# Action and agency in norm-governed multi-agent systems

Marek Sergot

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

**Abstract.** There is growing interest in the idea that, in some cases, interactions among multiple, independently acting agents in a multi-agent system can be regulated and managed by norms (or 'social laws') which, if respected, allow the agents to co-exist in a shared environment. We present a formal (modal-logical) language for describing and analysing such systems. We distinguish between system norms, which express a system designer's view of what system behaviours are deemed to be legal, permitted, desirable, and so on, and agent-specific norms which constrain and guide an individual agent's behaviours and which are supposed to be incorporated, in one way or another, in the agent's implementation. The language provides constructs for expressing properties of states and transitions in a transition system, and modalities of the kind found in logics of action/agency for expressing that an agent brings it about that, or is responsible for, its being the case that $A$. The novel feature is that an agent, or group of agents, brings it about that a transition has a certain property rather than bringing it about that a certain state of affairs obtains, as is usually the case. The aim of the paper is to motivate the technical development and illustrate the use of the formal language by means of a simple example in which there there are both physical and normative constraints on agents' behaviours. We discuss some relationships between system norms and agent-specific norms, and identify several different categories of non-compliant behaviour that can be expressed and analysed using the formal language. The final part of the paper presents some transcripts of output from a model-checker for the language.

## 1 Introduction

There has been growing interest in recent years in norm-governed multi-agent systems. References to normative concepts (obligation, permission, commitment, social commitment, ...) feature prominently in the literature. One reason for this interest is clear, for there are important classes of applications, in e-commerce, contracting, trading, e-government, and so on, where the domain of application is defined by and regulated by laws, regulations, codes of practice, and standards of various kinds whose existence is an essential ingredient of any application. Another, somewhat different, motivation is the idea that, in some cases, agent

interactions generally 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. The question of what happens to system behaviour when norms or social laws are not respected, however, has received little or no serious attention. It is also not entirely clear from works in this area whether these norms are intended to express only the system designer's view of what behaviours are legal, permitted, desirable, and so on, or whether they are supposed to be taken into account, explicitly or implicitly, in the implementation of the agents themselves, or both.

In a recent paper [1] we presented a formal framework, called there a 'coloured agent-stranded transition system', which adds two components to a labelled transition system. The first component partitions states and transitions according to various 'colourings', used to represent norms (or 'social laws'), of two different kinds. *System norms* express a system designer's point of view of what system states and system transitions are legal, permitted, desirable, and so on. A separate set of individual *agent-specific* norms are intended to guide or constrain an individual agent's behaviours. They are assumed to be taken into account in the agent's implementation, or in the case of deliberative agents with reasoning and planning capabilities, in the processes an agent uses to determine its choice of actions to be performed. The second component of a 'coloured agent-stranded transition system' is a way of picking out, from a global system transition representing many concurrent actions by multiple agents and possibly the environment, an individual agent's actions, or 'strand', in that transition. This is to enable us to say that in a particular transition it is specifically one agent's actions that are in compliance or non-compliance with a system or agent-specific norm rather than some other's. This framework allowed us in turn to identify and characterise several different categories of non-compliant behaviour, distinguishing 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. The other general aim, which is to consider how agent-specific norms can be incorporated into an agent's implementation, was not discussed. It is a topic of current work.

This paper presents a further development and refinement of those ideas. Specifically, we now prefer to separate the 'colourings' used to represent norms from the more general structure of an agent-stranded transition system. We present a formal (modal-logical) language for talking about properties of states and transitions, including but not restricted to their 'colourings', and for talking

about agent strands of transitions. The language has operators for expressing that a particular agent, or group of agents, *brings it about* that such-and-such is the case, in the sense that it is responsible for, or its actions are the cause of, such-and-such being the case. The resulting logic bears a strong resemblance to Ingmar Pörn's (1977) logic of 'brings it about' action/agency [2], except that we switch from talking about an agent's bringing about a certain state of affairs to an agent's bringing it about that a transition has a certain property. The general aim of the paper is to motivate the technical development and illustrate something of the expressiveness of the formal language. We use the same, rather simple, example discussed in the earlier paper [1] but present it now in terms of the new formal system. Technical details of the logic, comparisons with other works in the logic of action/agency, and discussion of various forms of collective or group agency are beyond the scope of this paper. These topics are covered elsewhere [3].

It is important to stress that we make *no assumptions* 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 here. 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 only what can be said about individual or collective agency when all such considerations are ignored. The result might be termed 'a logic of unwitting (collective) agency'—'unwitting' means both inadvertent and unaware. The logic of unwitting agency might be extended and strengthened in due course by bringing in other considerations such as (joint) intention; we do not discuss any such possibilities here. From the practical point of view, there is clearly 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.

