{red}(x)$$

The simplification of condition (2) is very significant. It will be discussed in more detail and justified more carefully when we look at forms of collective agency in Section 6.

Notice that the 'absence of moral luck' property $\mathcal{M} \models \mathit{red}(x) \leftrightarrow [x]\,\mathit{red}(x)$ means that a transition τ in a model $\mathcal{M}$ is also *sub-standard*(x) when

$$\mathcal{M}, \tau \models \mathit{red}(x) \wedge \neg[\backslash x]\,\mathit{red}(x)$$

that is, when the transition is $\mathit{red}(x)$ but not *unavoidably-red*(x). The 'absence of moral luck' property means that a $\mathit{red}(x)$ transition is either *sub-standard*(x) or *unavoidably-red*(x), but not both.

Hidden in the definition of *sub-standard*(x) is the idea that it is x, rather than some other agent y, who is responsible (perhaps unintentionally or even unwittingly) for the transition's being *red*(x): it is x's actions in the transition that are the cause, unintentional or not, of the transition's being *red*(x). We now make this aspect of *sub-standard*(x) explicit. We do this by looking more generally at expressions of the form

$$[x]\varphi \land \neg[\backslash x]\varphi$$

and

$$[x]\varphi \land \neg[\text{alt}]\varphi$$

Both can be seen as expressing a sense in which agent x 'brings it about that' a transition is of a particular type φ.

4 'Brings it about'

In logics of agency, expressions of the form 'agent x brings it about that φ' 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 concept of *agency*—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. The details vary according to the semantical structures employed, whether intentional or deliberate 'seeing to it that' agency is to be investigated, and so on.

4.1 Two 'brings it about' modalities

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 *prev*(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 *prev*(τ') = *prev*(τ) and $\tau_x \neq \tau'_x$ and $\mathcal{M}, \tau' \models \neg\varphi$.

The form of the second condition, the 'counteraction' condition, suggests that we need to introduce another set of operators for talking about transitions where an agent x acts differently. For every agent $x \in Ag$, let

$$\mathcal{M}, \tau \models [x^*]\varphi \quad \text{iff} \quad \mathcal{M}, \tau' \models \varphi \text{ for all } \tau' \in \mathcal{M} \text{ such that } \mathit{prev}(\tau') = \mathit{prev}(\tau) \text{ and } \tau_x \neq \tau'_x.$$

$\langle x^* \rangle$ are the respective duals.

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

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

The satisfaction conditions for $[x^*]$ are easily stated but a full axiomatisation of their logical properties would present a technical challenge: they can be seen as a species of 'difference modality' (de Rijke, 1992) or, as we will observe later in Section 6, as a special case of Boolean Modal Logic (Gargov and Passy, 1990). They can also be related to the 'window' operator of (van Benthem, 1979). (See also the discussion in (Blackburn et al., 2001, pp419–424, p479).)

However, if our purpose is only to construct the E_x modalities, which it is for the time being, then the same simplification is available as was used when dealing with *sub-standard*(x) in the previous section. The counterfactual condition (2) can be simplified because of the necessity condition (1): if there is a transition τ' in $\mathcal{M}$ such that $prev(\tau') = prev(\tau)$ and $\mathcal{M}, \tau' \models \neg\varphi$ but where $\tau_x = \tau'_x$, then the necessity condition (1) does not hold: $\mathcal{M}, \tau \not\models \varphi$. So for the counteraction condition we can take simply:

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

This is just $\mathcal{M}, \tau \models \langle\text{alt}\rangle\neg\varphi$, or equivalently, $\mathcal{M}, \tau \models \neg[\text{alt}]\varphi$.

The following simpler definition is thus equivalent to the original:

$$\mathrm{E}_x\varphi =_{\text{def}} [x]\varphi \wedge \neg[\text{alt}]\varphi$$

The conjunct $\neg[\text{alt}]\varphi$ says only that the transition might have been of type $\neg\varphi$: it is equivalent to $\langle\text{alt}\rangle\neg\varphi$. But in conjunction with the necessity condition $[x]\varphi$ it can be true at τ only if x acts differently than in τ. What we are saying, in other words, is that the following is valid:

$$[x]\varphi \to ([x^*]\varphi \leftrightarrow [\text{alt}]\varphi) \tag{5}$$

This is easily confirmed. It follows from the validity of

$$[x]\varphi \wedge [x^*]\varphi \leftrightarrow [\text{alt}]\varphi$$

which is also easily confirmed. We will return to these points later, in Section 6. For present purposes, the technically more complicated $[x^*]$ modalities can be ignored: the S5 modalities $[x]$ and $[\text{alt}]$ suffice.

In order to express the form of 'brings it about' agency implicit in *sub-standard*(x), and for other reasons, we need another 'brings it about' modality. We will write it as E^+_x: $\mathrm{E}^+_x\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) φ is necessary for how x acts in τ, $\mathcal{M}, \tau \models [x]\varphi$;

(2) (counteraction) had x acted differently than it did in τ then the transition might have been different *even if all other agents*, besides x, had acted in the same way as they did in τ: there exists a transition τ' in $\mathcal{M}$ such that $prev(\tau') = prev(\tau)$ and $\tau_x \neq \tau'_x$ and $\mathcal{M}, \tau' \models \neg\varphi$ with $\tau_y = \tau'_y$ for all $y \in Ag \setminus \{x\}$.

Again, by the same argument as above, the counteraction condition (2) can be simplified. Because of the necessity condition (1), it can be equivalently stated as:

(2′) there exists a transition τ' in $\mathcal{M}$ such that $prev(\tau') = prev(\tau)$ and $\mathcal{M}, \tau' \models \neg\varphi$ with $\tau_y = \tau'_y$ for all $y \in Ag \setminus \{x\}$.

This is just $\mathcal{M}, \tau \models \langle\backslash x\rangle\neg\varphi$, or equivalently, $\mathcal{M}, \tau \models \neg[\backslash x]\varphi$.

So again, the following simpler definition is equivalent to the original:

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

The simplification of the counteraction conditions is important because it simplifies very significantly the investigation of the logical properties of E_x and E^+_x. It is discussed in more detail in Section 6 when we look at the more general forms of collective agency.

The notation $\mathrm{E}_x \varphi$ is chosen because it bears a strong resemblance to Ingmar Pörn's (1977) logic of 'brings it about'—*except that* in Pörn's logic $\mathrm{E}_x p$ is used to express that agent x brings about the *state of affairs* represented by p. We are using $\mathrm{E}_x \varphi$ to express that x 'brings it about' that a *transition* has the property represented by φ. There are nevertheless some similarities, but also some very significant technical differences, in particular in regard to the semantical structures employed. We will comment further in Section 5.4. Pörn's logic does not have the analogue of $\mathrm{E}^+_x \varphi$.

It is very important not to confuse $[x^*]\varphi$ and $[\backslash x]\varphi$. They are different. The first is talking about alternative transitions in which x acts differently; the second is talking about alternative transitions in which all other agents $Ag \setminus \{x\}$ act the same way. The first is for E_x; the second is for E^+_x. We will have much more to say about this when we look at forms of collective agency in Section 6.

Example: sub-standard behaviours

As originally defined, a transition τ in a model $\mathcal{M}$ is *sub-standard*(x) when x fails to comply with its individual norms but could have complied with its norms even if all the other agents had behaved in the same way. That is when:

$$\mathcal{M}, \tau \models \mathit{red}(x) \land \neg[\backslash x]\,\mathit{red}(x)$$

Equivalently, because of the 'absence of moral luck' property, $\mathcal{M} \models \mathit{red}(x) \to [x]\,\mathit{red}(x)$, a transition τ is *sub-standard*(x) when

$$\mathcal{M}, \tau \models [x]\,\mathit{red}(x) \land \neg[\backslash x]\,\mathit{red}(x)$$

that is, when

$$\mathcal{M}, \tau \models \mathrm{E}^+_x\,\mathit{red}(x)$$

So a transition τ in a model $\mathcal{M}$ is *sub-standard*(x) when x brings it about that the transition is of type $\mathit{red}(x)$, that is, when it is the actions of x that are responsible for, or the cause of, the transition being $\mathit{red}(x)$, irrespectively of actions by any other agents.

What about $\mathrm{E}_x\,\mathit{red}(x)$? What kind of non-compliant behaviour does that express? $\mathrm{E}_x\,\mathit{red}(x)$ is $[x]\,\mathit{red}(x) \land \neg[_{\mathrm{alt}}]\,\mathit{red}(x)$. Assuming the 'absence of moral luck' property for $\mathit{red}(x)$, which we do, this is equivalent to $\mathit{red}(x) \land \neg[_{\mathrm{alt}}]\,\mathit{red}(x)$, which is just $\mathit{red}(x)$ but not *degenerately-red*(x) behaviour.

Other categories of non-compliant behaviours can similarly be expressed and investigated. To take just one example, we might look at $\mathrm{E}^+_x(\mathit{trans}{=}\mathit{red})$ and $\mathrm{E}_x(\mathit{trans}{=}\mathit{red})$ which express two different senses in which an agent x brings it about that a transition is (globally) red. These are not representations of agent-specific norms, but both express properties that might be of interest from the system designer's point of view.

Finally, as one last illustration, we might ask whether it is ever meaningful to talk about *sub-standard*(x) behaviour of an agent y other than x, that is, whether there can be transitions of type $\mathrm{E}_y \mathrm{E}^+_x \mathit{red}(x)$ or $\mathrm{E}^+_y \mathrm{E}^+_x \mathit{red}(x)$ for agents $x \neq y$. Certainly the simpler expressions $\mathrm{E}^+_y \mathit{red}(x)$ and $\mathrm{E}_y \mathit{red}(x)$ are meaningful for pairs of agents $x \neq y$ and may also represent properties of agent-specific colourings/norms that are of interest from the system designer's point of view. But *sub-standard*(x) behaviour of an agent $y \neq x$ is different: it is easy to check (as we will see later) that $\mathrm{E}_y \mathrm{E}^+_x \mathit{red}(x)$ and $\mathrm{E}^+_y \mathrm{E}^+_x \mathit{red}(x)$ are not satisfiable in any model $\mathcal{M}$; both of the following are valid

$$\neg \mathrm{E}_y \mathrm{E}^+_x \mathit{red}(x) \quad \text{and} \quad \neg \mathrm{E}^+_y \mathrm{E}^+_x \mathit{red}(x) \qquad (x \neq y)$$

No agent y can bring about, or be responsible for, a transition's being *sub-standard*(x) other than x itself.

4.2 Discussion

For every agent $x \in \mathit{Ag}$, we have two defined 'brings it about' operators:

$$\begin{aligned} \mathrm{E}_x \varphi &=_{\text{def}} [x]\varphi \wedge \neg[\text{alt}]\varphi \\ \mathrm{E}^+_x \varphi &=_{\text{def}} [x]\varphi \wedge \neg[\backslash x]\varphi \end{aligned}$$

We will not present a full account of their logical properties yet. Our aim in this section is merely to indicate that that their properties are those one would intuitively expect of 'brings it about' modalities, and are broadly in line with what is found in the literature on the logic of agency.

Both $\mathrm{E}_x \varphi$ and $\mathrm{E}^+_x \varphi$ express a sense in which agent x is 'responsible for' or 'brings it about that' (a transition is) φ. Clearly the following is valid:

$$\mathrm{E}^+_x \varphi \rightarrow \mathrm{E}_x \varphi$$

What is the difference? Since $[y]\varphi \rightarrow [\backslash x]\varphi$ is valid for any $x \neq y$, the following is valid

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

and hence also:

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

So $\mathrm{E}^+_x \varphi$ expresses that it is x, and x *alone*, who brings it about that φ. In contrast, $\mathrm{E}_x \varphi$ leaves open the possibility that some other agent $y \neq x$ also brings it about that φ: the conjunction $\mathrm{E}_x \varphi \wedge \mathrm{E}_y \varphi$ can be true even when $x \neq y$.

One might feel uncomfortable with the idea that two distinct agents, acting independently, can both be responsible for 'bringing about' the same thing. But it is easy to find examples. Notice that the conjunction $\mathrm{E}_x \varphi \wedge \mathrm{E}_y \varphi$ is equivalent to

$$[x]\varphi \wedge [y]\varphi \wedge \neg[\text{alt}]\varphi$$

Suppose that two agents are both pushing against a spring-loaded door and thereby keeping it shut. Suppose either one of them is strong enough by itself to keep the door shut. Both are then 'bringing it about' that the door is shut, or rather, that the transition is a 'keeping the door shut' transition. If x pushes, the door remains shut; if y pushes, the door remains shut. But 'keeping the door shut' is not unavoidable; there is a transition, viz., the one in which neither x nor y push, in which the door springs open. It is sufficient that it merely *might* spring open.

The conjunction $\mathrm{E}_x\varphi \wedge \mathrm{E}_y\varphi$ $(x \neq y)$ does *not* represent that x and y are acting in concert, or even that they are aware of each other's existence. We might as well be talking about two blind robots who have got themselves in a position where both are pushing against the same spring-loaded door. Neither can detect the other is there. This is not, and is not intended to be, a representation of genuine collective agency. We will discuss some forms of (unwitting) collective agency in Section 6.

In the same vein, there has been some discussion in the literature on whether the expression 'x brings it about that some other agent y brings it about that' is well formed (see e.g. Belnap and Perloff, 1993). Note that $\mathrm{E}_x\mathrm{E}_y\varphi$ when $x \neq y$ is well formed. We can see that it is, and examples can readily be found to demonstrate that it is meaningful. The 'keeping the door shut' example is easily modified.

As it turns out, the 'transfer of agency' property:

$$\mathrm{E}_x\mathrm{E}_y\varphi \rightarrow \mathrm{E}_x\varphi \tag{6}$$

is valid for E_x. This has also been seen as an undesirable feature of (some forms of) agency. It is perhaps surprising that it is valid here, but that is readily confirmed. Both $\mathrm{E}_x\mathrm{E}_y\varphi \rightarrow [x]\mathrm{E}_y\varphi$ and $[x]\mathrm{E}_y\varphi \rightarrow [x][y]\varphi \wedge [x]\neg[\mathrm{alt}]\varphi$ are clearly valid. Now $[x][y]\varphi \rightarrow [x]\varphi$ is valid (because $[y]\varphi \rightarrow \varphi$ is valid and $[x]$ is normal), and so is $[x]\neg[\mathrm{alt}]\varphi \rightarrow \neg[\mathrm{alt}]\varphi$. So $\mathrm{E}_x\mathrm{E}_y\varphi \rightarrow [x]\varphi \wedge \neg[\mathrm{alt}]\varphi$ is valid. And really there is nothing suspicious about the validity of (6). All it is saying is that if x acts in such a way that it unwittingly brings it about that y unwittingly brings it about that φ, then x also unwittingly brings it about that φ, which seems non-problematic.

What of E^+_x and E^+_y for different x and y? $\mathrm{E}^+_x\mathrm{E}^+_y\varphi$ is syntactically well formed, but it is not meaningful, in the sense that the following is valid (for $x \neq y$):

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

No agent x can by itself bring it about that some other agent y by itself brings something about. Moreover both of the following are also valid (for $x \neq y$):

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

As for 'transfer of (sole) agency', $\mathrm{E}^+_x\mathrm{E}^+_y\varphi \rightarrow \mathrm{E}^+_x\varphi$ *is* valid, but only trivially so: for any $x \neq y$, $\mathrm{E}^+_x\mathrm{E}^+_y\varphi \rightarrow \bot$ is valid, and so therefore, trivially, is $\mathrm{E}^+_x\mathrm{E}^+_y\varphi \rightarrow \mathrm{E}^+_x\varphi$.

What is the difference between E_x and E^+_x? In the case where there is a singleton agent, $Ag = \{x\}$, there is none, because in that case $[\backslash x]\varphi \leftrightarrow [\mathrm{alt}]\varphi$ is valid and therefore so is:

$$\mathrm{E}^+_x\varphi \leftrightarrow \mathrm{E}_x\varphi \qquad (Ag = \{x\})$$

In the case of two or more distinct agents $x \neq y$, we know that $\models \mathrm{E}^+_x\varphi \rightarrow \mathrm{E}_x\varphi$, and $\models \mathrm{E}^+_x\varphi \rightarrow \neg\mathrm{E}^+_y\varphi$. If we compute $\mathrm{E}_x\varphi \wedge \neg\mathrm{E}^+_x\varphi$ we find the following validity:

$$\mathrm{E}_x\varphi \wedge \neg\mathrm{E}^+_x\varphi \;\leftrightarrow\; [x]\varphi \wedge [\backslash x]\varphi \wedge \neg[\mathrm{alt}]\varphi \tag{7}$$

If we have exactly two agents, $Ag = \{x, y\}$, then $[\backslash x]\varphi \leftrightarrow [y]\varphi$ is valid, and then

$$\mathrm{E}_x\varphi \wedge \neg\mathrm{E}^+_x\varphi \;\leftrightarrow\; [x]\varphi \wedge [y]\varphi \wedge \neg[\mathrm{alt}]\varphi$$

is valid. (Cf. the 'keeping the door shut' example.) But that is merely a special case. In general, when there are more than two distinct agents in Ag, $[\backslash x]\varphi$ can be true even if $[y]\varphi$ is not true for any individual agent $y \in Ag \setminus \{x\}$. $[\backslash x]\varphi$ expresses that *between them* the actions of the other agents $Ag \setminus \{x\}$ are such that φ is necessarily true in the transition.

Clearly E_x and E^+_x express a notion of *successful* action: if agent x brings it about that (a transition is of type) φ then it is indeed the case that φ. Or to put it another way (paraphrasing Hilpinen (1997) quoting Chellas (1969)): x can be held responsible for its being the case that φ only if it is the case that φ. E_x and E^+_x are both 'success' operators: both of the following schemes are valid:

$$\mathrm{E}_x\varphi \rightarrow \varphi \qquad \mathrm{E}^+_x\varphi \rightarrow \varphi$$

In axiomatic presentations of logics of agency, the negative 'counteraction' or counterfactual feature of agency, that had x not acted in the way it did then the world would, or might, have been different, is usually reflected (among other things) by axioms that say no agent can bring about what is logically true, or more generally, what was unavoidable anyway. For E_x and E^+_x as defined here, the following are valid

$$\neg\mathrm{E}_x\top \qquad \neg\mathrm{E}^+_x\top$$

and more generally, so are

$$[\mathrm{alt}]\varphi \rightarrow \neg\mathrm{E}_x\varphi \qquad [\mathrm{alt}]\varphi \rightarrow \neg\mathrm{E}^+_x\varphi$$

It follows from the above that (as usual for logics of agency) E_x and E^+_x are not normal modalities. We do not have, for instance, validity of the following

$$\not\models \mathrm{E}_x(\varphi \wedge \varphi') \rightarrow (\mathrm{E}_x\varphi \wedge \mathrm{E}_x\varphi') \qquad \not\models \mathrm{E}^+_x(\varphi \wedge \varphi') \rightarrow (\mathrm{E}^+_x\varphi \wedge \mathrm{E}^+_x\varphi')$$

except in a restricted form. We do have validity of the following schema (often called the schema 'C') for both E_x and E^+_x:

$$\mathrm{E}_x\varphi \wedge \mathrm{E}_x\varphi' \rightarrow \mathrm{E}_x(\varphi \wedge \varphi') \qquad \mathrm{E}^+_x\varphi \wedge \mathrm{E}^+_x\varphi' \rightarrow \mathrm{E}^+_x(\varphi \wedge \varphi')$$

as is easily checked. This property is generally accepted as an intuitively reasonable feature of notions of agency; both E_x and E^+_x have this property. Furthermore, it can be checked that

$$[\mathrm{alt}]\varphi \wedge \mathrm{E}_x\varphi' \rightarrow \mathrm{E}_x(\varphi \wedge \varphi') \qquad [\mathrm{alt}]\varphi \wedge \mathrm{E}^+_x\varphi' \rightarrow \mathrm{E}^+_x(\varphi \wedge \varphi')$$

are valid. These last properties will be of particular interest in Section 4.4.

What about iterations of agency modalities, and other properties? We have already discussed iterations of the form 'x brings it about that y brings it about' for pairs of distinct agents $x \neq y$. We will comment on just one more characteristic feature of agency. First, the expressions $\neg \mathrm{E}_x \varphi$ and $\mathrm{E}_x \neg \mathrm{E}_x \varphi$ are clearly not equivalent. The first expresses merely that x does not bring it about that (a transition is) φ; the second is stronger, and represents a sense in which x *refrains* from bringing it about that (a transition is) φ. $\neg \mathrm{E}_x \varphi \to \mathrm{E}_x \neg \mathrm{E}_x \varphi$ is not valid, though there is a valid restricted form on which we will comment in Section 5.3. $\neg \mathrm{E}_x^+ \varphi \to \mathrm{E}_x \neg \mathrm{E}_x^+ \varphi$ is not valid either.

$\mathrm{E}_x \varphi \to \mathrm{E}_x \mathrm{E}_x \varphi$ is valid. This expresses a fundamental feature of agency, and is surely to be expected in any plausible account: if x brings it about that φ, then x brings it about that x brings it about that φ. In contrast, $\mathrm{E}_x^+ \varphi \to \mathrm{E}_x^+ \mathrm{E}_x^+ \varphi$ is *not* valid. This is rather surprising at first sight, though not if we look more closely at what $\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi$ is saying. $\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi$ depends on a notion of *independence* of x's actions (with respect to φ) from the actions of others. And indeed: we can show

$$\models (\mathrm{E}_x^+ \varphi \to \mathrm{E}_x^+ \mathrm{E}_x^+ \varphi) \leftrightarrow (\neg [\backslash x] \varphi \to [x] \neg [\backslash x] \varphi)$$

$\neg [\backslash x] \varphi \to [x] \neg [\backslash x] \varphi$ expresses a form of independence of x's actions (with respect to φ) from the actions of others: if in a transition τ the actions of the others do not make it unavoidable that φ, then in any other transition from the same state in which x acts in the same way as in τ, the actions of others do not make it unavoidable that φ either. It might be helpful to note that the following is valid[1]:

$$\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi \leftrightarrow [x] \mathrm{E}_x^+ \varphi$$

In applications of the language (see e.g. (Sergot, 2008)) one encounters many instances of transitions at which $\mathrm{E}_x^+ \varphi$ is true but $[x] \mathrm{E}_x^+ \varphi$, and hence $\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi$, is not. We will comment briefly on this feature of E_x^+ after examining the logic.

4.3 Example: 'The others made me do it'

Claims that 'the others made me do it' are common in disputes about the ascription of responsibility. Just for illustration of the language, here are three different senses in which it can be said that 'the others made me do it'.

One possibility:

$$[x]\varphi \wedge [\backslash x]\varphi \wedge \neg [\mathrm{alt}]\varphi \tag{8}$$

This might be read as 'x did φ, but the others $Ag \setminus \{x\}$ between them acted in such a way as to make φ unavoidable'. As discussed above, (8) is equivalent to

$$\mathrm{E}_x \varphi \wedge \neg \mathrm{E}_x^+ \varphi \tag{9}$$

This might be read as saying 'x did φ, but was not solely responsible'.

'The others made me do it': another possibility:

$$[\backslash x]\,[x]\varphi \wedge \neg [\mathrm{alt}]\varphi \tag{10}$$

[1] One half is immediate: validity of $\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi \to [x] \mathrm{E}_x^+ \varphi$ follows immediately from the definition of E_x^+. The other direction is less obvious. Here is one derivation. Notice first that $\models \neg [\backslash x] \mathrm{E}_x^+ \varphi$. This is because both $[\backslash x][x]\varphi \to [\backslash x]\varphi$ and $[\backslash x] \neg [\backslash x]\varphi \to \neg [\backslash x]\varphi$ are valid. Now $\models \mathrm{E}_x^+ \mathrm{E}_x^+ \varphi \leftrightarrow [x] \mathrm{E}_x^+ \varphi \wedge \neg [\backslash x] \mathrm{E}_x^+ \varphi$ by definition, but since $\neg [\backslash x] \mathrm{E}_x^+ \varphi$ is valid, we have the validity of $\mathrm{E}_x^+ \mathrm{E}_x^+ \varphi \leftrightarrow [x] \mathrm{E}_x^+ \varphi$.

We mean by this that between them the others $Ag \setminus \{x\}$ acted in such a way as to make it necessary for what x does that the transition was φ. Again this does not imply any joint action, or even that the agents $Ag \setminus \{x\}$ are aware of each other's existence, or of x's. The second conjunct is because the others did not 'do' φ if there was no alternative for them, or for anyone else. In the case of a singleton set $Ag = \{x\}$ there are no 'others' and the expression (10) is false.

Notice that $\models \neg[_{\mathrm{alt}}]\varphi \rightarrow [\backslash x]\neg[_{\mathrm{alt}}]\varphi$, and indeed that $\neg[_{\mathrm{alt}}]\varphi$ and $[\backslash x]\neg[_{\mathrm{alt}}]\varphi$ are equivalent. So $[\backslash x][x]\varphi \wedge \neg[_{\mathrm{alt}}]\varphi$ is equivalent to $[\backslash x][x]\varphi \wedge [\backslash x]\neg[_{\mathrm{alt}}]\varphi$, which is $[\backslash x]\mathrm{E}_x\varphi$. (10) can thus be expressed equivalently as

$$[\backslash x]\mathrm{E}_x\varphi \tag{11}$$

which is also equivalent to $[\backslash x]\mathrm{E}_x\varphi \wedge \neg[_{\mathrm{alt}}]\varphi$.

Further, $\models [\backslash x][x]\varphi \rightarrow [x]\varphi \wedge [\backslash x]\varphi$. So $\models ([\backslash x][x]\varphi \wedge \neg[_{\mathrm{alt}}]\varphi) \rightarrow ([x]\varphi \wedge [\backslash x]\varphi \wedge \neg[_{\mathrm{alt}}]\varphi)$, i.e., the following is valid:

$$[\backslash x]\mathrm{E}_x\varphi \rightarrow (\mathrm{E}_x\varphi \wedge \neg\mathrm{E}^{+}_x\varphi)$$

In other words, 'the others made me do it' (10)–(11) implies 'the others made me do it' (8)–(9), but not the other way round.

A third possibility would be to say that 'the others made me do it' means that there is some individual agent $y \in Ag \setminus \{x\}$ who brought it about that $\mathrm{E}_x\varphi$, in other words that the following is true:

$$\bigvee_{y \in Ag \setminus \{x\}} \mathrm{E}_y\mathrm{E}_x\varphi \tag{12}$$

Now, $\models \mathrm{E}_y\mathrm{E}_x\varphi \rightarrow [y]\mathrm{E}_x\varphi$ and $\models [y]\mathrm{E}_x\varphi \rightarrow [\backslash x]\mathrm{E}_x\varphi$ $(y \neq x)$. So (12) implies, but is not implied by, (11).

In summary: we can distinguish at least three different senses in which it can be said that 'the others made me do it': the third (12) implies the second (10)–(11) which implies the first (8)–(9).

4.4 Bringing about a state of affairs

$\mathrm{E}_x\varphi$ and $\mathrm{E}^{+}_x\varphi$ represent that x brings it about that a transition is of type φ. This is unusual. Usually, logics of agency do not talk about properties of transitions in this way. What falls in the scope of a 'brings it about' or 'sees to it that' operator is a formula representing a *state of affairs*: an agent 'brings it about' or 'sees to it that' such-and-such a state of affairs exists. How might this sense of 'brings it about' be expressed using the resources of the language presented here?

$\mathrm{E}_x(0{:}F \wedge 1{:}G)$ expresses that x brings about a transition from a state where F holds to one where G holds, and $\mathrm{E}^{+}_x(0{:}F \wedge 1{:}G)$ that x is solely responsible for such a transition. $\mathrm{E}_x 1{:}F$ and $\mathrm{E}^{+}_x 1{:}F$ express that x brings about (resp., solely) that a transition results in a state where F holds. These formulas express *one sense* in which it might be said that x 'brings about' such-and-such a state of affairs F exists. It is not the only sense, because it says that F holds in the state immediately following the transition, whereas we might want to say merely that

F holds at some (unspecified) state in the future. Logics of agency usually do not insist that what is brought about is immediate; indeed, since transitions are not elements of the semantics, references to 'immediate' or the 'next state' are not meaningful. There is one other essential difference: $\mathrm{E}_x 1{:}F$ and $\mathrm{E}^+_x 1{:}F$ are *transition* formulas; they cannot be used to say that in a particular state s, x brings it about that such-and-such a state of affairs F holds.

How might we express that in a given state s, x brings it about that F holds (in some state)? One natural way is to say that every transition from state s is a $\mathrm{E}_x 1{:}F$ transition (resp., $\mathrm{E}^+_x 1{:}F$ transition); more precisely, that there is a $\mathrm{E}_x 1{:}F$ transition (resp., $\mathrm{E}^+_x 1{:}F$ transition) from state s, and that every transition from s is of this type. This is:

$$\mathcal{M}, s \models \langle \mathrm{E}_x 1{:}F \rangle \top \wedge \neg \langle \neg \mathrm{E}_x 1{:}F \rangle \top$$

or equivalently

$$\mathcal{M}, s \models \langle \mathrm{E}_x 1{:}F \rangle \top \wedge [\neg \mathrm{E}_x 1{:}F] \bot$$

This version of 'brings it about that' is not a 'success operator': $\not\models \langle \mathrm{E}_x 1{:}F \rangle \top \wedge [\neg \mathrm{E}_x 1{:}F] \bot \to F$. That is to be expected. But nor do we have (as might be expected) $\models \langle \mathrm{E}_x 1{:}F \rangle \top \wedge [\neg \mathrm{E}_x 1{:}F] \bot \to [\top] F$. (If $[\top] F$ were true in a state s, then $1{:}F$ would be unavoidable, that is, $1{:}F$ and hence $[\mathrm{alt}] 1{:}F$ would be true at every transition from this state; and x could not bring it about that $1{:}F$.) We will not develop these ideas here. As we have remarked already, generally speaking state formulas are awkward to read and manipulate, and we find it much clearer and more useful to say things using transition formulas.

What about $\mathrm{E}_x 0{:}F$ and $\mathrm{E}^+_x 0{:}F$? These are not meaningful: neither is satisfiable in any model $\mathcal{M}$. Clearly, $\models 0{:}F \to [\mathrm{alt}] 0{:}F$, and we have $\models [\mathrm{alt}] \varphi \to \neg \mathrm{E}_x \varphi$. However, $\models [\mathrm{alt}] \varphi \wedge \mathrm{E}_x \varphi' \to \mathrm{E}_x (\varphi \wedge \varphi')$ was noted earlier (and similarly for E^+_x), so the following pair are valid:

$$\begin{aligned} 0{:}F \wedge \mathrm{E}_x 1{:}G &\leftrightarrow \mathrm{E}_x (0{:}F \wedge 1{:}G) \\ 0{:}F \wedge \mathrm{E}^+_x 1{:}G &\leftrightarrow \mathrm{E}^+_x (0{:}F \wedge 1{:}G) \end{aligned}$$

This seems very satisfactory: if in a transition where F holds in the initial state, x brings it about that G holds in the resulting state, then x brings it about that the transition is a transition from a state where F to a state where G, and vice versa.

Now, this observation makes it possible to formalise, in a rather natural way, a suggestion made by von Wright (1968; 1983), Segerberg (1992), and Hilpinen (1997). We will follow the terminology of Hilpinen's version; the others are essentially the same. He sketches an account with two components: first, the idea that actions are associated with transitions between states; and second, to provide the counterfactual 'counteraction' condition required to capture the notion of agency, he distinguishes between transitions corresponding to the agent's activity from transitions corresponding to the agent's inactivity. The latter are transitions where the agent lets 'nature take its own course'. There are then eight possible modes of agency, and because of the symmetry between F and $\neg F$, four basic forms to consider:

- x brings it about that F ($\neg F$ to F, x active);
- x lets it become the case that F ($\neg F$ to F, x inactive);
- x sustains the case that F (F to F, x active);
- x lets it remain the case that F (F to F, x inactive).

The first two correspond to a transition from a state where $\neg F$ to a state where F. The first is a type of bringing about that F by agent x; the second corresponds to inactivity by x (with respect to F)—here the agent x lets nature take its own course. The last two correspond to a transition from a state where F to a state where F. Again, the first of them is a type of bringing about that F by agent x; the second corresponds to inactivity by x (with respect to F).

As discussed by Segerberg and Hilpinen there remain a number of fundamental problems to resolve in this account. Moreover, not discussed by those authors, the picture is considerably more complicated when there are the actions of other agents to take into account and not just the effect of nature's taking its course. However, these distinctions are easily, and rather naturally, expressed in the language we have presented here.

The first ('brings it about that') and third ('sustains the case that') are straightforward: they are

$$\mathrm{E}_x(0{:}\neg F \wedge 1{:}F) \qquad \text{or} \qquad \mathrm{E}^+_x(0{:}\neg F \wedge 1{:}F)$$

and

$$\mathrm{E}_x(0{:}F \wedge 1{:}F) \qquad \text{or} \qquad \mathrm{E}^+_x(0{:}F \wedge 1{:}F)$$

respectively, depending on whether it is x's sole agency that we want to express or not.

The second and fourth cases, where x is inactive, can be expressed as follows

$$(0{:}\neg F \wedge 1{:}F) \wedge \neg\mathrm{E}_x(0{:}\neg F \wedge 1{:}F)$$
$$(0{:}F \wedge 1{:}F) \wedge \neg\mathrm{E}_x(0{:}F \wedge 1{:}F)$$

(Or as above, but with E^+_x in place of E_x.)

It remains to check that these latter expressions do indeed correspond to what Hilpinen was referring to by his term 'inactive'. Whether or not that is the case, other, finer distinctions can be expressed. For example (we do not give an exhaustive exploration of all the possibilities here), supposing that $0{:}\neg F$ is true and that the transition to $1{:}F$ is not unavoidable or inevitable (in other words, that $\neg[\mathrm{alt}]1{:}F$ is true), then we can distinguish:

$$\begin{array}{l} \mathrm{E}^+_x(0{:}\neg F \wedge 1{:}F) \\ \mathrm{E}_x(0{:}\neg F \wedge 1{:}F) \wedge \neg\mathrm{E}^+_x(0{:}\neg F \wedge 1{:}F) \\ 0{:}\neg F \wedge \neg[x]1{:}F \wedge [\backslash x]1{:}F \\ 0{:}\neg F \wedge 1{:}F \wedge \neg[x]1{:}F \wedge \neg[\backslash x]1{:}F \end{array}$$

The reading of the first two is clear. The third and fourth both say that x lets it become the case that F; the first of them says that the other agents between them act in such a way that it becomes the case that F, and the last one that ‘nature takes its own course’. And similarly for the ‘sustains’ and ‘lets it remain’ transitions, i.e., those of type $0{:}F \wedge 1{:}F$.

Note that intuitively x brings it about that F *simpliciter*, $\mathrm{E}_x 1{:}F$, should be equivalent to the disjunction of ‘x brings it about that F’ in Hilpinen’s terminology and ‘x sustains the case that F’. This is easily confirmed:

$$\begin{aligned} \models\ \mathrm{E}_x 1{:}F &\leftrightarrow (0{:}F \vee \neg 0{:}F) \wedge \mathrm{E}_x 1{:}F \\ &\leftrightarrow (0{:}F \wedge \mathrm{E}_x 1{:}F) \vee (\neg 0{:}F \wedge \mathrm{E}_x 1{:}F) \\ &\leftrightarrow \mathrm{E}_x(0{:}F \wedge 1{:}F) \vee \mathrm{E}_x(0{:}\neg F \wedge 1{:}F) \end{aligned}$$

(and likewise for E^+_x).

As an example of some of the things we might want to express using formulas of this kind consider transitions of type $0{:}\mathit{status}{=}\mathit{red} \wedge 1{:}\mathit{status}{=}\mathit{green}$. These correspond to a recovery from a red system state to a green system state. $\mathrm{E}_x(0{:}\mathit{status}{=}\mathit{red} \wedge 1{:}\mathit{status}{=}\mathit{green})$ expresses that agent x brings it about that the system recovers to a green system state, $\mathrm{E}_x(0{:}\mathit{status}{=}\mathit{red} \wedge 1{:}\mathit{status}{=}\mathit{red})$ that agent x sustains the case that the system is in a red state, $\mathrm{E}_x(0{:}\mathit{status}{=}\mathit{green} \wedge 1{:}\mathit{status}{=}\mathit{green})$ that agent x sustains the case that the system is in a green state, $\mathrm{E}_x(0{:}\mathit{status}{=}\mathit{green} \wedge 1{:}\mathit{status}{=}\mathit{red})$ that agent x brings it about, not necessarily by itself, that the system moves from a green state to a red state, and so on for the other categories where x is inactive (x lets it become the case that the system is in a red state, x lets it remain the case that the system is in a red state, and so on). We write E^+_x in place of E_x if we wish to express that x is the sole agent responsible in each case.

4.5 A note on counteraction conditions

We have defined two ‘brings it about’ operators, E_x and E^+_x, which differ only in their counteraction conditions: one is stronger than the other. It is helpful to note, for future reference, that these are the endpoints of a spectrum of ‘brings it about’ operators that could be defined in similar fashion by varying the counteraction conditions. They are, listed in order of decreasing strength:

strongest	$[x]\varphi \wedge \neg[\backslash x]\varphi$	had all the others in $Ag \setminus \{x\}$ acted the same way they did, it might have been otherwise
	$\vdots$	
	$[x]\varphi \wedge \neg[G]\varphi$	had all the others in $G \subseteq Ag \setminus \{x\}$ acted the same way they did, it might have been otherwise
	$[x]\varphi \wedge \neg \bigwedge_{y \in G}[y]\varphi$	had at least one of the others in $G \subseteq Ag \setminus \{x\}$ acted the same way it did, it might have been otherwise
	$\vdots$	
	$[x]\varphi \wedge \neg \bigwedge_{y \in Ag \setminus \{x\}}[y]\varphi$	had at least one of the others in $Ag \setminus \{x\}$ acted the same way it did, it might have been otherwise
weakest	$[x]\varphi \wedge \neg[\textsc{alt}]\varphi$	it might have been otherwise

In each case, the stronger conditions imply the weaker. The strongest is $\mathrm{E}^+_x\varphi$ and the weakest $\mathrm{E}_x\varphi$. The ones in between can all be expressed as formulas of the language (as above) but since they do not seem to be deserving of any special attention we will not name them nor examine their properties. Nevertheless, it will be helpful to refer to this range of possibilities later, when we look at forms of collective agency in Section 6.