The formal language presented here has been implemented, in the form of a model-checker that can be used to evaluate formulas on a given transition system. It is included as part of the ICCALC system[1], which at its core is a re-implementation of the 'Causal Calculator' CCALC[2] developed at the University of Texas and made available as a means of performing computational tasks using the action language $\mathcal{C}+$. $\mathcal{C}+$ [4] is a formalism for defining transition systems of a certain kind. It provides a treatment of default persistence ('inertia'), non-deterministic and concurrent actions, and indirect effects of actions ('ramifications'). CCALC can be used (among other things) to generate (a symbolic representation of) a transition system defined by means of $\mathcal{C}+$ laws. ICCALC re-

[1] `http://www.doc.ic.ac.uk/~rac101/iccalc/`
[2] `http://www.cs.utexas.edu/users/tag/cc`

tains the core functionality of CCALC, and the core implementation techniques, and adds a number of other features, such as the ability to pass the transition system to standard CTL model checking systems (specifically NuSMV). ICCALC also supports a number of extended forms of $\mathcal{C}+$, of which the language $n\mathcal{C}+$ is the most relevant here. $n\mathcal{C}+$ [5, 6] is an extended form of $\mathcal{C}+$ designed specifically for representing simple normative and institutional concepts. An action description in $n\mathcal{C}+$ defines a coloured (agent-stranded) transition system of a certain kind. The examples discussed in this paper are constructed by formulating them as $n\mathcal{C}+$ action descriptions, using ICCALC to generate (a symbolic representation of) the transition system so defined, and then passing the transition system to the model checker that evaluates formulas of the language presented in this paper. However, the framework presented in this paper is more general, and is *not* restricted to transition systems of the kind defined by $\mathcal{C}+$ or $n\mathcal{C}+$.

## 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 $(s, \varepsilon, s')$ is a transition in $R$, $s$ is the initial state and $s'$ is the resulting state, or end state, of the transition. $\varepsilon$ is *executable* in a state $s$ when there is a transition $(s, \varepsilon, s')$ in $R$, and *non-deterministic* in $s$ when there are transitions $(s, \varepsilon, s')$ and $(s, \varepsilon, s'')$ in $R$ with $s' \neq s''$. A *path* or *run* of length $m$ of the labelled transition system $\langle S, A, R\rangle$ is a sequence $s_0\, \varepsilon_0\, s_1 \cdots s_{m-1}\, \varepsilon_{m-1}\, s_m \quad (m \geq 0)$ such that $(s_{i-1}, \varepsilon_{i-1}, s_i) \in R$ for $i \in 1..m$. 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 $\tau$, 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 $\tau_1 \cdots \tau_{m-1}\, \tau_m \quad (m \geq 0)$ 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* (LTS) 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 $prev(\tau) = prev(\tau')$ and $post(\tau) = post(\tau')$ and $label(\tau) = label(\tau')$ implies $\tau = \tau'$, and in which $prev((s, \varepsilon, s')) = s$, $post((s, \varepsilon, s')) = s'$, and $label((s, \varepsilon, s')) = \varepsilon$. The more abstract account is of little practical significance but is helpful in that it allows a more concise statement of some things we want to say about transition systems. It is also more general: transitions are not identified by $(s, \varepsilon, s')$ 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 $\langle S, A, R, prev, post, label\rangle$ 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 $h^{\mathrm{f}}$ 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 $\sigma^{\mathrm{f}}$ of propositional atoms for expressing properties of states, and a disjoint set $\sigma^{\mathrm{a}}$ 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 $h^{\mathrm{f}}$ is a valuation function for atomic propositions $\sigma^{\mathrm{f}}$ in states $S$ and $h^{\mathrm{a}}$ is a valuation function for atomic propositions $\sigma^{\mathrm{a}}$ 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 $\varphi$ 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 $\varphi$. The transition formulas $0{:}F$ and $1{:}G$ are satisfied by a transition $\tau$ when the initial state of $\tau$ satisfies state formula $F$ and the resulting state of $\tau$ satisfies state formula $G$, respectively. The details are summarised presently.

It is not clear whether evaluating formulas on transitions in this fashion is novel or not. Große and Khalil [7] evaluate formulas on state-event pairs $(s, \varepsilon)$ when the transition system is a set of triples $(s, \varepsilon, s')$ but that is not the same as we have here. Venema [8] 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.

We also find it convenient to add a little more structure to the underlying propositional language. This is not essential but makes the formulation of typical examples clearer and more concise. It is also the propositional language that is supported by $\mathcal{C}+$ and $n\mathcal{C}+$, and the CCalc and iCCalc implementations.