5 The logic of unwitting agency

It is obvious that the logic of the modalities $[\mathrm{alt}]$, $[x]$, $[\backslash x]$ and the defined modalities E_x and E^+_x is determined by the fact that $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\tau_x = \tau'_x$ define equivalence relations on the set of transitions R. The obvious abstraction is to consider frames of the form

$$\langle W, \sim, \{\sim_x\}_{x \in Ag}\rangle$$

where $\sim$ and each $\sim_x$ are equivalence relations and where $\sim_x \subseteq \sim$ for every $x \in Ag$. We assume throughout that Ag is a finite and non-empty set (of agent names). The restriction to finite sets could be removed but the extra complication is not worth it.

5.1 Generated transition frames

As usual, we make the nature of the abstraction precise by speaking of the relational ('Kripke') frame that is generated by an agent-stranded LTS. There is a question of what level of generality to aim for at the beginning. For our immediate purposes $\tau \sim \tau'$ represents that τ and τ' are alternative transitions from the same state, and $\tau \sim_x \tau'$ that they are alternatives in which x acted the same way in both. It is natural that they are equivalence relations. However there are other kinds of frames, and other notions of agency, that could be defined on an agent-stranded LTS, and for those different notions the relations may not always be equivalences.

A reasonable compromise for present purposes is this.

Definition Let $\mathcal{T} = \langle S, A, R, Ag, \mathit{strand}\rangle$ be an agent-stranded LTS. The *frame generated by* $\mathcal{T}$, written $F(\mathcal{T})$, is:

$$\langle R, \sim, \{\sim_x\}_{x \in Ag}\rangle$$

and the *generalised frame generated by* $\mathcal{T}$, $F_g(\mathcal{T})$, is:

$$\langle R, \sim, \{R_x\}_{x \in Ag}\rangle$$

where $\sim$ and each R_x and each $\sim_x$ are binary relations on R such that $\tau \sim \tau'$ iff $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ in $\mathcal{T}$, $\tau\, R_x\, \tau'$ iff $\tau_x = \tau'_x$ in $\mathcal{T}$, and $\sim_x = \sim \cap R_x$ for every $x \in Ag$.

This is more general than we need but it adds a bit more flexibility. It simplifies some of what follows, and it leaves open the possibility of introducing another set of operators $\Box_x$ to talk about R_x rather than $\sim_x$. We will not bother with that in this paper.

The model generated by an agent-stranded LTS model $\langle S, A, R, Ag, \mathit{strand}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$ is a little more complicated because we have the transition formulas $0{:}F$ and $1{:}F$ to deal with. However, since we restricted the language so that F in these expressions is always a propositional (Boolean) formula of $\mathcal{P}_{\mathrm{f}}$, we can treat $0{:}F$ and $1{:}F$ simply as propositional atoms to be evaluated on the elements of R. In other words, we have a set of propositional atoms $\mathcal{P}_R =_{\mathrm{def}} \mathcal{P}_{\mathrm{a}} \cup \{0{:}F \mid F \text{ is a propositional formula of } \mathcal{P}_{\mathrm{f}}\} \cup \{1{:}F \mid F \text{ is a propositional formula of } \mathcal{P}_{\mathrm{f}}\}$, and we define the valuation function h^R for the atoms $\mathcal{P}_R$ on elements of R in the obvious way.

Definition Given an agent-stranded LTS $\mathcal{T} = \langle S, A, R, Ag, \mathit{strand}\rangle$ and a model $\mathcal{M} = \langle \mathcal{T}, h^{\mathrm{f}}, h^{\mathrm{a}}\rangle$, the *model generated by* $\mathcal{M}$ is $F(\mathcal{M}) = \langle F(\mathcal{T}), h^R\rangle$ where the valuation function h^R is defined as follows:

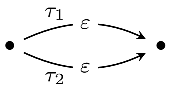
$$\begin{aligned} h^R(\alpha) &=_{\mathrm{def}} h^{\mathrm{a}}(\alpha) \quad \text{when } \alpha \in \mathcal{P}_{\mathrm{a}} \\ h^R(0{:}F) &=_{\mathrm{def}} \{\tau \mid \mathcal{M}, \tau \models 0{:}F\} = \{\tau \mid \mathcal{M}, \mathit{prev}(\tau) \models F\} \\ h^R(1{:}F) &=_{\mathrm{def}} \{\tau \mid \mathcal{M}, \tau \models 1{:}F\} = \{\tau \mid \mathcal{M}, \mathit{post}(\tau) \models F\} \end{aligned}$$

with the additional constraint that

$$\text{if } \tau \sim \tau' \text{ then } \tau \in h^R(0{:}F) \text{ iff } \tau' \in h^R(0{:}F)$$

The constraint is to reflect the validity of $0{:}F \leftrightarrow [{\scriptstyle \mathrm{alt}}]0{:}F$ in LTS models. We also have the validity of $[1{:}F]F$ in those models but since that is a state formula not a transition formula we can ignore it.

Notice that if there are two distinct transitions τ_1 and τ_2 in $\mathcal{T}$ with the same end states and the same label:



then a propositional atom in $\mathcal{P}_R$ is true at τ_1 if and only if it is true at τ_2: if it is an atom of $\mathcal{P}_{\mathrm{a}}$ then it is true at these transitions if only if it is true at their labels, and the labels are the same; if it is of the form $0{:}F$ then it is true at the transitions if and only if it is true in their initial states, and these are the same. And likewise for $1{:}F$.

Definition. Given some (countable) set of propositional atoms $\mathcal{P}_R$ and a non-empty, finite set Ag of agent names, $\mathcal{L}^{Ag}$ is the language with (unary) modal operators $[G]$ for every $G \subseteq Ag$.

Let $\mathcal{M}$ be a model based on the frame $\langle W, \sim, \{R_x\}_{x \in Ag}\rangle$ and let τ be any element of W. Let G be any subset of Ag.

$\mathcal{M}, \tau \models [G]\varphi$ iff $\mathcal{M}, \tau' \models \varphi$ for every τ' such that $\tau \sim \tau'$ and $(\tau, \tau') \in \bigcap_{x \in G} R_x$
iff $\mathcal{M}, \tau' \models \varphi$ for every τ' such that $(\tau, \tau') \in {\sim} \cap \bigcap_{x \in G} \sim_x$ where $\sim_x =_{\mathrm{def}} {\sim} \cap R_x$
iff $\mathcal{M}, \tau' \models \varphi$ for every τ' such that $\tau \sim_G \tau'$ where $\sim_G =_{\mathrm{def}} {\sim} \cap \bigcap_{x \in G} R_x$

$\langle G\rangle$ are the respective duals.

The satisfaction definition is stated in this way for generality. The frames of interest are those where $R_x \subseteq \sim$ for every $x \in Ag$, and hence where $\sim_x = R_x$.

As before, $[x]\varphi$, $[\backslash x]\varphi$, and $[\text{alt}]\varphi$ are abbreviations for $[\{x\}]\varphi$, $[Ag\backslash\{x\}]\varphi$, and $[\emptyset]\varphi$, respectively. We also generalise the 'brings it about' operators. For any subset $G \subseteq Ag$. We define:

$$\mathrm{E}_G\varphi =_{\text{def}} [G]\varphi \wedge \neg[\text{alt}]\varphi$$
$$\mathrm{E}^+_G\varphi =_{\text{def}} [G]\varphi \wedge \neg[Ag\backslash G]\varphi$$

$\mathrm{E}_x\varphi$ and $\mathrm{E}^+_x\varphi$ are abbreviations for $\mathrm{E}_{\{x\}}\varphi$ and $\mathrm{E}^+_{\{x\}}\varphi$, respectively. We will discuss later whether E_G and E^+_G express any useful notion of collective agency of a set of agents G. (They do not.) For the degenerate case $G = \emptyset$, $\mathrm{E}_G\varphi$ and $\mathrm{E}^+_G\varphi$ are both equivalent to $\bot$. Moreover the following is valid:

$$\mathrm{E}^+_G\varphi \leftrightarrow \mathrm{E}_G\varphi \wedge \neg\mathrm{E}_{Ag\backslash G}\varphi$$

This and other properties of E_G and E^+_G will be discussed separately in Section 6.

It is sometimes useful to refer to the relation $\sim_{\backslash x} =_{\text{def}} \sim \cap \bigcap_{y\in Ag\backslash\{x\}} \sim_y$. Then $\mathcal{M}, \tau \models [\backslash x]\varphi$ iff $\mathcal{M}, \tau' \models \varphi$ for every τ' such that $\tau \sim_{\backslash x} \tau'$.

We have to be careful here with the informal reading of these relations, which can be misleading if read too casually. $\sim_x$ is clear enough: it represents the alternative transitions from any given state in which x acts the same way. $\sim_{\backslash x}$ can be read as representing the alternative transitions from any given state in which everyone *except perhaps* x acts the same way. $\sim_{\backslash x}$ is not to be confused with $\sim \backslash \sim_x$.

Now clearly

$$\models [G']\varphi \rightarrow [G]\varphi \quad \text{for every } G' \subseteq G \tag{13}$$

The case $G' = \emptyset$ is just $\models [\emptyset]\varphi \rightarrow [G]\varphi$, which by definition of $[\text{alt}]$ is

$$\models [\text{alt}]\varphi \rightarrow [G]\varphi$$

From (13) it also follows that:

$$\models [G]\varphi \leftrightarrow \bigvee_{G'\subseteq G}[G']\varphi \vee [\text{alt}]\varphi \tag{14}$$

Right-to-left follows from (13); left-to-right is a tautology. It was in order to avoid infinitely long disjunctions that we assumed Ag is a finite (and non-empty) set. This is not an unreasonable assumption for the intended applications. We need the second disjunct because if $G = \emptyset$ then $\models [\emptyset]\varphi \leftrightarrow \bot \vee [\text{alt}]\varphi$. (The empty disjunction is false.)

Theorem 5.1. Let $\mathcal{T}$ be an agent-stranded LTS and let $F(\mathcal{T}) = \langle R, \sim, \{\sim_x\}_{x\in Ag}\rangle$ and $F_g(\mathcal{T}) = \langle R, \sim, \{R_x\}_{x\in Ag}\rangle$ be the frame and the generalised frame generated by $\mathcal{T}$, respectively. The relations $\sim$, R_x, and $\sim_x$ are all equivalence relations on R with $\sim_x \subseteq \sim$. For every $G \subseteq Ag$, $\sim_G =_{\text{def}} \sim \cap \bigcap_{x\in G} \sim_x$ is also an equivalence relation on R with $\sim_G \subseteq \sim$. For $G = \emptyset$, $\sim_G = \sim$.

Proof. Trivial. (The intersection of an empty set of subsets of U is U.) □

Theorem 5.2. Let $\mathcal{T}$ be an agent-stranded LTS and $\mathcal{M} = \langle \mathcal{T}, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ an LTS model based on $\mathcal{T}$. Let $F(\mathcal{T})$ be the frame generated by $\mathcal{T}$ and $\mathcal{F}(\mathcal{M}) = \langle F(\mathcal{T}), h^R \rangle$ the model generated $\mathcal{M}$. Let φ be any formula of $\mathcal{L}^{Ag}$. Then φ is valid in $\mathcal{T}$ iff φ is valid in $F(\mathcal{T})$, and φ is valid in the model $\mathcal{M} = \langle \mathcal{T}, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ iff φ is valid in the model $F(\mathcal{M}) = (F(\mathcal{T}), h^R)$.

Proof. By induction on the structure of φ. The case where φ is of the form $0{:}F$ or $1{:}F$ is discussed above. □

Completing the picture

Given an agent-stranded LTS $\mathcal{T}$, we have checked that any formula is valid in $\mathcal{T}$ iff it is valid in the frame $F(\mathcal{T})$ generated by $\mathcal{T}$. The last step is to show that this particular class of frames provides the right abstraction, in other words, that it does not lose any essential features of an LTS. The usual strategy is to show that every frame in this class is the p-morphic image of a frame generated by some LTS. That is, given a frame $\mathcal{F}$ in this class define an agent-stranded LTS $\mathcal{T}_{\mathcal{F}}$, then show that there is a frame p-morphism from the frame $F(\mathcal{T}_{\mathcal{F}})$ generated by $\mathcal{T}_{\mathcal{F}}$ to $\mathcal{F}$. The rest is just an application of standard results.

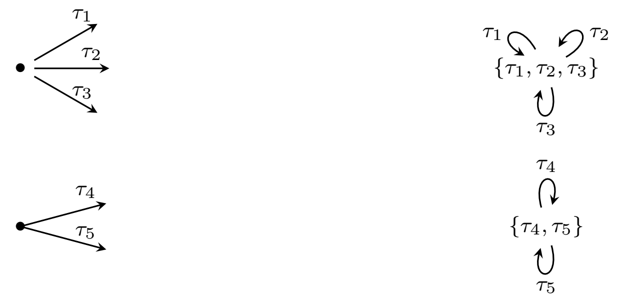
In our case it is simpler. Given a frame $\mathcal{F}$ of the form $\langle W, \sim, \{R_x\}_{x \in Ag} \rangle$ with no assumptions about $\sim$ or R_x other than that they are equivalence relations on W, we can define an LTS $\mathcal{T}_{\mathcal{F}}$ such that there is an isomorphism between $\mathcal{F}$ and the generalised frame $F_g(\mathcal{T}_{\mathcal{F}})$ generated by $\mathcal{T}_{\mathcal{F}}$. That is more than we need.

In what follows we will write $[\tau]^{\sim}$ to denote the equivalence class of $\sim$ to which τ belongs, and similarly for the equivalence classes of R_x.

Definition. Let $\mathcal{F}$ be any frame $\langle W, \sim, \{R_x\}_{x \in Ag} \rangle$ in which $\sim$ and every R_x are equivalence relations on W. The canonical LTS for $\mathcal{F}$, written $\mathcal{T}_{\mathcal{F}}$, is the agent-stranded LTS $\langle S, A, R', Ag, \mathit{strand} \rangle$ in which:

- the set of states S is the set of equivalence classes of $\sim$ in $\mathcal{F}$, $S = W/{\sim}$;
- the set of transition labels A is the set W in $\mathcal{F}$;
- the set of transitions R' is the set of tuples $([\tau]^{\sim}, \tau, [\tau]^{\sim})$ such that $\tau \in W$;
- *strand* is defined so that $\mathit{strand}(x, \tau) =_{\mathrm{def}} [\tau]^{R_x}$, i.e., $\tau_x =_{\mathrm{def}} [\tau]^{R_x}$;
- $\mathit{prev}((s, \tau, s')) = s$; $\mathit{post}((s, \tau, s')) = s'$.

Perhaps a picture will make the construction clearer.



$\langle W, \sim, \{R_x\}_{x \in Ag} \rangle$ $\langle S, A, R', Ag, \mathit{strand} \rangle$

Lemma 5.3. Let $\mathcal{F}$ be any frame $\langle W, \sim, \{R_x\}_{x \in Ag}\rangle$ in which $\sim$ and every R_x are equivalence relations on W. Let $\mathcal{T}_{\mathcal{F}} = \langle S, A, R', Ag, \mathit{strand}\rangle$ be the canonical LTS for $\mathcal{F}$, and let $F_g(\mathcal{T}_F) = \langle R', \sim', \{R'_x\}_{x \in Ag}\rangle$ be the generalised frame generated by $\mathcal{T}_F$. Then we have:

- if $\tau \in W$ then $(s, \tau, s) \in R'$ for $s = [\tau]^{\sim}$;
- if $(s, \tau, s') \in R'$ then $s = s'$.

And further:

- $(s_1, \tau_1, s'_1) \sim' (s_2, \tau_2, s'_2)$ iff $\tau_1 \sim \tau_2$;
- $(s_1, \tau_1, s'_1)\, R'_x\, (s_2, \tau_2, s'_2)$ iff $\tau_1\, R_x\, \tau_2$.

Proof. The first part is just a restatement of the definition. For the second part:

- $(s_1, \tau_1, s'_1) \sim' (s_2, \tau_2, s'_2)$ iff $s_1 = s_2$, but $s_1 = [\tau_1]^{\sim}$ and $s_2 = [\tau_2]^{\sim}$, and $[\tau_1]^{\sim} = [\tau_2]^{\sim}$ iff $\tau_1 \sim \tau_2$.
- $(s_1, \tau_1, s'_1)\, R'_x\, (s_2, \tau_2, s'_2)$ iff $(\tau_1)_x = (\tau_2)_x$ iff $[\tau_1]^{R_x} = [\tau_2]^{R_x}$ iff $\tau_1\, R_x\, \tau_2$.

□

Now we can define an isomorphism p from $F_g(\mathcal{T}_{\mathcal{F}})$ to $\mathcal{F}$: let $p\colon R' \to W$ be a mapping such that $p((s, \tau, s')) = \tau$. By the lemma above, p is a 1-1 correspondence; and further, for any u and v in R', $u \sim' v$ in $F_g(\mathcal{T}_F)$ iff $p(u) \sim p(v)$ in $\mathcal{F}$, and $u\, R'_x\, v$ in $F_g(\mathcal{T}_F)$ iff $p(u)\, R_x\, p(v)$ in $\mathcal{F}$.

We have to complete the picture to include the frame $F(\mathcal{T}_{\mathcal{F}})$ generated by $\mathcal{T}_{\mathcal{F}}$, as shown in the following diagram:

$$\begin{tikzcd}
\langle R', \sim', \{\sim'_x\}_{x \in Ag}\rangle & \langle R', \sim', \{R'_x\}_{x \in Ag}\rangle & \\
F(\mathcal{T}_{\mathcal{F}}) \arrow[d, leftrightarrow] \arrow[dr, phantom, "\mathit{isomorphisms}"] & F_g(\mathcal{T}_{\mathcal{F}}) \arrow[l] \arrow[d, leftrightarrow] & \\
\langle W, \sim, \{\sim_x\}_{x \in Ag}\rangle & \mathcal{F} \arrow[l] \arrow[r] & \mathcal{T}_{\mathcal{F}} \arrow[ul, "\mathit{generates}"'] \\
& \langle W, \sim, \{R_x\}_{x \in Ag}\rangle & \langle S, A, R', Ag, \mathit{strand}\rangle
\end{tikzcd}$$

For the case where the frame $\langle W, \sim, \{R_x\}_{x \in Ag}\rangle$ already has the property that $R_x \subseteq \sim$ for every $x \in Ag$, $\sim_x = \sim \cap R_x = R_x$ and there is nothing more to do. Where it does not have that property, it remains to show that there is an isomorphism between $\langle R', \sim', \{\sim'_x\}_{x \in Ag}\rangle$ and $\langle W, \sim, \{\sim_x\}_{x \in Ag}\rangle$ where every $\sim_x = \sim \cap R_x$. The same mapping p will do: $u \sim'_x v$ iff $u \sim' v$ and $u\, R'_x\, v$, and $p(u) \sim_x p(v)$ iff $p(u) \sim p(v)$ and $p(u)\, R_x\, p(v)$, so $u \sim'_x v$ iff $p(u) \sim_x p(v)$.

Theorem 5.4. Let Ag be a finite, non-empty set of agent names. Any formula φ of $\mathcal{L}^{Ag}$ is valid in all agent-stranded LTSs iff it is valid in the class of frames $\langle W, \sim, \{\sim_x\}_{x \in Ag}\rangle$ such that $\sim$ and every $\sim_x$ are equivalence relations on W with $\sim_x \subseteq \sim$ for every $x \in Ag$.

5.2 Axiomatisations

We are on very familiar territory. For $G \neq \emptyset$, $[G]\varphi$ has exactly the same satisfaction conditions as 'distributed knowledge': if $[x]\varphi$ is read as 'agent x knows that φ', then $[G]\varphi$ is read as 'there is distributed knowledge in group G that φ'. This means that all the well-known results on soundness and completeness (and decidability and complexity) apply equally here.[2] The only small adjustment we need to make is to deal with the case where $G = \emptyset$, which is normally not considered but which arises naturally here in the form of the counteraction condition $\neg[{\scriptstyle\mathrm{alt}}]\varphi$.

Clearly $[G]$ for every $G \subseteq Ag$ is a modality of type S5.

Let Σ be any set of formulas of $\mathcal{L}^{Ag}$. As usual, Σ is a modal logic (of signature $\mathcal{L}^{Ag}$) iff it contains all tautologies of propositional logic and is closed under uniform substitution and modus ponens.

Definition. Let K^{Ag} be the smallest (set inclusion) modal logic of signature $\mathcal{L}^{Ag}$ containing all instances of the following schemas and closed under the following rules RN.$[G]$, for all $G \subseteq Ag$:

$$\begin{array}{rc} \text{RN.}[G] & \dfrac{\varphi}{[G]\varphi} \\[2ex] \text{K.}[G] & [G](\varphi \to \varphi') \to ([G]\varphi \to [G]\varphi') \\[1ex] \text{Sub.}[G] & [G]\varphi \to [G']\varphi \qquad \text{if } G \subseteq G' \\[1ex] \text{Def.E}_G & \mathrm{E}_G\varphi \;\leftrightarrow\; [G']\varphi \land \neg[Ag\backslash G]\varphi \end{array}$$

Let S5^{Ag} be the smallest (set inclusion) modal logic of signature $\mathcal{L}^{Ag}$ containing K^{Ag} and all instances of the following schemas:

$$\begin{array}{rc} \text{T.}[G] & [G]\varphi \to \varphi \\[1ex] \text{4.}[G] & [G]\varphi \to [G][G]\varphi \\[1ex] \text{5.}[G] & \neg[G]\varphi \to [G]\neg[G]\varphi \end{array}$$

Theorem 5.5. *K^{Ag} is sound and complete with respect to the class of frames $\langle W, \sim, \{\sim_x\}_{x \in Ag}\rangle$ in which every $\sim_x \subseteq \sim$.*

S5^{Ag} is sound and complete with respect to the class of such frames in which $\sim$ and every $\sim_x$ are equivalence relations on W.

Proof. The cases for $G \neq \emptyset$ are exactly the same as for 'distributed knowledge' and the proofs can be found in any standard text on epistemic logic. See e.g. (Fagin et al., 1995, Ch. 3, Thm. 3.4.1). The case $G = \emptyset$ is just S5 for $[{\scriptstyle\mathrm{alt}}]$. $[{\scriptstyle\mathrm{alt}}]\varphi \to [x]\varphi$ means that the canonical model has the property $\sim_x \subseteq \sim$. □

[2] By 'knowledge' we are referring here to the idealized form of knowledge sometimes called 'implicit knowledge', or knowledge that can be ascribed to an agent by an external observer. 'Distributed knowledge' is the knowledge that can be ascribed to a group if all its members were to pool their information. In the literature on epistemic logic the notation $\mathrm{E}_G\varphi$ is often used to denote mutual knowledge ('everyone in group G knows that φ'), which is the conjunction of $[x]\varphi$ for all $x \in G$. This is *not* how the E_G modalities in this paper are to be read. In the present context we prefer to emphasise the connections to Pörn's 'brings it about' operator. We will not need the analogue of 'mutual knowledge' in this paper.

The above result can of course be generalised to provide soundness and completeness results for other classes of frames, not just those in which $\sim$ and all $\sim_x$ are equivalence relations. We will not bother recording those. They can be found in any standard text on modal logic.

Theorem 5.6. $\mathrm{S5}^{Ag}$ is sound and complete with respect to the class of agent-stranded LTSs $\langle S, A, R, Ag, \mathit{strand}\rangle$.

Remark: The only one who knows

We have observed that the logic of $[G]$ is the same as that of (distributed) knowledge. It is instructive to look at the non-validity of

$$\mathrm{E}^+_x\varphi \to \mathrm{E}^+_x\mathrm{E}^+_x\varphi$$

from that point of view.

Suppose we read $[x]\varphi$ as 'agent x knows that φ'. $\mathrm{E}^+_x\varphi = [x]\varphi \land \neg[\backslash x]\varphi$ would then express that x is the only one who knows that φ, in the sense that x knows that φ but the other agents $Ag \setminus \{x\}$ do not, and would not know that φ even if they pooled their information: $[\backslash x]\varphi$ is shorthand for $[Ag\backslash\{x\}]\varphi$ which represents in this reading *distributed* knowledge in the set of agents $Ag \setminus \{x\}$. Now, if x is the only one who knows that φ, would we infer that x is the only one who knows that x is the only one who knows that φ? Surely not, and from that point of view, it is not at all surprising that $\mathrm{E}^+_x\varphi \to \mathrm{E}^+_x\mathrm{E}^+_x\varphi$ is not valid. What is more surprising perhaps is that the following is valid:

$$\mathrm{E}^+_x\mathrm{E}^+_x\varphi \leftrightarrow [x]\mathrm{E}^+_x\varphi$$

On the knowledge reading, the right-to-left half is saying that if x knows that x is the only one who knows that φ, then x is the only one who knows that x is the only one who knows that φ. This might seem odd. But suppose some other y, different from x, also knows that x is the only one who knows that φ, in other words, suppose $[y]\mathrm{E}^+_x\varphi$ is true. Then, because $[y]$ is normal (for the idealized form of 'knows' under consideration) and E^+_x is a 'success' operator, it follows that $[y]\varphi$ is true, that is, that y also knows that φ, in contradiction to $\mathrm{E}^+_x\varphi$. So in summary: if x is the only one who knows that φ, this does not imply that x knows that x is the only one who knows that φ. But if x does know that x is the only one who knows that φ, then x is the only one who knows this. This seems quite plausible.

An alternative axiomatisation

In order to expose the properties of E_G and E^+_G, Section 5.3 presents an axiomatisation of K^{Ag} and $\mathrm{S5}^{Ag}$ directly in terms of E_G. It is helpful for that exercise to refer to an alternative axiomatisation of K^{Ag}. It relies only on properties of normal systems found in any standard text on modal logic. (See e.g. (Chellas, 1980, Ch. 4).)

Theorem 5.7. K^{Ag} is the smallest modal logic of signature $\mathcal{L}^{Ag}$ containing all instances of the following schemas and closed under the rules $\mathrm{RN.}[_{\mathrm{alt}}]$ and

RE.$[G]$, where $[\text{alt}]\varphi =_{\text{def}} [\emptyset]\varphi$:

$$\text{RN.}[\text{alt}] \qquad \frac{\varphi}{[\text{alt}]\varphi}$$

$$\text{K.}[\text{alt}] \qquad [\text{alt}](\varphi \to \varphi') \to ([\text{alt}]\varphi \to [\text{alt}]\varphi')$$

$$\text{RE.}[G] \qquad \frac{\varphi \leftrightarrow \varphi'}{[G]\varphi \leftrightarrow [G]\varphi'}$$

$$\text{K}'.[G] \qquad [\text{alt}](\varphi \to \varphi') \to ([G]\varphi \to [G]\varphi')$$

$$\text{C.}[G] \qquad [G]\varphi \wedge [G]\varphi' \to [G](\varphi \wedge \varphi')$$

$$\text{Sub.}[G] \qquad [G']\varphi \to [G]\varphi \qquad \text{if } G' \subseteq G$$

Proof. Deriving K$'$.$[G]$ in K^{Ag} is very easy; the other axioms are trivially in K^{Ag}. It only remains to show that this alternative set of axioms makes every $[G]$ normal. $[\text{alt}]$.$[G]$ and RN.$[\text{alt}]$ gives us $[G]\top$. RN.$[\text{alt}]$ and K$'$.$[G]$ gives us the rule RM.$[G]$:

$$\frac{\varphi \to \varphi'}{[G]\varphi \to [G]\varphi'}$$

which is all we need. (See e.g. (Chellas, 1980, Ch. 4).) □

Notice that the above could be compressed: K.$[\text{alt}]$ is subsumed by K$'$.$[G]$. However, when we move to the axiomatisation in terms of E_G, $[\text{alt}]$ will have special status and it is helpful to anticipate that here.

5.3 Axiomatisations of E_G

E_G and E^+_G are not normal modal operators, and in particular they are not 'closed under logical consequence': if $\models \varphi \to \varphi'$ that does not imply that $\models \text{E}_G\varphi \to \text{E}_G\varphi'$. The reason is that $\neg[\text{alt}]\varphi \wedge [\text{alt}]\varphi'$ could be satisfiable.

The modalities E_G are not normal, but they are very nearly normal, and in fact, very nearly of type S5. It is instructive to construct a complete axiomatisation of S5^{Ag}, and therefore of agent-stranded LTS frames, in terms of E_G and $[\text{alt}]$ directly.

First, let us consider the Sub axioms. If $G' \subseteq G$ then the following is valid, or equivalently, derivable in K^{Ag}:

$$\text{E}_{G'}\varphi \to \text{E}_G\varphi \qquad (G' \subseteq G) \tag{15}$$

The case where $G' = \emptyset$ is simply the tautology $\bot \to \text{E}_G\varphi$.

Notice that the definition of E_G together with $[\text{alt}]\varphi \to [G]\varphi$ gives

$$[G]\varphi \leftrightarrow \text{E}_G\varphi \vee [\text{alt}]\varphi$$

Now let us consider axioms for E_G that will make each $[G]$ normal.

First, we can easily check that the following schema is valid (or derivable in K^{Ag}): $[\text{alt}](\varphi \to \varphi') \wedge \neg[\text{alt}]\varphi' \to (\text{E}_G\varphi \to \text{E}_G\varphi')$. And similarly for E^+_G

and $[Ag \backslash G]$. So we have the following restricted form of 'closure under logical consequence':

$$[\textsc{alt}](\varphi \to \varphi') \land \neg[\textsc{alt}]\varphi' \to (\mathrm{E}_G\varphi \to \mathrm{E}_G\varphi') \tag{16}$$

For E^+_G it is:

$$[\textsc{alt}](\varphi \to \varphi') \land \neg[Ag \backslash G]\varphi' \to (\mathrm{E}^+_G\varphi \to \mathrm{E}^+_G\varphi') \tag{17}$$

(17) can be derived from (16), because $[\textsc{alt}](\varphi \to \varphi') \to [Ag \backslash G](\varphi \to \varphi')$.

As a special case of (16) we also have, for example, the following property:

$$\neg[\textsc{alt}]\varphi \to (\mathrm{E}_G(\varphi \land \varphi') \to \mathrm{E}_G\varphi)$$

and similarly for E^+_G but with $[Ag \backslash G]\varphi$ in place of $[\textsc{alt}]\varphi$.

The schema 'C' which is often taken as a non-controversial property of 'brings it about' operators

$$\mathrm{E}_G\varphi \land \mathrm{E}_G\varphi' \to \mathrm{E}_G(\varphi \land \varphi')$$

is valid for E_G and also for E^+_G:

$$\mathrm{E}^+_G\varphi \land \mathrm{E}^+_G\varphi' \to \mathrm{E}^+_G(\varphi \land \varphi')$$

In fact it can be strengthened; we have validity of the more general pair

$$\mathrm{E}_G\varphi \land [G]\varphi' \to \mathrm{E}_G(\varphi \land \varphi') \tag{18}$$

$$\mathrm{E}^+_G\varphi \land [G]\varphi' \to \mathrm{E}^+_G(\varphi \land \varphi') \tag{19}$$

Here is the derivation[3] of (18):

$$\begin{array}{rlll}
\mathrm{E}_G\varphi \land [G]\varphi' & \to & [G]\varphi \land \neg[\textsc{alt}]\varphi \land [G]\varphi' & \\
& \to & [G](\varphi \land \varphi') \land \neg[\textsc{alt}]\varphi & ([G] \text{ normal}) \\
& \to & [G](\varphi \land \varphi') \land \neg[\textsc{alt}](\varphi \land \varphi') & ([\textsc{alt}] \text{ normal}) \\
& \to & \mathrm{E}_G(\varphi \land \varphi') &
\end{array}$$

The corresponding derivation for E^+_G in (19) works in exactly the same way: we have $\neg[Ag \backslash G]\varphi \to \neg[Ag \backslash G](\varphi \land \varphi')$. (19) can also be derived from (18) using the fact that $[Ag \backslash G]$ is normal.

Notice that from $[\textsc{alt}]\varphi' \to [G]\varphi'$ and (18) and (19) it also follows that

$$\mathrm{E}_G\varphi \land [\textsc{alt}]\varphi' \to \mathrm{E}_G(\varphi \land \varphi')$$

$$\mathrm{E}^+_G\varphi \land [\textsc{alt}]\varphi' \to \mathrm{E}^+_G(\varphi \land \varphi')$$

We made use of this property earlier in Section 4.4 in connection with bringing about and sustaining a state of affairs.

Since $[G]\varphi' \leftrightarrow \mathrm{E}_G\varphi' \lor [\textsc{alt}]\varphi'$, (18) can be expressed equivalently as the pair

$$\mathrm{E}_G\varphi \land \mathrm{E}_G\varphi' \to \mathrm{E}_G(\varphi \land \varphi') \tag{20}$$

$$\mathrm{E}_G\varphi \land [\textsc{alt}]\varphi' \to \mathrm{E}_G(\varphi \land \varphi') \tag{21}$$

[3] Derivations are presented in this schematic form for clarity. The full details are easily reconstructed.

This is useful when we construct the axiomatisation in terms of E_G. We also have the corresponding properties of E^+_G:

$$\mathrm{E}^+_G\varphi \wedge \mathrm{E}_G\varphi' \to \mathrm{E}^+_G(\varphi \wedge \varphi') \tag{22}$$

$$\mathrm{E}^+_G\varphi \wedge [\mathrm{alt}]\varphi' \to \mathrm{E}^+_G(\varphi \wedge \varphi') \tag{23}$$

(23) was noted earlier.

Notice finally that

$$\mathrm{E}^+_G\varphi \wedge \mathrm{E}^+_G\varphi' \to \mathrm{E}^+_G(\varphi \wedge \varphi')$$

is implied by (22), by strengthening of the antecedent: $\mathrm{E}^+_G\varphi' \to \mathrm{E}_G\varphi'$.

Now we have an alternative axiomatisation of K^{Ag} in terms of E_G and $[\mathrm{alt}]$: with $[G]\varphi \leftrightarrow \mathrm{E}_G\varphi \vee [\mathrm{alt}]\varphi$, (16) and (20)–(21) make $[G]$ normal when $[\mathrm{alt}]$ is normal.

Theorem 5.8. *K^{Ag} is the smallest modal logic of signature $\mathcal{L}^{Ag}$ containing all instances of the following schemas and closed under the rules RN.$[\mathrm{alt}]$ and RE.E_G:*

RN.$[\mathrm{alt}]$	$\dfrac{\varphi}{[\mathrm{alt}]\varphi}$
K.$[\mathrm{alt}]$	$[\mathrm{alt}](\varphi \to \varphi') \to ([\mathrm{alt}]\varphi \to [\mathrm{alt}]\varphi')$
RE.E_G	$\dfrac{\varphi \leftrightarrow \varphi'}{\mathrm{E}_G\varphi \leftrightarrow \mathrm{E}_G\varphi'}$
noN.E_G	$[\mathrm{alt}]\varphi \to \neg\mathrm{E}_G\varphi$
K′.E_G	$[\mathrm{alt}](\varphi \to \varphi') \wedge \neg[\mathrm{alt}]\varphi' \to (\mathrm{E}_G\varphi \to \mathrm{E}_G\varphi')$
C.E_G	$\mathrm{E}_G\varphi \wedge \mathrm{E}_G\varphi' \to \mathrm{E}_G(\varphi \wedge \varphi')$
C′.E_G	$\mathrm{E}_G\varphi \wedge [\mathrm{alt}]\varphi' \to \mathrm{E}_G(\varphi \wedge \varphi')$
Sub.E_G	$\mathrm{E}_{G'}\varphi \to \mathrm{E}_G\varphi \qquad \text{if } G' \subseteq G$
Def.$[G]$	$[G]\varphi \leftrightarrow \mathrm{E}_G\varphi \vee [\mathrm{alt}]\varphi$

Proof. One half requires us to show that each of the above listed axioms is sound—either by showing validity of each in the class of frames of interest, or equivalently, by showing that each is derivable in K^{Ag}. The first three are obvious. noN.E_G follows from the definition of E_G. Derivations of all the others were outlined in the preceding discussion.

To complete the proof, we use the axiomatisation of K^{Ag} in Theorem 5.7, and show that each of the axioms listed there is derivable from the axioms above. RE.$[G]$ follows easily from RE.E_G and Def.$[G]$. Def.E_G is derived from Def.$[G]$ and noN.E_G. Sub.$[G]$ follows from Sub.E_G and Def.$[G]$.

This leaves K′.$[G]$ and C.$[G]$. K′.$[G]$ can be derived from K′.$[G]$, Def.$[G]$ and K.$[\mathrm{alt}]$. C.$[G]$ is easily derived from C.E_G and C′.E_G using Def.$[G]$ and the fact that $[\mathrm{alt}]$ is normal. □

Now for the S5 properties. The axiomatisation can be stated quite succinctly, as in Theorem 5.10 below. However, the results are easily generalised to other classes of frames, those where the relations $\sim_x$ are reflexive, transitive, reflexive and transitive, euclidean, and so on. Indeed, if we assume only that $\sim$ is an equivalence relation (and even that assumption can be relaxed) then the corresponding axioms for $[G]$ and E_G are in fact equivalences (theorems of K^{Ag}). The only slight complication is the transitivity axiom $\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi$. Although this implies $[G]\varphi \to [G][G]\varphi$ in K^{Ag} the converse does not hold.

Theorem 5.9. All instances of the following schemas are theorems of K^{Ag}.

$$\begin{array}{ll}
\text{(T)} & ([G]\varphi \to \varphi) \;\leftrightarrow\; (\mathrm{E}_G\varphi \to \varphi) \\
\text{(4a)} & ([\mathrm{alt}][G]\varphi \to [\mathrm{alt}]\varphi) \to \\
 & \qquad (\,([G]\varphi \to [G][G]\varphi) \to (\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi)\,) \\
\text{(4b)} & (\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi) \to ([G]\varphi \to [G][G]\varphi) \\
\text{(5)} & (\neg[G]\varphi \to [G]\neg[G]\varphi) \;\leftrightarrow\; (\neg\mathrm{E}_G\varphi \to [G]\neg\mathrm{E}_G\varphi)
\end{array}$$

Proof. These are all straightforward derivations in K^{Ag}. See Appendix A. □

Notice (4a). $\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi$ is not valid in the class of transitive frames, but it is valid in the class of reflexive, transitive frames. (4a) is a statement of a more general property.

Theorem 5.10. $\mathrm{S5}^{Ag}$ is the smallest K^{Ag} system containing all instances of the following schemas:

$$\begin{array}{ll}
\mathrm{T.E}_G & \mathrm{E}_G\varphi \to \varphi \\
\mathrm{4.E}_G & \mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi \\
5'.\mathrm{E}_G & \neg\mathrm{E}_G\varphi \to [G]\neg\mathrm{E}_G\varphi
\end{array}$$

Proof. From Theorem 5.9. Since all instances of $[G]\varphi \to \varphi$ are theorems of $\mathrm{S5}^{Ag}$, so are $[\mathrm{alt}]([G]\varphi \to \varphi)$ and $[\mathrm{alt}][G]\varphi \to [\mathrm{alt}]\varphi$, which is the antecedent for the property 4(a) of Theorem 5.9. □

We conclude with a comment on axiom $5'.\mathrm{E_G}$. Because $[G]\varphi \leftrightarrow \mathrm{E}_G\varphi \vee [\mathrm{alt}]\varphi$, $5'.\mathrm{E_G}$ can be written equivalently as:

$$\neg[\mathrm{alt}]\neg\mathrm{E}_G\varphi \to (\neg\mathrm{E}_G\varphi \to \mathrm{E}_G\neg\mathrm{E}_G\varphi) \tag{24}$$

It might appear that this says that, unless $\neg\mathrm{E}_G\varphi$ is unavoidable, then not bringing about φ is equivalent to *refraining* from bringing about φ. That seems wrong. It is plain that simply not doing something is not the same as refraining from doing it. However, 'refraining' has a very strong connotation of intentional (intended) action. $5'.\mathrm{E_G}$ says only that if G acts in such a way that it does not, unwittingly, bring about φ, then G also, unwittingly, brings about that it does not, unwittingly, bring about φ. *Unwittingly* refraining from something, if it means anything, surely means the same as simply not doing it. From that point of view, $5'.\mathrm{E_G}$ seems perfectly reasonable.

5.4 Comparison with Pörn's (1977) logic of action

Readers familiar with Ingmar Pörn's logic of 'brings it about' may have been surprised by earlier remarks that there is a strong resemblance between our operator E_x and Pörn's. The constructions are similar but there are also some very significant differences, not least because Pörn's account is much more abstract than the transition-based one developed here. We summarise the presentation in Pörn's book (1977), which also provides useful references to earlier work along the same lines by Chellas, Kanger, Hilpinen, and others. The main elements of the logic can also be found in (Pörn, 1974).

Pörn's logic has two normal operators:

$\mathrm{D}_x p$	'it is necessary for something which x does that p'
$\mathrm{D}'_x p$	'but for x's action it would not be the case that p', or 'p is dependent on x's action'

and a defined (non-normal) operator

$$\mathrm{E}^{\mathrm{P}}_x p \ =_{\mathrm{def}} \ \mathrm{D}_x p \land \neg \mathrm{D}'_x \neg p$$

Here p represents a state of affairs: that a window is open, that a vase is broken, that agent x is at location l, that a system is in a 'red' system state, and so on.

The counteraction condition $\neg \mathrm{D}'_x \neg p$ in the definition of $\mathrm{E}^{\mathrm{P}}_x p$ may be read as 'but for x's activity it might not be the case that p' or 'p is not independent of x's action'.

There is also a defined (normal) operator

$$\mathrm{N}_x p \ =_{\mathrm{def}} \ \mathrm{D}_x p \land \mathrm{D}'_x \neg p$$

which may be read as 'it is unavoidable for x that p'.

Notice that from these definitions we have

$$\models \ \mathrm{E}^{\mathrm{P}}_x p \leftrightarrow \mathrm{D}_x p \land \neg \mathrm{N}_x p$$

The resemblance to the E_x operator of this paper is even clearer if we write $\mathrm{D}^*_x p$ for $\mathrm{D}'_x \neg p$. $\mathrm{D}^*_x p$ may be read as 'had x acted differently, it would be the case that p'. Then:

$$\begin{aligned} \mathrm{E}^{\mathrm{P}}_x p \ &=_{\mathrm{def}} \ \mathrm{D}_x p \land \neg \mathrm{D}^*_x p \\ \mathrm{N}_x p \ &=_{\mathrm{def}} \ \mathrm{D}_x p \land \mathrm{D}^*_x p \\ \mathrm{E}^{\mathrm{P}}_x p \ &\leftrightarrow \ \mathrm{D}_x p \land \neg \mathrm{N}_x p \end{aligned}$$

However we should not jump to the assumption that D_x and D^*_x correspond exactly to what we write as $[x]$ and $[x^*]$ in this paper. The difference is in the semantics.

Pörn considers frames of the form

$$\langle W, R_x, R^*_x \rangle$$

(his notation is different) where W is a set of possible worlds or 'situations'. D_x and D'_x are interpreted on R_x and R^*_x, respectively:

$$\begin{array}{lll} \mathcal{M}, u \models \mathrm{D}_x p & \text{iff} & \mathcal{M}, v \models p \ \text{ for each } v \text{ such that } (u,v) \in R_x \\ \mathcal{M}, u \models \mathrm{D}'_x p & \text{iff} & \mathcal{M}, v \models \neg p \ \text{ for each } v \text{ such that } (u,v) \in R^*_x \end{array}$$

that is

$$\mathcal{M}, u \models \mathrm{D}^*_x p \quad \text{iff} \quad \mathcal{M}, v \models p \ \text{ for each } v \text{ such that } (u,v) \in R^*_x$$

For R_x, the reading suggested is that $(u,v) \in R_x$ when x 'does in v at least as much as he does in u', or when 'everything x does in u is the case in v' (Segerberg, 1992). R_x is reflexive and transitive.

This semantics is rather abstract however, and it is not entirely clear how phrases like 'does at least as much' are to be understood, or indeed, why an agent's doing at least as much in one situation as in another is a useful concept. Not doing something is also a kind of action or activity. If an agent lifts a table-end in one situation v and does not lift it in another situation u, but in every other respect acts in exactly the same way in u as it does in v (assuming that is possible, which is itself far from obvious) then presumably it does in v at least as much as it does in u. The consequences of its actions in u are quite different from the consequences of its actions in v, however. If an agent acts in the same way in situations u and v, except that in v it moves to the left and in u it does not move at all, then (presumably) it does in v as least as much as it does in u; but again, the consequences of its actions in u and in v could be quite different. We should be careful, however, not to take the suggested readings too literally. Reflexivity and transitivity are referred to by Pörn as the minimal assumptions one could make about the properties of R_x; the readings suggested for R_x are perhaps merely indicative of possible ways that R_x could be understood. We do not want to speculate here about what Pörn did or did not intend R_x to represent. But we can surely say this. Let $u \sim_x v$ represent that agent x acts identically in situations u and v. If x acts identically in u and in v then surely x does in v at least as much as he does in u: $\sim_x \subseteq R_x$. And let $u \sim v$ represent simply that v is an alternative situation to u, irrespectively of how x (or any other agent) acts in u and v. Then:

$$\sim_x \subseteq R_x \subseteq \sim \tag{25}$$

The intended reading of R^*_x is rather more obscure. Pörn suggests: $(u,v) \in R^*_x$ when x 'does not do in v any of the things he does in u', or 'the opposite of everything that x does in u is the case in v'. Again, we should be very careful not to read these expressions too literally. R^*_x is irreflexive (and serial). The only other clue to what Pörn has in mind is that he imposes the following condition on R_x and R^*_x:

(OM7) if $(u, v_1) \in R_x$ and $(u, v_2) \in R_x$ then $(v_1, w) \in R^*_x$ if and only if $(v_2, w) \in R^*_x$

Since R_x is reflexive, (OM7) implies (but is not implied by) two further conditions (labelled as in (Pörn, 1977)):

(OM7.1) if $(u,v) \in R_x$ and $(v,w) \in R^*_x$ then $(u,w) \in R^*_x$

(OM7.2) if $(u,v) \in R_x$ and $(u,w) \in R^*_x$ then $(v,w) \in R^*_x$

Note that by (OM7.2):

$$\text{if } (u,v) \in R_x \text{ and } (u,v) \in R^*_x \text{ then } (v,v) \in R^*_x$$

Since R^*_x is irreflexive, this means that $R_x \cap R^*_x = \emptyset$.

The phrase 'does not do in v any of the things he does in u' suggests that $(u,v) \in R^*_x$ might mean simply 'acts differently in u and v', i.e., $R^*_x = {\sim} \setminus {\sim_x}$. But that cannot be: R_x and R^*_x would then not be disjoint, and would fail to satisfy the conditions (OM7.2) and (OM7), unless $R_x = {\sim_x}$. But we can surely say this: if x 'does not do in v any of the things he does in u', or 'does the opposite' in v and u, then certainly he does not do the same in u and v: $R^*_x \subseteq {\sim} \setminus {\sim_x}$. And so:

$$R^*_x \subseteq {\sim} \setminus {\sim_x} \subseteq {\sim} \tag{26}$$

With $[\mathrm{alt}]$, $[x]$ and $[x^*]$ defined as in previous sections, ${\sim_x} \subseteq R_x \subseteq {\sim}$ gives

$$\models\ [\mathrm{alt}]p \to \mathrm{D}_X p \qquad \text{and} \qquad \models \mathrm{D}_X p \to [x]p$$

and $R^*_x \subseteq {\sim} \setminus {\sim_x} \subseteq {\sim}$ gives

$$\models\ [\mathrm{alt}]p \to [x^*]p \qquad \text{and} \qquad \models [x^*]p \to \mathrm{D}^*_X p$$

Now: $\models\ [x]p \wedge [x^*]p \to \mathrm{D}_X p \wedge \mathrm{D}^*_X p$, and hence

$$\models\ [\mathrm{alt}]\varphi \to \mathrm{N}_X p$$

from which follows:

$$\models\ \mathrm{E}^{\mathrm{P}}_X p \to \mathrm{E}_X p$$

This is as much as we want to say. The question of how $\mathrm{E}^{\mathrm{P}}_X p$ relates to $\mathrm{E}^{+}_X p$ is not meaningful since there is no analogue of E^{+}_X in Pörn's logic.

Note that the soundness and completeness results of Section 5.3 are not applicable to Pörn's logic without some further modification. They depend on the assumption that $\sim$ is an equivalence relation ($[\mathrm{alt}]$ is of type S5). Even if we add that assumption here, N_X is not of type S5: $R_x \cup R^*_x$ is not necessarily an equivalence relation (and not even transitive in general).

6 Unwitting collective agency

Consider again the table-vase example from the introductory section. To recap: a vase stands on a table at whose ends are positioned two agents a and b. Each can lift or lower its end. If one lifts and the other does not, or if one lowers its end and the other does not, the table tilts and the vase falls and breaks.

Consider a transition in which a lifts and b does not and the vase falls and breaks. Neither of the agents a or b individually brings it about that the vase falls. The falling of the vase is not necessary for how a acts in this transition, and it is not necessary for how b acts in this transition. Yet intuitively it seems right to say that the agents a and b collectively, though perhaps unwittingly, bring it about that the vase falls. The falling is necessary for how they act collectively, and the falling is not unavoidable, for there are alternative transitions, where a does not lift or where b also lifts, in which the vase does not fall. If we add another agent c into the picture, an agent whose actions cannot affect the table or the vase, or interfere in any way with the actions by a and b, it also seems right to say that in the transition where a lifts and b does not, it is the set $\{a, b\}$ of agents that brings it about that the vase falls and not the set of agents $\{a, b, c\}$. c had nothing to do with it.

We now consider how the treatment of individual 'brings it about' agency can be generalised to collective action by sets of agents. We will build up the account in stages. Section 6.3 will identify an analogue of E_x for a set of agents, and Section 6.4 an analogue of E^+_x.

Henceforth, when we say 'we have φ' or sometimes just 'φ' we mean that φ is a theorem of $\mathrm{S5}^{Ag}$, or equivalently, valid in all LTS frames.

6.1 The modalities E^+_G

We have discussed the logic of

$$\mathrm{E}_G\varphi \ =_{\text{def}}\ [G]\varphi \wedge \neg[\text{alt}]\varphi$$

By analogy with the definition of E^+_x for an individual agent x, let us consider first, for any $G \subseteq Ag$:

$$\mathrm{E}^+_G\varphi \ =_{\text{def}}\ [G]\varphi \wedge \neg[Ag\backslash G]\varphi$$

For the degenerate case $G = \emptyset$, $[G]\varphi$ is $[\text{alt}]\varphi$, and since $[\text{alt}]\varphi \to [Ag\backslash G]\varphi$, we have $\mathrm{E}^+_\emptyset\varphi \leftrightarrow \bot$.

It is straightforward to confirm that we also have

$$\mathrm{E}^+_G\varphi \ \leftrightarrow\ \mathrm{E}_G\varphi \wedge \neg\mathrm{E}_{Ag\backslash G}\varphi \tag{27}$$

Since $\mathrm{E}_\emptyset\varphi \leftrightarrow \bot$, from (27) we have, e.g.

$$\mathrm{E}^+_{Ag}\varphi \leftrightarrow \mathrm{E}_{Ag}\varphi$$

Does $\mathrm{E}^+_G\varphi$ provide a plausible representation of collective agency? No. If $G \subseteq H$ then $\mathrm{E}_G\varphi \to \mathrm{E}_H\varphi$. And if $G \subseteq H$ then $Ag \setminus H \subseteq Ag \setminus G$, and so $\mathrm{E}_{Ag\backslash H}\varphi \to \mathrm{E}_{Ag\backslash G}\varphi$. So from (27) follows also

$$\mathrm{E}^+_G\varphi \to \mathrm{E}^+_H\varphi \qquad \text{if } G \subseteq H$$

and therefore also

$$\mathrm{E}^+_G\varphi \to \mathrm{E}^+_{Ag}\varphi$$

Clearly $\mathrm{E}^+_G\varphi$ does not express that it is the set G of agents that is solely responsible for bringing it about that φ. At best E_G and E^+_G express a very weak sense of collective agency indeed. But there is something we can say. Consider $\mathrm{E}^+_G\varphi \wedge \mathrm{E}_H\varphi$ for any two subsets G and H of Ag. From (27) we get $\mathrm{E}^+_G\varphi \to \neg\mathrm{E}_{Ag\backslash G}\varphi$. If $G \cap H = \emptyset$, then $H \subseteq Ag \setminus G$, and $\mathrm{E}_H\varphi \to \mathrm{E}_{Ag\backslash G}\varphi$.

So then we have both $\mathrm{E}^+_G\varphi \rightarrow \neg\mathrm{E}_{Ag\setminus G}\varphi$ and $\mathrm{E}_H\varphi \rightarrow \mathrm{E}_{Ag\setminus G}\varphi$, and hence $\mathrm{E}^+_G\varphi \wedge \mathrm{E}_H\varphi \rightarrow \bot$ if $G \cap H = \emptyset$. *A fortiori*:

$$\mathrm{E}^+_G\varphi \wedge \mathrm{E}^+_H\varphi \rightarrow \bot \qquad \text{if } G \cap H = \emptyset \tag{28}$$

So although G and H need not be *unique* they must have some members in common. Notice the special case: $\mathrm{E}^+_x\varphi \wedge \mathrm{E}^+_y\varphi \rightarrow \bot$ if $\{x\} \cap \{y\} = \emptyset$, i.e., if $x \neq y$:

$$\mathrm{E}^+_x\varphi \wedge \mathrm{E}^+_y\varphi \rightarrow \bot \qquad \text{if } x \neq y$$

This is something. But still we have $\mathrm{E}^+_G\varphi \rightarrow \mathrm{E}^+_H\varphi$ for any $G \subseteq H$, which is hopeless.

6.2 Counteraction modalities

The characterisation of collective agency depends critically on the formulation of appropriate counteraction conditions, and in particular on what it means to say that a set G of agents, collectively, acts differently in two alternative transitions. The differences are quite subtle, and not easy to express unambiguously in natural language.

It is convenient to switch to a functional notation. Let:

$$\begin{aligned} alt(\tau) &=_{\text{def}} \{\tau' \mid \tau \sim \tau'\} \\ alt_x(\tau) &=_{\text{def}} \{\tau' \mid \tau \sim_x \tau'\} \end{aligned}$$

Then

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

where

$$alt_G(\tau) \;=_{\text{def}}\; alt(\tau) \cap \textstyle\bigcap_{x \in G} alt_x(\tau)$$

Let us now abandon (temporarily) the definitions of E_G and E^+_G used so far, and re-examine their treatment with more careful attention to the counteraction conditions and their simplification.

For E_x, we argued as follows. $\mathrm{E}_x\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) $\mathcal{M}, \tau \models [x]\varphi$, that is, $alt_x(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$: all alternative transitions 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 $\mathcal{M}, \tau' \models \neg\varphi$, that is, $(alt(\tau) \setminus alt_x(\tau)) \cap \|\neg\varphi\|^{\mathcal{M}} \neq \emptyset$.

To express the second condition, the 'counteraction' condition, we introduced another set of operators for talking about transitions in which an agent x acts differently.

$$\begin{array}{lll}\mathcal{M}, \tau \models [x^*]\varphi & \text{iff} & \mathcal{M}, \tau' \models \varphi \text{ for all } \tau' \in \mathcal{M} \text{ such that } \tau \sim \tau' \text{ and} \\ & & \tau \not\sim_x \tau', \text{ i.e., iff } (alt(\tau) \setminus alt_x(\tau)) \subseteq \|\varphi\|^{\mathcal{M}}.\end{array}$$

$\langle x^*\rangle$ are the respective duals.

$\mathrm{E}_x\varphi$ is then defined as follows:

$$\mathrm{E}_x\varphi \;=_{\text{def}}\; [x]\varphi \wedge \neg[x^*]\varphi$$

We were then able to simplify the counteraction condition, replacing $\neg[x^*]\varphi$ in the definition of E_x by $\neg[_{\text{alt}}]\varphi$.

So the natural generalisation is this. Instead of the definition used so far, we will say that $\mathrm{E}_G\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) $\mathcal{M}, \tau \models [G]\varphi$, that is, $alt_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$;

(2) (counteraction) had G, collectively, 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 G, collectively, acts differently in τ' than in τ and $\mathcal{M}, \tau' \models \neg\varphi$.

There remains the question of how to express that a set G of agents, collectively, acts differently in alternative transitions $\tau \sim \tau'$. The simplest is to say it is when $\tau \sim \tau'$ and $\tau \not\sim_G \tau'$, i.e., when $\tau' \in (alt(\tau) \setminus alt_G(\tau))$. We will consider some other possibilities presently.

Definition. For any $G \subseteq Ag$ let

$$\mathrm{E}_G\varphi \;=_{\text{def}}\; [G]\varphi \wedge \neg[G^*]\varphi$$

where

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

For the degenerate case $G = \emptyset$ we have $\models [\emptyset^*]\bot$, i.e., $\models [\emptyset^*]\varphi \leftrightarrow \top$.

Since $(alt(\tau) \setminus alt_G(\tau)) \subseteq alt(\tau)$ (even when $G = \emptyset$) we have validity of

$$[_{\text{alt}}]\varphi \rightarrow [G^*]\varphi$$

More generally, if $G \neq \emptyset$ then $G \subseteq H$ implies $(alt(\tau) \setminus alt_H(\tau)) \subseteq (alt(\tau) \setminus alt_G(\tau))$, and we have validity of

$$[G^*]\varphi \rightarrow [H^*]\varphi \qquad \text{if } \emptyset \subset G \subseteq H$$

(When $G = \emptyset$, this is $\top \rightarrow [H^*]\varphi$, which is not valid, unless $H = \emptyset$.)

For $G \neq \emptyset$, for all τ:

$$alt(\tau) \setminus alt_G(\tau) \;=\; alt(\tau) \setminus \textstyle\bigcap_{x\in G} alt_x(\tau) \;=\; \textstyle\bigcup_{x\in G}(alt(\tau) \setminus alt_x(\tau))$$

So, as defined above, an alternative transition in which G, collectively, acts differently than it does in τ is one where any agent $x \in G$ acts differently than it does in τ. The following is valid for any $G \subseteq Ag$:

$$[G^*]\varphi \;\leftrightarrow\; \textstyle\bigwedge_{x\in G}[x^*]\varphi \tag{29}$$

This is because, if $G \neq \emptyset$:

$$\begin{aligned}(\mathit{alt}(\tau) \setminus \mathit{alt}_G(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \quad &\text{iff} \quad \textstyle\bigcup_{x\in G}(\mathit{alt}(\tau) \setminus \mathit{alt}_x(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \\ &\text{iff} \quad (\mathit{alt}(\tau) \setminus \mathit{alt}_x(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \quad \text{for all } x \in G\end{aligned}$$

If $G = \emptyset$ then (29) is $[\emptyset^*]\varphi \leftrightarrow \top$, whose validity was noted above.

This seems quite natural. We will consider a slightly different, but as it turns out a much stronger, statement of what it means for G, collectively, to act differently in alternative transitions, later in Section 6.4.

From (29) we have the validity of:

$$\mathrm{E}_G\varphi \;\leftrightarrow\; [G]\varphi \wedge \neg \textstyle\bigwedge_{x\in G}[x^*]\varphi \tag{30}$$

(If $G = \emptyset$ this is $\mathrm{E}_\emptyset\varphi \leftrightarrow [\mathrm{alt}]\varphi \wedge \neg\top$.)

How does this new definition of E_G relate to the simpler one examined in previous sections? They are equivalent.

Proposition 6.1. The following is valid for any $G \subseteq Ag$:

$$[G]\varphi \rightarrow ([G^*]\varphi \leftrightarrow [\mathrm{alt}]\varphi)$$

This is easily confirmed. It is a generalisation of the observation (5) in Section 4, and a special case of Proposition 6.8 below. (For $G = \emptyset$ it is $[\mathrm{alt}]\varphi \rightarrow ([\emptyset^*]\varphi \leftrightarrow [\mathrm{alt}]\varphi)$, which is a tautology since $\models [\emptyset^*]\varphi \leftrightarrow \top$.)

From this follows immediately the simplification of E_G.

Proposition 6.2. The following is valid for any $G \subseteq Ag$:

$$\mathrm{E}_G\varphi \;\leftrightarrow\; [G]\varphi \wedge \neg[\mathrm{alt}]\varphi$$

$\mathrm{E}^+_G\varphi$ can be treated in exactly the same way, though we need to generalise the $[G^*]$ operators, as follows.

$\mathrm{E}^+_G\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) $\mathcal{M}, \tau \models [G]\varphi$, that is, $\mathit{alt}_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$;

(2) (counteraction) had G, collectively, acted differently than it did in τ then the transition might have been different, *even if* all other agents $Ag \setminus G$ acted the same way as they did in τ: there exists a transition τ' in $\mathcal{M}$ such that $\tau \sim \tau'$ and G acts differently in τ' than in τ and $\mathcal{M}, \tau' \models \neg\varphi$, and $\tau \sim_y \tau'$ for all $y \in Ag \setminus G$, i.e., $(\mathit{alt}_{Ag\setminus G}(\tau) \setminus \mathit{alt}_G(\tau)) \cap \|\neg\varphi\|^{\mathcal{M}} \neq \emptyset$.

Definition. For any $G \subseteq Ag$ let

$$\mathrm{E}^{+}_{G}\varphi \;=_{\mathrm{def}}\; [G]\varphi \wedge \neg[\,Ag \backslash G \,|\, G^{*}\,]\varphi$$

where, for any $H \subseteq Ag$:

$$\mathcal{M}, \tau \models [\,H \,|\, G^{*}\,]\varphi \quad \text{iff} \quad (alt_{H}(\tau) \setminus alt_{G}(\tau)) \subseteq \|\varphi\|^{\mathcal{M}}$$

For the case $H = \emptyset$, $alt_{H}(\tau) = alt(\tau)$, and so $[G^{*}]$ can be defined as a special case of the general form $[\,H \,|\, G^{*}\,]$: the following is valid

$$[G^{*}]\varphi \;\leftrightarrow\; [\,\emptyset \,|\, G^{*}\,]\varphi$$

For $G = \emptyset$ and any $H \subseteq Ag$, we have $\models [\,H \,|\, \emptyset^{*}\,]\varphi \leftrightarrow \top$, and hence $\mathrm{E}^{+}_{\emptyset}\varphi = [\mathrm{alt}]\varphi \wedge \neg[\,Ag \,|\, \emptyset^{*}\,]\varphi$, which is equivalent to $\bot$.

There is a very close connection here to Boolean Modal Logic (Gargov and Passy, 1990). (See also (Blackburn et al., 2001, pp 424–425).) We will not develop that connection. We only need to consider some special cases, and for that purpose it is not necessary to introduce a new set of definitions and notations. Moreover, we are considering a special case in which $G \subseteq H$ implies $\sim_{H} \subseteq \sim_{G}$, and it is easier to employ the special-purpose notation $[\,H \,|\, G^{*}\,]\varphi$ for convenience.[4]

The following generalises $\models [\mathrm{alt}]\varphi \to [G^{*}]\varphi$.

Proposition 6.3. For any subsets G and H of Ag

$$\models\; [H]\varphi \to [\,H \,|\, G^{*}\,]\varphi$$

Proof. For all τ: $alt_{H}(\tau) \setminus alt_{G}(\tau) \subseteq alt_{H}(\tau)$ for any G (including $G = \emptyset$). □

We have already noted $\models [\,G \,|\, \emptyset^{*}\,]\varphi \leftrightarrow \top$, but more generally $[\,G \,|\, G^{*}\,]\varphi \leftrightarrow \top$ is valid (because $alt_{G}(\tau) \backslash alt_{G}(\tau) = \emptyset$). A more general version still will be useful later.

Proposition 6.4 ('triviality'). For any subsets G and H of Ag:

$$\models\; [\,H \,|\, G^{*}\,]\varphi \leftrightarrow \top \qquad \text{if } G \subseteq H$$

Proof. For any τ, $alt_{H}(\tau) \setminus alt_{G}(\tau) = \emptyset$ if $G \subseteq H$ (including when $G = \emptyset$). □

Proposition 6.5 ('monotony').

$$\begin{array}{lll} \models & [\,H' \,|\, G^{*}\,]\varphi \to [\,H \,|\, G^{*}\,]\varphi & \text{if } H \subseteq H' \\ \models & [\,H \,|\, {G_1}^{*}\,]\varphi \to [\,H \,|\, {G_2}^{*}\,]\varphi & \text{if } \emptyset \subset G_1 \subseteq G_2 \end{array}$$

Proof. If $H \subseteq H'$ then $(alt_{H}(\tau) \setminus alt_{G}(\tau)) \subseteq (alt_{H'}(\tau) \setminus alt_{G}(\tau))$.
If $G_1 \neq \emptyset$ and $G_1 \subseteq G_2$ then $(alt_{H}(\tau) \setminus alt_{G_2}(\tau)) \subseteq (alt_{H}(\tau) \setminus alt_{G_1}(\tau))$. □

Now the very useful 'distribution theorem'.

[4] For readers familiar with Boolean Modal Logic: $[\mathrm{alt}]\varphi$ would be written $[\sim]\varphi$, $[G]\varphi$ would be $[\sim \cap \bigcap_{x \in G} \sim_{x}]\varphi$, $[G^{*}]\varphi$ would be $[\sim \setminus \bigcap_{x \in G} \sim_{x}]\varphi$, and $[\,H \,|\, G^{*}\,]\varphi$ would be $[\bigcap_{x \in H} \sim_{x} \setminus \bigcap_{x \in G} \sim_{x}]\varphi$.

Proposition 6.6 ('distribution')**.** For any subsets G_1, G_2 and H of Ag:

$$[\,H \,|\, (G_1 \cup G_2)^*\,]\varphi \;\leftrightarrow\; [\,H \,|\, {G_1}^*\,]\varphi \wedge [\,H \,|\, {G_2}^*\,]\varphi$$

Proof. For all τ:

$$\begin{array}{rcl} alt_H(\tau) \setminus alt_{G_1 \cup G_2}(\tau) & = & alt_H(\tau) \setminus (alt_{G_1} \cap alt_{G_2}) \\ & = & (alt_H(\tau) \setminus alt_{G_1}(\tau)) \cup (alt_H(\tau) \setminus alt_{G_1}(\tau)) \end{array}$$

□

The following is an immediate consequence, and is generally useful in providing alternative characterisations of the collective agency operators.

Proposition 6.7. For any subsets H and G of Ag:

$$\models\; [\,H \,|\, G^*\,]\varphi \;\leftrightarrow\; \textstyle\bigwedge_{x \in G}[\,H \,|\, x^*\,]\varphi$$

$[\,H \,|\, x^*\,]$ is the obvious abbreviation for $[\,H \,|\, \{x\}^*\,]$.

Proof. From Proposition 6.6 ('distribution'). For the degenerate case $G = \emptyset$, we need the validity of $[\,H \,|\, \emptyset^*\,]\varphi \leftrightarrow \top$, and that was noted earlier. □

From $\mathrm{E}_G\varphi =_{\mathrm{def}} [G]\varphi \wedge \neg[G^*]\varphi$ and $\mathrm{E}^+_G\varphi =_{\mathrm{def}} [G]\varphi \wedge \neg[\,Ag\backslash G \,|\, G^*\,]\varphi$ we thus obtain

$$\models\; \mathrm{E}_G\varphi \;\leftrightarrow\; [G]\varphi \wedge \neg \textstyle\bigwedge_{x \in G}[x^*]\varphi$$

which was noted earlier, at (30), and also

$$\models\; \mathrm{E}^+_G\varphi \;\leftrightarrow\; [G]\varphi \wedge \neg \textstyle\bigwedge_{x \in G}[\,Ag\backslash G \,|\, \{x\}^*\,] \tag{31}$$

(For $G = \emptyset$, the right hand side is just $[G]\varphi \wedge \neg\top$.)

Theorem 6.8 ('Simplification theorem')**.** The following is valid for any subsets G and H of Ag:

$$[G]\varphi \to ([\,H \,|\, G^*\,]\varphi \leftrightarrow [H]\varphi)$$

Proof. Validity of $[H]\varphi \to [\,H \,|\, G^*\,]\varphi$ follows from

$$(alt_H(\tau) \setminus alt_G(\tau)) \;\subseteq\; alt_H(\tau)$$

which is obviously true for all τ (even if $G = \emptyset$). Validity of

$$[G]\varphi \wedge [\,H \,|\, G^*\,]\varphi \to [H]\varphi$$

follows from

$$alt_H(\tau) \;\subseteq\; alt_G(\tau) \cup (alt_H(\tau) \setminus alt_G(\tau)) \tag{32}$$

which is also true for all τ (even if $G = \emptyset$, because then $alt_H(\tau) \setminus alt(\tau) = \emptyset$, but $alt_H(\tau) \;\subseteq\; alt(\tau)$ for all $H \subseteq G$). Then:

$$\begin{array}{ll} \mathcal{M}, \tau \models [G]\varphi \wedge [\,H \,|\, G^*\,]\varphi & \\ \quad \text{implies} & alt_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}} \text{ and } (alt_H(\tau) \setminus alt_G(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \\ \quad \text{implies} & alt_G(\tau) \cup (alt_H(\tau) \setminus alt_G(\tau)) \subseteq \|\varphi\|^{\mathcal{M}} \\ \quad \text{implies} & alt_H(\tau) \subseteq \|\varphi\|^{\mathcal{M}} \quad \text{by (32)} \end{array}$$

□

The above can be generalised: if $G' \subseteq G$ then $[G']\varphi \to [G]\varphi$ is valid, and so $[G']\varphi \to ([\,H \,|\, G^*\,]\varphi \leftrightarrow [H]\varphi)$ is valid.

This simple observation is so useful that we will refer to it as the 'simplification theorem' in what follows. For example, if $G \neq \emptyset$ then

$$[G]\varphi \to ([\,Ag\backslash G \,|\, G^*\,]\varphi \leftrightarrow [Ag\backslash G]\varphi)$$

which confirms that the definition of E^+_G is equivalent to the simpler form discussed in previous sections.

Proposition 6.9. For any $G \subseteq Ag$, the following is valid

$$\mathrm{E}^+_G\varphi \;\leftrightarrow\; [G]\varphi \land \neg[Ag\backslash G]\varphi$$

Finally, we record two further simple properties which we will want to refer to later.

Proposition 6.10 ('reduction')**.** For any subsets G and H of Ag:

$$[\,H \,|\, G^*\,]\varphi \;\leftrightarrow\; [\,H \,|\, (G\backslash H)^*\,]\varphi$$

Proof. If $G = \emptyset$ the above is equivalent to $\top \leftrightarrow \top$. In general $G = (G \setminus H) \cup (G \cap H)$. So by Proposition 6.6 ('distribution') we have

$$\models\; [\,H \,|\, G^*\,]\varphi \;\leftrightarrow\; [\,H \,|\, (G\backslash H)^*\,]\varphi \land [\,H \,|\, (G \cap H)^*\,]\varphi$$

But $G\cap H \subseteq H$, so by Proposition 6.4 ('triviality') $\models [\,H \,|\, (G\cap H)^*\,]\varphi \leftrightarrow \top$. □

Proposition 6.11. For any subsets G, H and X of Ag:

$$\models\; [\,H\backslash X \,|\, G^*\,]\varphi \;\leftrightarrow\; [\,H\backslash X \,|\, (G \cap X)^*\,]\varphi \qquad \text{if } G \subseteq H$$

Proof. $G \setminus (H \setminus X) = (G \setminus H) \cup (G \cap X)$. So if $G \subseteq H$, $G \setminus (H \setminus X) = G \cap X$. Now apply Proposition 6.10. □

6.3 Minimal sets of agents

E_G and E^+_G express a very weak kind of collective agency. If the set G of agents collectively brings it about that φ then so, in a very weak sense, does every superset of G; indeed the set Ag also collectively brings it about that φ. But this is not what we are aiming at. In the table-vase example when a and b collectively tilt the table and break the vase, c has nothing to do with it.

One way of looking at it is that the necessary condition $[G]\varphi$ is too weak: G can be 'too big'—it might contain x who contributes nothing to the bringing about of φ: after all, for $x \in G$, $[G\backslash\{x\}]\varphi \land [G]\varphi$ is satisfiable, and $[G\backslash\{x\}]\varphi \to [G]\varphi$. It seems inescapable to look at the subsets of G in an expression $[G]\varphi$, and insist that for the necessity condition, G should be *minimal* in some sense. There are several possible ways of expressing this requirement. Let us consider the obvious one first.

Let

$$[G]_{\min}\varphi \;=_{\text{def}}\; [G]\varphi \wedge \neg \bigvee\nolimits_{H\subset G}[H]\varphi$$

and

$$\Delta_G\varphi \;=_{\text{def}}\; [G]_{\min}\varphi \wedge \neg[G^*]\varphi$$

that is, in full:

$$\Delta_G\varphi \;=_{\text{def}}\; [G]\varphi \wedge \neg \bigvee\nolimits_{H\subset G}[H]\varphi \wedge \neg[G^*]\varphi$$

In words: $\mathcal{M},\tau \models \Delta_G\varphi$ iff (1) $\mathcal{M},\tau \models [G]\varphi$, (1′) there is no proper subset $H \subset G$ such that $\mathcal{M},\tau \models [H]\varphi$, and (2) had G, collectively, acted differently than it did in τ the transition might have been different.

The minimality condition (1′) could be regarded as part of the 'necessity' condition or as part of the counteraction condition. It makes no difference.

When $G = \emptyset$, $\models [\emptyset]_{\min}\varphi \leftrightarrow [\emptyset]\varphi$ (the empty disjunction is false), i.e., $\models [\emptyset]_{\min}\varphi \leftrightarrow [\text{alt}]\varphi$. And $\Delta_\emptyset\varphi \leftrightarrow [\text{alt}]\varphi \wedge \neg[\emptyset^*]\varphi$, which is equivalent to $\bot$ since $\models [\emptyset^*]\varphi \leftrightarrow \top$.

When $G = \{x\}$, $[\{x\}]_{\min}\varphi = [x]\varphi \wedge \neg \bigvee_{H\subset\{x\}}[H]\varphi$, and so $\models [\{x\}]_{\min}\varphi \leftrightarrow [x]\varphi \wedge \neg[\emptyset]\varphi$, that is, $\models [\{x\}]_{\min}\varphi \leftrightarrow [x]\varphi \wedge \neg[\text{alt}]\varphi$, i.e.:

$$\models\; [\{x\}]_{\min}\varphi \;\leftrightarrow\; \mathrm{E}_x\varphi$$

The simplification of the counteraction condition $\neg[G^*]\varphi$ is still available: it requires only $[G]\varphi$. So by the 'simplification theorem' we have:

$$\Delta_G\varphi \;\leftrightarrow\; [G]_{\min}\varphi \wedge \neg[\text{alt}]\varphi \tag{33}$$

For the degenerate case $G = \emptyset$, $\Delta_\emptyset\varphi \leftrightarrow [\emptyset]_{\min}\varphi \wedge \neg[\text{alt}]\varphi$, which is equivalent to $\bot$. For the case $G = \{x\}$, $\Delta_{\{x\}}\varphi \leftrightarrow \mathrm{E}_x\varphi \wedge \neg[\text{alt}]\varphi$, and so:

$$\Delta_{\{x\}}\varphi \leftrightarrow \mathrm{E}_x\varphi$$

Notice also that we have $\neg[H]\varphi \rightarrow \neg[\text{alt}]\varphi$ for every $H \subseteq Ag$ and so for the case $G \neq \emptyset$, (33) can be simplified:

$$\models\; \Delta_G\varphi \leftrightarrow [G]_{\min}\varphi \qquad (G \neq \emptyset)$$

For $G = \emptyset$, $\models \Delta_G\varphi \leftrightarrow \bot$.

It is very useful to have an alternative characterisation of $[G]_{\min}\varphi$ and hence of $\Delta_G\varphi$.

Proposition 6.12. For every $G \subseteq Ag$:

$$\bigvee\nolimits_{H\subset G}[H]\varphi \;\leftrightarrow\; \bigvee\nolimits_{x\in G}[G\backslash\{x\}]\varphi$$

Proof. The degenerate case $G = \emptyset$ is trivial: it is $\bot \leftrightarrow \bot$.

For left-to-right. If $H \subset G$ then $H \subseteq G \setminus \{x\}$ for some $x \in G$, and then $[H]\varphi \rightarrow [G\backslash\{x\}]\varphi$. So we have:

$$\bigwedge\nolimits_{H\subset G}\big([H]\varphi \;\rightarrow\; \bigvee\nolimits_{x\in G}[G\backslash\{x\}]\varphi\big)$$

which is equivalent to

$$\bigvee\nolimits_{H\subset G}[H]\varphi \;\rightarrow\; \bigvee\nolimits_{x\in G}[G\backslash\{x\}]\varphi$$

The other direction is immediate. $G \setminus \{x\} \subset G$ for every $x \in G$. □

From Proposition 6.12 we have the following.

Proposition 6.13.

$$\models\ [G]_{\min}\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi$$

and so

$$\models\ \Delta_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi \wedge \neg[{\scriptstyle\mathrm{alt}}]\varphi$$

The conjunct $\neg[{\scriptstyle\mathrm{alt}}]\varphi$ is to deal with the special case when $G=\emptyset$. When $G=\emptyset$, $\neg[G\backslash\{x\}]\varphi \to \neg[{\scriptstyle\mathrm{alt}}]\varphi$ and it is redundant.

Notice the relative strength of E_G and Δ_G. From Proposition 6.13 we can see immediately:

$$\models\ \Delta_G\varphi \ \leftrightarrow\ \mathrm{E}_G\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi$$

Let us now generalise. Let

$$\Delta^+_G\varphi \ =_{\mathrm{def}}\ [G]_{\min}\varphi \wedge \neg[\,Ag\backslash G \mid G^*\,]\varphi$$

or, in full:

$$\Delta^+_G\varphi \ =_{\mathrm{def}}\ [G]\varphi \wedge \neg \textstyle\bigvee_{H\subset G}[H]\varphi \wedge \neg[\,Ag\backslash G \mid G^*\,]\varphi$$

or equivalently

$$\Delta^+_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi \wedge \neg[\,Ag\backslash G \mid G^*\,]\varphi$$

For the degenerate case $G=\emptyset$, we have $\Delta^+_\emptyset\varphi = [\emptyset]_{\min}\varphi \wedge \neg[\,Ag \mid \emptyset^*\,]\varphi$, which is equivalent to $\bot$ since $\models [\,Ag \mid \emptyset^*\,]\varphi \leftrightarrow \top$.

Again, the simplification of the counteraction condition is available, since it depends only on $[G]\varphi$. So we have the validity of:

$$\Delta^+_G\varphi \ \leftrightarrow\ [G]_{\min}\varphi \wedge \neg[Ag\backslash G]\varphi$$

and

$$\Delta^+_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi \wedge \neg[Ag\backslash G]\varphi$$

which is also

$$\Delta^+_G\varphi \ \leftrightarrow\ \mathrm{E}^+_G\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi$$

and also

$$\Delta^+_G\varphi \ \leftrightarrow\ \Delta_G\varphi \wedge \neg[Ag\backslash G]\varphi$$

What about some of the properties of Δ_G and Δ^+_G?

- Obviously we have both of the following:

  $$\Delta_G\varphi \to \mathrm{E}_G\varphi \quad \text{and} \quad \Delta^+_G\varphi \to \mathrm{E}^+_G\varphi$$

  They require only $[G]_{\min}\varphi \to [G]\varphi$. In fact we have already observed that both of the following are valid:

  $$\begin{aligned}\Delta_G\varphi \ &\leftrightarrow\ \mathrm{E}_G\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi\\ \Delta^+_G\varphi \ &\leftrightarrow\ \mathrm{E}^+_G\varphi \wedge \textstyle\bigwedge_{x\in G}\neg[G\backslash\{x\}]\varphi\end{aligned}$$

- It is no longer the case that if $G \subseteq H$ then $\Delta_G \varphi \to \Delta_H \varphi$. Indeed, we have for all subsets G and H of Ag:

  $$\Delta_G \varphi \to \neg \Delta_H \varphi \qquad \text{if } G \subset H$$

  If $G \neq \emptyset$ then $[G]_{\min} \varphi \to \neg [H]_{\min} \varphi$ for $G \subset H$ by definition of $[H]_{\min} \varphi$. If $G = \emptyset$, then $\Delta_G \varphi \leftrightarrow \bot$, and so trivially $\Delta_G \varphi \to \neg \Delta_H \varphi$ for any $H \subseteq Ag$.

- $\Delta_G \varphi \wedge \Delta_H \varphi$ is satisfiable for $G \neq H$.

  Suppose there are three agents a, b, and c pushing against a spring-loaded door and keeping it shut. Suppose any two of them collectively are strong enough to keep the door shut. Let k represent that the transition is a 'keeping-the-door-shut' transition. If we do the calculation we find that $[\{a,b\}]_{\min} k$, $[\{b,c\}]_{\min} k$, and $[\{a,c\}]_{\min} k$ are all true. Clearly $[\{a,b,c\}]k$ is true but $[\{a,b,c\}]_{\min} k$ is not. $\Delta_{\{a,b\}} k$ is $[\{a,b\}]_{\min} k \wedge \neg [\mathrm{alt}] k$. Since there is an alternative transition (we are supposing) in which $\neg k$ is true, we have $\Delta_{\{a,b\}} k$. And likewise for $\Delta_{\{b,c\}} k$ and $\Delta_{\{a,c\}} k$.

- The above example demonstrates that:

  $$\not\models \ \Delta_G \varphi \to \mathrm{E}^+_G \varphi$$

- Since $\Delta^+_G \varphi \wedge \Delta^+_H \varphi \to \mathrm{E}^+_G \varphi \wedge \mathrm{E}^+_H \varphi$, we inherit the property (28), that is, we have:

  $$\Delta^+_G \varphi \wedge \Delta^+_H \varphi \to \bot \qquad \text{if } G \cap H = \emptyset \tag{34}$$

  So although again G and H need not be unique, they cannot be disjoint.

In the 'keeping-the-door-shut' example, we have $\Delta_{\{a,b\}} k$ and $\neg [c] k$ (c is not strong enough to keep the door shut by itself). So $\Delta^+_{\{a,b\}} k$ is true in the transition. And likewise $\Delta^+_{\{b,c\}} k$ and $\Delta^+_{\{a,c\}} k$ are also true. The sets $\{a,b\}$, $\{b,c\}$, and $\{a,c\}$ who bring about k are not unique, but also not disjoint, as expected.

But consider a variation of the example. Suppose there are four agents a, b, c, and d pushing against the door, and any two of them are strong enough to keep the door shut. If we do the calculation for this version of the example, we find (assuming the door can spring open) that $\Delta_G k$ is true in this transition for any pair G from $\{a,b,c,d\}$. But consider $\Delta^+_{\{a,b\}} k$. That is false, because it requires $\neg [\{c,d\}] k$, and $[\{c,d\}] k$ is true. $\Delta^+_{\{c,d\}} k$ is false too, and so is $\Delta^+_G k$ for any pair G of agents from $\{a,b,c,d\}$.

This seems very odd, at best. Suppose we have some set Ag of agents, any two of whom are strong enough to keep the spring-loaded door shut if they push together. If three of them are pushing the door shut, then $\Delta^+_G k$ is true for any pair G of them. If four or more are pushing, however, $\Delta^+_G k$ is false for any $G \subseteq Ag$. That cannot be right. Perhaps there is some kind of informal reading of Δ^+_G that makes these properties intuitively plausible, but it is far from clear what it might be.

6.4 Collective agency

In Section 4.5 we noted that E_x and E_x^+ are the endpoints of a range of possible 'brings it about' operators that differ only in the strength of the counteraction conditions. Here, in a slightly compressed form, is the list of the possible counteraction conditions considered before, but now with G in place of the individual agent x:

strongest	$[G]\varphi \wedge \neg[Ag \backslash G]\varphi$	had all the others in $Ag \setminus G$ acted the same way they did, it might have been otherwise
	$\vdots$	
	$[G]\varphi \wedge \neg[H]\varphi$	had all the others in some subset $H \subseteq Ag \setminus G$ acted the same way they did, it might have been otherwise
	$\vdots$	
weakest	$[G]\varphi \wedge \neg[\mathrm{alt}]\varphi$	it might have been otherwise

Or equivalently, because of the 'simplification theorem':

strongest	$[G]\varphi \wedge \neg[\, Ag \backslash G \mid G^* \,]\varphi$	had G acted differently, and all the others in $Ag \setminus G$ the same way they did, it might have been otherwise
	$\vdots$	
	$[G]\varphi \wedge \neg[\, H \mid G^* \,]\varphi$	had G acted differently, and all the others in some subset $H \subseteq Ag \setminus G$ the same way they did, it might have been otherwise
	$\vdots$	
weakest	$[G]\varphi \wedge \neg[G^*]\varphi$	had G acted differently, it might have been otherwise

The strongest is $\mathrm{E}_G^+\varphi$ and the weakest is $\mathrm{E}_G\varphi$.

But here is a question: should we not consider also sets H other than subsets of $Ag \setminus G$? It made no sense to do this earlier where we were considering only individual agents and where G was always a singleton, but in this more general setting there are more possibilities.

Here is one way to look at it. Consider constructions of this general form:

$$[G]\varphi \wedge \neg[\, H \mid G^* \,]\varphi \tag{35}$$

The bigger we make H, the stronger is the counteraction condition $\neg[\, H \mid G^* \,]\varphi$: for any $H' \subseteq H$ we have $\neg[\, H \mid G^* \,]\varphi \rightarrow \neg[\, H' \mid G^* \,]\varphi$ by Proposition 6.5 ('monotony'). We can ignore all supersets of G because if $G \subseteq H$ then by Proposition 6.4 ('triviality') we have $\neg[\, H \mid G^* \,]\varphi \leftrightarrow \bot$. So what are the maximal (set inclusion) H such that $G \not\subseteq H$? They are any $H = Ag \setminus \{x\}$ where $x \in G$.

So the *strongest* counteraction condition we can construct this way suggests the following as a plausible candidate for the expression of (unwitting) collective agency of the set $G \subseteq Ag$ of agents:

$$[G]\varphi \wedge \textstyle\bigwedge_{x \in G} \neg[\, Ag \backslash \{x\} \mid G^* \,]\varphi$$

or equivalently by the 'simplification theorem'

$$[G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[Ag\backslash\{x\}]\varphi$$

which is

$$[G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[\backslash x]\varphi$$

in the shorthand notation used earlier.

This construction does indeed give us a form of collective agency with the desired properties, as will be demonstrated below. But how do we read these constructions? Their intuitive reading is far from clear.

We have by Proposition 6.11, that for any subsets G, H and X of Ag:

$$\models\ [\,H\backslash X \,|\, G^*\,]\varphi \ \leftrightarrow\ [\,H\backslash X \,|\, (G\cap X)^*\,]\varphi \qquad \text{if } G\subseteq H$$

And so, as a special case, for any subsets G and X of Ag:

$$\models\ [\,Ag\backslash X \,|\, G^*\,]\varphi \ \leftrightarrow\ [\,Ag\backslash X \,|\, (G\cap X)^*\,]\varphi$$

From this follows immediately:

Proposition 6.14. For any subset G of Ag:

$$[\,Ag\backslash\{x\} \,|\, G^*\,]\varphi \ \leftrightarrow\ [\,Ag\backslash\{x\} \,|\, x^*\,]\varphi \qquad \text{for any } x\in G$$

Proof. If $x \in G$, then $G \cap \{x\} = \{x\}$. □

So we can re-express the counteraction conditions above in the following equivalent form.

Definition. For any $G \subseteq Ag$, let

$$\Gamma_G\varphi \ =_{\text{def}}\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[\,Ag\backslash\{x\} \,|\, \{x\}^*\,]\varphi \wedge \neg[\text{alt}]\varphi$$

In the shorthand notation of previous sections this is:

$$\Gamma_G\varphi \ =_{\text{def}}\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[\,\backslash x \,|\, x^*\,]\varphi \wedge \neg[\text{alt}]\varphi$$

As usual, the conjunct $\neg[\text{alt}]\varphi$ is included to deal with the degenerate case $G = \emptyset$. It is redundant (implied) if $G \neq \emptyset$.

In words: $\Gamma_G\varphi$ is satisfied by a transition τ in a model $\mathcal{M}$ when:

(1) (necessity) $\mathcal{M}, \tau \models [G]\varphi$, that is, $alt_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$;

(2) (counteraction) for every $x \in G$, had x acted differently than it did in τ then the transition might have been different, even if all other agents $Ag \setminus \{x\}$ besides x acted in the same way as they did in τ.

One can see that the counteraction condition above is similar, though slightly different, from those we have considered earlier. In particular it expresses a slightly different sense of what it means to say that a set G of agents, collectively, acts differently in a pair of transitions $\tau \sim \tau'$, and a slightly different sense in which every $x \in G$ must contribute to G's collectively bringing it about that φ.

It is worth observing that the definition of Γ_G incorporates the strongest counteraction conditions of the kind we have been looking at. In considering the general form (35), we chose maximal sets H to make the counteraction condition $\neg[\,H \mid G^*\,]\varphi$ as strong as possible. By Proposition 6.5 ('monotony'), however, we get stronger counteraction conditions $\neg[\,H \mid G'^*\,]\varphi$ by choosing smaller sets G'. The smallest such sets G' we can choose are any sets $\{x\}$ where $x \in G$. $G' = \emptyset$ is no good, because then $\neg[\,H \mid G'^*\,]\varphi$ is equivalent to $\bot$.

Now let us record the simplification of the definition.

Proposition 6.15.

$$\models\ \Gamma_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[Ag\backslash\{x\}]\varphi \wedge \neg[\text{alt}]\varphi$$

which in the shorthand notation of previous sections is

$$\models\ \Gamma_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[\backslash x]\varphi \wedge \neg[\text{alt}]\varphi$$

Proof. The degenerate case $G = \emptyset$ is trivial. For the general case the argument is given above. In summary: by the 'simplification theorem', $[G]\varphi \wedge \bigwedge_{x\in G} \neg[Ag\backslash\{x\}]\varphi$ is equivalent to $[G]\varphi \wedge \bigwedge_{x\in G} \neg[\,Ag\backslash\{x\} \mid G^*\,]\varphi$; and by Proposition 6.10 ('reduction') and the corollary Proposition 6.14, this is equivalent to $[G]\varphi \wedge \bigwedge_{x\in G} \neg[\,Ag\backslash\{x\} \mid \{x\}^*\,]\varphi$, which is $\Gamma_G\varphi$ by definition. □

Now two properties of Γ_G. First, compare (for the case $G \neq \emptyset$):

$$\Gamma_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[Ag\backslash\{x\}]\varphi$$

and

$$\Delta^+_G\varphi \ \leftrightarrow\ [G]\varphi \wedge \textstyle\bigwedge_{x\in G} \neg[G\backslash\{x\}]\varphi \wedge \neg[G\backslash\{x\}]\varphi$$

Proposition 6.16. For any $G \subseteq Ag$, the following is a theorem of $\mathrm{S5}^{Ag}$, or equivalently, valid in all LTS frames.

$$\Gamma_G\varphi \rightarrow \Delta^+_G\varphi$$

Proof. For $G = \emptyset$, the above is equivalent to $\bot \leftrightarrow \bot$.

For the general case, by Proposition 6.15 it is enough to show:

$$\textstyle\bigwedge_{x\in G} \neg[Ag\backslash\{x\}]\varphi \ \rightarrow\ (\bigwedge_{x\in G} \neg[G\backslash\{x\}]\varphi \wedge \neg[Ag\backslash G]\varphi)$$

We show that, for every $x \in G$

$$\neg[Ag\backslash\{x\}]\varphi \ \rightarrow\ (\neg[G\backslash\{x\}]\varphi \wedge \neg[Ag\backslash G]\varphi)$$

First: $G \setminus \{x\} \subseteq Ag \setminus \{x\}$ and so $[G\backslash\{x\}]\varphi \rightarrow [Ag\backslash\{x\}]\varphi$. Second: $Ag \setminus G \subseteq Ag \setminus \{x\}$ when $x \in G$, and so $\neg[Ag\backslash\{x\}]\varphi \rightarrow \neg[Ag\backslash G]\varphi$. □

Proposition For any subsets G and H of Ag:

$$\Gamma_G\varphi \wedge \Gamma_H\varphi \rightarrow \bot \qquad \text{if } G \neq H$$

Proof. If $G = \emptyset$ then the above is equivalent to $\bot \wedge \Gamma_H\varphi \rightarrow \bot$, which is a tautology. And likewise for the case $H = \emptyset$.

Suppose that G and H are both non-empty. If $G \neq H$ then either $G \not\subseteq H$ or $H \not\subseteq G$. Suppose that $G \not\subseteq H$. Consider any $x \in G \setminus H$. Then $H \subseteq Ag \setminus \{x\}$. Now: $x \in G$ means that $\Gamma_G\varphi \rightarrow \neg[Ag\backslash\{x\}]\varphi$ from the definition of Γ_G. But $H \subseteq Ag\backslash\{x\}$ implies $[H]\varphi \rightarrow [Ag\backslash\{x\}]\varphi$, and so we also get $\Gamma_H\varphi \rightarrow [Ag\backslash\{x\}]\varphi$. So then $\Gamma_G\varphi \wedge \Gamma_H\varphi \rightarrow \bot$. □

In the table-vase example where there are three agents a, b, and c, and the atom v represents that the vase falls, $\Gamma_{\{a,b\}}v$ is true in the transition where a lifts, b does not, and c has nothing to do with it. In the pushing-the-door-shut example in a transition where any two agents pushing together are required to keep the door shut, we have $\Delta_G k$ true for any pair G from Ag (at least two agents are required to keep the door shut, and if $\Delta_G k$ is true then $\Delta_H k$ is not true for any $H \supset G$). $\Gamma_G k$ on the other hand is not true for any set $G \subseteq Ag$. This seems quite natural and satisfactory.

Γ_G is analogous to Δ_G^+ and E_G^+ in that the counteraction conditions for these modalities refer to all other agents besides G and their actions in alternative transitions. Is there similarly a variant of Γ_G where, as in the case of Δ_G and E_G, no reference is made in the counteraction conditions to the actions of the other agents? It would be:

(1) (necessity) $\mathcal{M}, \tau \models [G]\varphi$, that is, $alt_G(\tau) \subseteq \|\varphi\|^{\mathcal{M}}$;

(2) (counteraction) for every $x \in G$, had x acted differently than it did in τ then the transition might have been different.

Let us call it $\Delta_G^- \varphi$. It is (ignoring the degenerate case $G = \emptyset$):

$$\Delta_G^- \varphi \;=_{\mathrm{def}}\; [G]\varphi \wedge \textstyle\bigwedge_{x \in G} \neg [x^*]\varphi$$

instead of

$$\Gamma_G \varphi \;=_{\mathrm{def}}\; [G]\varphi \wedge \textstyle\bigwedge_{x \in G} \neg [\, Ag \backslash \{x\} \mid x^* \,]\varphi$$

as we have for $\Gamma_G \varphi$. But what is $\Delta_G^- \varphi$? Proposition 6.10 ('reduction') and the corollary Proposition 6.14 do not help here.

We can see that, by Proposition 6.5 ('monotony'), $[\, \emptyset \mid x^* \,]\varphi \to [\, Ag \backslash \{x\} \mid x^* \,]\varphi$ is valid and so

$$\models \; \Gamma_G \varphi \to \Delta_G^- \varphi$$

And $[\text{alt}]\varphi \to [x^*]\varphi$ is valid and so

$$\models \; \Delta_G^- \varphi \to \mathrm{E}_G \varphi$$

On the other hand, Δ_G^- does not seem to express any useful notion of collective agency at all. Consider again the table-vase example with agents a, b, and c, and a transition in which a and b collectively tilt the table and break the vase (v). In this transition $\Delta_{\{a,b\}}^- v$ is true because $[\{a,b\}]v$ and $\langle a^* \rangle \neg v$ and $\langle b^* \rangle \neg v$ are all true. However, $\Delta_{\{a,b,c\}}^- v$ is also true, because $[\{a,b,c\}]v$ is true, and so are $\langle a^* \rangle \neg v$, $\langle b^* \rangle \neg v$, and $\langle c^* \rangle \neg v$ (we are supposing). So $\Delta_G^- \varphi \to [G]_{\min}\varphi$ is not valid, in other words:

$$\not\models \; \Delta_G^- \varphi \to \Delta_G \varphi$$

It is similarly easy to construct examples to show that

$$\not\models \; \mathrm{E}_G^+ \varphi \to \Delta_G^- \varphi$$

Other properties of Δ_G^- can be explored and established quite easily but since Δ_G^- seems not to express any useful notion of collective agency we will not bother with them.

Finally, we noted earlier that

$$\mathrm{E}^+_G\varphi \;\leftrightarrow\; \mathrm{E}_G\varphi \wedge \neg\mathrm{E}_{Ag\backslash G}\varphi$$

Perhaps the construction

$$\Delta_G\varphi \wedge \neg\Delta_{Ag\backslash G}\varphi \tag{36}$$

yields something interesting? It does not. First, observe that:

$$\Gamma_G\varphi \to \Delta_G\varphi \wedge \neg\Delta_{Ag\backslash G}\varphi \tag{37}$$

$\Gamma_G\varphi \to \Delta_G\varphi$, obviously. The other part is also easy to check: $\Gamma_G\varphi \to \mathrm{E}^+_G\varphi$, $\mathrm{E}^+_G\varphi \to \neg[Ag\backslash G]\varphi$, and $\neg[Ag\backslash G] \to \neg\Delta_{Ag\backslash G}\varphi$. But the property (37) is not an equivalence. In the pushing-the-door-shut example we have $\Delta_{\{a,b\}}k$ and $\neg\Delta_{\{c\}}k$ but not $\Gamma_{\{a,b\}}k$. The construction (36) seems to have no particular significance. Similarly, we have

$$\Gamma_G\varphi \to \Delta^+_G\varphi \wedge \neg\Delta^+_{Ag\backslash G}\varphi$$

But this is not an equivalence, and is not interesting either: since G and $Ag \setminus G$ are disjoint, $\Delta^+_G\varphi \to \neg\Delta^+_{Ag\backslash G}\varphi$. The above says only that $\Gamma_G\varphi \to \Delta^+_G\varphi$.

6.5 Summary

We have explored a range of possible forms of collective agency with implications between them as summarised in the following diagram:

$$\begin{tikzcd}
 & \mathrm{E}^+_G \arrow[rr] & & \mathrm{E}_G \\
 & & & \Delta^-_G \arrow[u] \\
\Gamma_G \arrow[uur] \arrow[rr] & & \Delta^+_G \arrow[uul] \arrow[r] & \Delta_G \arrow[u]
\end{tikzcd}$$

Of these, it is Δ_G and Γ_G that are deserving of attention. They are the analogues of E_x and E^+_x, respectively, in as much as $\Delta_G\varphi$ allows for the possibility that several sets G of agents bring it about that φ, while $\Gamma_G\varphi$ expresses that any such set G is unique. Both imply that the set G of agents is minimal: $\Delta_G\varphi \leftrightarrow [G]_{\min}\varphi$ when $G \neq \emptyset$, and $\Gamma_G\varphi \to \Delta_G\varphi$. Of the others, Δ^-_G has a natural definition but is too weak to express any useful notion of collective agency. We called Δ^+_G 'odd'.

These complications arise if we are interested in collective agency of sets $G \subseteq Ag$, or in individual agency of an agent $x \in Ag$ where there are multiple agents in Ag whose actions need to be taken into account. They do not arise if we restrict attention to a single agent acting in an environment, that is, the case where $Ag = \{x\}$ is a singleton. In that case all the above forms collapse to one:

$$\Gamma_{\{x\}}\varphi \leftrightarrow \Delta_{\{x\}}\varphi \leftrightarrow \mathrm{E}^+_x\varphi \leftrightarrow \mathrm{E}_x\varphi \qquad (\text{for } Ag = \{x\})$$

7 Conclusion

The formal framework presented in this paper has been implemented in the form of a model checker that can evaluate formulas expressing properties of interest on (a symbolic representation of) an agent-stranded transition system. That and some illustrations of how the language can be applied to some (very simple) examples can be found in (Sergot, 2008).

Generally, the logic of norms and the logic of action/agency have been studied together. We touched on some examples from norm-governed multi-agent systems but only as a means of motivating parts of the technical development. It remains to explore how the resources of the language can be used to represent norms, and to investigate distinctions and issues that have previously been discussed in the literature.

As regards the logic of (unwitting) agency itself, the main contributions of the paper can be summarised as follows.

(1) Combining transition-based treatments of action with 'brings it about' agency turns out to be quite natural and straightforward, if we switch attention from talking about an agent's bringing it about that a certain state of affairs exists to talking about an agent's bringing it about that a transition has a certain property. This change of perspective also provides a natural way of formalising distinctions between bringing it about, sustaining, and letting it remain the case that a certain state of affairs exists. One could try to do something similar by looking at paths/runs of a transition system rather than just single transitions, perhaps in combination with a temporal logic. We leave that for future work.

(2) Agency of an individual agent, and weak forms of collective agency, are easy to deal with technically. Counterfactual counteraction conditions can be reduced to combinations of simple S5 modalities. The characterisation of their logical properties is then a straightforward exercise; indeed it requires no more than trivial modifications of the S5 logic of 'distributed knowledge'. We looked at two defined 'brings it about' modalities, one which treats the actions of other agents as part of the environment, and a stronger one, expressing sole agency, which takes the actions of other agents into account. We were also able to relate our account to Pörn's (1977) logic of 'brings it about'.

(3) Collective agency, even of the unwitting kind, is more challenging. There are many different ways of formulating the counterfactual counteraction conditions, and in particular, various senses in which a set G of agents, collectively, acts differently in alternative transitions. We picked out two forms for special attention, Δ_G and Γ_G, corresponding roughly to the two forms of 'bringing about' by individual agents, E_x and E^+_x.

The logic of unwitting collective agency is surprisingly strong. Our intention in due course is to add more features, such as communication and joint intention, to look at genuine joint action.

Finally, there is another, weaker kind of brings it about agency that seems to be deserving of attention.

Consider the table-vase example again, but now suppose that if the table tilts, the vase might fall, but might not, and if the vase falls, it might break, or might not. To simplify the example suppose there is just one agent a who can lift and lower its end of the table. We do not need b to make the point. If a lifts its end of the table and the vase falls and breaks, a does not bring it about that the vase breaks—it is not necessary for what a does that the vase falls or that it breaks. But suppose that if a does not lift then the table does not tilt, and if the table does not tilt then the vase does not fall, and if the vase does not fall it does not break. When a lifts the table and the vase falls and breaks, there is a sense in which a brings it about: but for a's actions the vase would not fall and would not break.

What we seem to have here is another, weaker form of 'brings it about' agency. x (weakly) brings it about that φ if φ is true but would not be true had x acted differently. In the language of this paper this would be

$$\varphi \wedge [x^*]\neg\varphi$$

or equivalently

$$\varphi \wedge \neg\langle x^*\rangle\varphi$$

where $[x^*]$ means 'had x acted differently' or 'but for x's actions'. A similar construction is mentioned by Ingmar Pörn (1977, p16) in connection with the ascription of responsibility. The details seem to be deserving of investigation, but are technically more challenging because of the nature of the $[x^*]$ modality and its generalisations.

References

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

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

Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic.* Cambridge University Press, 2001.

B. F. Chellas. *Modal Logic—An Introduction.* Cambridge University Press, 1980.

B. F. Chellas. *The Logical Form of Imperatives.* Dissertation, Stanford University, 1969.

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

Maarten de Rijke. The modal logic of inequality. *Journal of Symbolic Logic*, 57: 566–584, 1992.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge.* MIT Press, Cambridge, 1995.

G. Gargov and S. Passy. A note on Boolean modal logic. In P. P. Petkov, editor, *Mathematical Logic. Proceedings of the 1988 Heyting Summer School*, pages 311–321. Plenum Press, 1990.

Enrico Giunchiglia, Joohyung Lee, Vladimir Lifschitz, Norman McCain, and Hudson Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153(1–2):49–104, 2004.

Gerd Große and Hesham Khalil. State Event Logic. *Journal of the IGPL*, 4(1):47–74, 1996.

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.

Ingmar Pörn. Some basic concepts of action. In S. Stenlund, editor, *Logical Theory and Semantic Analysis*, number 63 in Synthese Library, pages 93–101. D. Reidel, Dordrecht, 1974.

Luigi Sauro, Jelle Gerbrandy, Wiebe van der Hoek, and Michael Wooldridge. Reasoning about action and cooperation. In *Proceedings of the Fifth International Joint Conference on Autonomous agents and Multiagent Systems: AAMAS'06*, pages 185–192, New York, NY, USA, 2006. ACM.

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. O'Hare, K. Stathis, and G. Vouros, editors, *Proceedings 8th Annual International Workshop "Engineering Societies in the Agents World" (ESAW'07), Athens, October 2007*, LNAI 4995. Springer, 2008. Forthcoming.

Marek Sergot and Robert Craven. The deontic component of action language nC+. In L. Goble and J-J. Ch. Meyer, editors, *Deontic Logic and Artificial Normative Systems. Proc. 8th International Workshop on Deontic Logic in Computer Science (DEON'06), Utrecht, July 2006*, LNAI 4048, pages 222–237. Springer Verlag, 2006.

J. F. van Benthem. Minimal deontic logics. *Bulletin of the Section of Logic*, 8(1): 36–42, 1979.

Y. Venema. Points, lines and diamonds: a two-sorted modal logic for projective planes. *Journal of Logic and Computation*, 9(5):601–621, 1999.

Georg Henrik von Wright. *An essay in deontic logic and the general theory of action*. Number 21 in Acta Philosophica Fennica. 1968.

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

Georg Henrik von Wright. *Practical Reason*. Blackwell, Oxford, 1983.

A Proofs of Theorem 5.9, Section 5.2

As in the other parts of the paper, derivations are presented in a schematic form for clarity, from which full details are easily reconstructed.

Proposition All instances of the following schema are theorems of K^{Ag}.

$$(\mathrm{T})([G]\varphi \to \varphi) \;\leftrightarrow\; (\mathrm{E}_G\varphi \to \varphi)$$

Proof.

$$\begin{aligned}
([G]\varphi \to \varphi) \;&\leftrightarrow\; ((\mathrm{E}_G\varphi \lor [\mathrm{alt}]\varphi) \to \varphi)\\
&\leftrightarrow\; (\mathrm{E}_G\varphi \to \varphi) \land ([\mathrm{alt}]\varphi \to \varphi)\\
&\leftrightarrow\; (\mathrm{E}_G\varphi \to \varphi)
\end{aligned}$$

The last step is because all instances of $[\mathrm{alt}]\varphi \to \varphi$ are theorems of K^{Ag}. □

Proposition All instances of the following schema are theorems of K^{Ag}.

$$(4\mathrm{a}) \qquad ([\mathrm{alt}][G]\varphi \to [\mathrm{alt}]\varphi) \to (\,([G]\varphi \to [G][G]\varphi) \to (\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi)\,)$$

Proof. All instances of the following schemas are theorems of K^{Ag}.

$$\begin{array}{lc}
(\mathrm{i}) & \neg[\mathrm{alt}]\varphi \to [G]\neg[\mathrm{alt}]\varphi\\
(\mathrm{ii}) & ([\mathrm{alt}][G]\varphi \to [\mathrm{alt}]\varphi) \;\leftrightarrow\; ([\mathrm{alt}]\mathrm{E}_G\varphi \to [\mathrm{alt}]\varphi)
\end{array}$$

(i) follows from $\neg[\mathrm{alt}]\varphi \to [\mathrm{alt}]\neg[\mathrm{alt}]\varphi$ and $[\mathrm{alt}]\neg[\mathrm{alt}]\varphi \to [G]\neg[\mathrm{alt}]\varphi$.

(ii) can be derived as follows:

$$\begin{aligned}
&([\mathrm{alt}]\mathrm{E}_G\varphi \to [\mathrm{alt}]\varphi)\\
&\quad\leftrightarrow\; (\,([\mathrm{alt}]([G]\varphi \land \neg[\mathrm{alt}]\varphi) \to [\mathrm{alt}]\varphi\,)\\
&\quad\leftrightarrow\; (\,([\mathrm{alt}][G]\varphi \land [\mathrm{alt}]\neg[\mathrm{alt}]\varphi) \to [\mathrm{alt}]\varphi\,) && ([\mathrm{alt}]\text{ normal})\\
&\quad\leftrightarrow\; (\,([\mathrm{alt}][G]\varphi \land \neg[\mathrm{alt}]\varphi) \to [\mathrm{alt}]\varphi\,) && ([\mathrm{alt}]\neg[\mathrm{alt}]\varphi \leftrightarrow \neg[\mathrm{alt}]\varphi)\\
&\quad\leftrightarrow\; (\,[\mathrm{alt}][G]\varphi \to [\mathrm{alt}]\varphi\,) && (\text{by (i)})
\end{aligned}$$

Now a derivation of (4a) (or rather, of a schema equivalent to (4a)):

$$\begin{aligned}
&(\neg[\mathrm{alt}]\varphi \to \neg[\mathrm{alt}][G]\varphi) \land ([G]\varphi \to [G][G]\varphi) \to\\
&\quad(\,\mathrm{E}_G\varphi \;\to\; [G]\varphi \land \neg[\mathrm{alt}]\varphi\,)\\
&\qquad\qquad\to\; [G][G]\varphi \land [G]\neg[\mathrm{alt}]\varphi \land \neg[\mathrm{alt}]\varphi\,) && (\text{by (i)})\\
&\qquad\qquad\to\; [G](\,[G]\varphi \land \neg[\mathrm{alt}]\varphi) \land \neg[\mathrm{alt}]\varphi\,) && ([G]\text{ normal})\\
&\qquad\qquad\to\; [G]\mathrm{E}_G\varphi \land \neg[\mathrm{alt}]\varphi\,)\\
&\qquad\qquad\to\; [G]\mathrm{E}_G\varphi \land \neg[\mathrm{alt}]\mathrm{E}_G\varphi\,) && (\text{by (ii)})\\
&\qquad\qquad\to\; \mathrm{E}_G\mathrm{E}_G\varphi\,)
\end{aligned}$$

□

Proposition All instances of the following schema are theorems of K^{Ag}.

$$\text{(4b)} \qquad (\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi) \to ([G]\varphi \to [G][G]\varphi)$$

Proof. All instances of the following schemas are theorems of K^{Ag}.

$$\begin{array}{ll} \text{(i)} & [\mathrm{alt}]\varphi \to [G][G]\varphi \\ \text{(ii)} & \mathrm{E}_G\mathrm{E}_G\varphi \to [G][G]\varphi \end{array}$$

(i) is from $[\mathrm{alt}]\varphi \to [\mathrm{alt}][\mathrm{alt}]\varphi$, $[\mathrm{alt}][\mathrm{alt}]\varphi \to [\mathrm{alt}][G]\varphi$, and $[\mathrm{alt}][G]\varphi \to [G][G]\varphi$.

(ii) can be derived as follows:

$$\begin{array}{rcll} \mathrm{E}_G\mathrm{E}_G\varphi & \to & [G]\mathrm{E}_G\varphi & \\ & \to & [G]([G]\varphi \wedge \neg[\mathrm{alt}]\varphi) & \\ & \to & [G][G]\varphi & ([G] \text{ normal}) \end{array}$$

Derivation of (4b) comes from (i) and (ii) above:

$$\begin{array}{rcl} (\mathrm{E}_G\varphi \to \mathrm{E}_G\mathrm{E}_G\varphi) & \to & (\, [G]\varphi \to \mathrm{E}_G\varphi \vee [\mathrm{alt}]\varphi\,) \\ & & \to \mathrm{E}_G\mathrm{E}_G\varphi \vee [\mathrm{alt}]\varphi\,) \\ & & \to [G][G]\varphi\,) \end{array}$$

□

Proposition All instances of the following schema are theorems of K^{Ag}.

$$\text{(5)} \qquad (\neg[G]\varphi \to [G]\neg[G]\varphi) \;\leftrightarrow\; (\neg\mathrm{E}_G\varphi \to [G]\neg\mathrm{E}_G\varphi)$$

Proof. All instances of the following schema are theorems of K^{Ag}.

$$\text{(i)} \qquad \langle G\rangle\neg[\mathrm{alt}]\varphi \to \neg[\mathrm{alt}]\varphi$$

(i) follows from $[\mathrm{alt}]\varphi \to [\mathrm{alt}][\mathrm{alt}]\varphi$ and $[\mathrm{alt}][\mathrm{alt}]\varphi \to [G][\mathrm{alt}]\varphi$.

The following schema is equivalent to the left-to-right half of (5):

$$(\langle G\rangle[G]\varphi \to [G]\varphi) \to (\langle G\rangle\mathrm{E}_G\varphi \to \mathrm{E}_G\varphi)$$

It can be derived as follows:

$$\begin{array}{rcll} (\langle G\rangle[G]\varphi \to [G]\varphi) & \to & (\,\langle G\rangle\mathrm{E}_G\varphi \to \langle G\rangle([G]\varphi \wedge \neg[\mathrm{alt}]\varphi)\,) & \\ & & \to \langle G\rangle[G]\varphi \wedge \langle G\rangle\neg[\mathrm{alt}]\varphi\,) & ([G] \text{ normal}) \\ & & \to [G]\varphi \wedge \neg[\mathrm{alt}]\varphi\,) & (\text{by (i)}) \\ & & \to \mathrm{E}_G\varphi\,) & \end{array}$$

For the other half of (5): all instances of the following schema are theorems of K^{Ag}.

$$\text{(ii)} \qquad \neg[\mathrm{alt}]\varphi \to [G]\neg[\mathrm{alt}]\varphi$$

(ii) follows from $\neg[\mathrm{alt}]\varphi \to [\mathrm{alt}]\neg[\mathrm{alt}]\varphi$ and $[\mathrm{alt}]\neg[\mathrm{alt}]\varphi \to [G]\neg[\mathrm{alt}]\varphi$.

The right-to-left half of (5) can be derived as follows:

$$\begin{array}{rcll} (\neg\mathrm{E}_G\varphi \to [G]\neg\mathrm{E}_G\varphi) & \to & (\,\neg[G]\varphi \to \neg\mathrm{E}_G\varphi \wedge \neg[\mathrm{alt}]\varphi\,) & \\ & & \to [G]\neg\mathrm{E}_G\varphi \wedge [G]\neg[\mathrm{alt}]\varphi\,) & (\text{by (ii)}) \\ & & \to [G](\neg\mathrm{E}_G\varphi \wedge \neg[\mathrm{alt}]\varphi)\,) & ([G] \text{ normal}) \\ & & \to [G]\neg\mathrm{E}_G\varphi\,) & \end{array}$$

□