*Multi-valued signatures* The following is adapted from [4]. A *multi-valued propositional signature* $\sigma$ is a set of symbols called *constants*. For each constant $c$ in $\sigma$ there is a non-empty set $dom(c)$ of values called the *domain* of $c$. For simplicity, in this paper we will assume that each $dom(c)$ is finite and has at least two elements. An *atom* of a signature $\sigma$ is an expression of the form $c{=}v$ where $c$ is a

constant in $\sigma$ and $v \in dom(c)$. A *formula* of signature $\sigma$ is any truth-functional compound of atoms of $\sigma$.

A *Boolean* constant is one whose domain is the set of truth values $\{\mathsf{t}, \mathsf{f}\}$. If $c$ is a Boolean constant, $c$ is shorthand for the atom $c{=}\mathsf{t}$ and $\neg c$ for the atom $c{=}\mathsf{f}$. More generally, if $c$ is a constant whose domain is $\{v_1, \ldots, v_n, \mathsf{f}\}$, then by convention we write $\neg c$ as shorthand for the atom $c{=}\mathsf{f}$.

An *interpretation* of a multi-valued signature $\sigma$ is a function that maps every constant $c$ in $\sigma$ to some value $v$ in $dom(c)$; an interpretation $I$ *satisfies* an atom $c{=}v$ if $I(c) = v$. We write $\mathrm{I}(\sigma)$ for the set of interpretations of $\sigma$.

As observed in [4], a multi-valued signature of this type can always be translated to an equivalent Boolean signature. Use of a multi-valued signature makes the formulation of examples more concise.

### Syntax and semantics

The base propositional language is constructed from a set $\sigma^{\mathrm{f}}$ of *state constants* (also known as 'fluents' or 'state variables') and a disjoint set $\sigma^{\mathrm{a}}$ of *event constants*. In previous work we followed the terminology of [4] and called the constants of $\sigma^{\mathrm{a}}$ 'action constants'. This terminology is misleading however. Although event constants *are* used to name actions and attributes of actions, they are also used to express properties of an event or transition as a whole. An example of an event constant might be $x{:}move$ with domain $\{l, r, \mathsf{f}\}$ : the atom $x{:}move{=}l$ represents that agent $x$ moves in direction $l$, $x{:}move{=}r$ that $x$ moves in direction $r$, and $\neg x{:}move$ (which, recall, is shorthand for $x{:}move{=}\mathsf{f}$) that $x$ does not move in a given transition. In ICCALC we employ an (informal) convention that event constants with a prefix '$x$:' are intended to represent actions by an agent $x$. The (Boolean) event constant $falls(vase)$ might be used to represent transitions in which the object *vase* falls from a table to the ground (say). Here there is no prefix '*vase*:'—'falls' is not an action that is meaningfully performed by the object *vase*. Event constants are also used to express properties of a transition as whole, for instance, whether it is desirable or undesirable, timely or untimely, permitted or not permitted, and so on. For this reason we prefer the term 'event constant' for the elements of $\sigma^{\mathrm{a}}$, and we reserve the term 'action constant' for referring informally to those event constants that are intended to represent actions by an agent. In general, an event (or transition label) will represent multiple concurrent actions by agents and the environment, concurrent actions, such as the falling of an object, that cannot be ascribed to any agent, and other properties of the event, such as whether it is desirable or undesirable, desirable or undesirable from the point of view of an agent $x$, timely or untimely, and so on.

For example, the formula

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

might represent an event in which $a$ moves to the left, $b$ does not move to the left, $c$ does not move at all, and the object *vase* falls. The atom $trans{=}red$ might represent that the event is illegal (say), or undesirable, or not permitted.

Propositional formulas of $\sigma^{\mathrm{a}}$ are evaluated on transition labels/events. When an event satisfies a propositional formula $\varphi$ of $\sigma^{\mathrm{a}}$ we say that the event is an event of type $\varphi$. So, all events of type $a{:}move{=}l \wedge \neg\, c{:}move$ are also events of type $a{:}move{=}l$, and events of type $\neg\, c{:}move$, and so on. By extension, we also say that a transition is of type $\varphi$ when its label (event) is of type $\varphi$. 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 $\mathcal{C}+$ [4], for example, and has also been used recently in [9] in discussions of agent 'ability'.

*Formulas* Formulas are state formulas and transition formulas.

*State formulas:*

$$F \;::=\; \top \mid \bot \mid \text{any atom } f{=}v \text{ of } \sigma^{\mathrm{f}} \mid \neg F \mid F \wedge F \mid [\varphi]F$$

*Transition formulas:*

$$\varphi \;::=\; \top \mid \bot \mid \text{any atom } a{=}v \text{ of } \sigma^{\mathrm{a}} \mid \neg\varphi \mid \varphi \wedge \varphi \mid 0{:}F \mid 1{:}F$$

where $F$ is any propositional state formula (i.e., a propositional formula of $\sigma^{\mathrm{f}}$). We refer to the propositional formulas of $\sigma^{\mathrm{a}}$ 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 $h^{\mathrm{f}}$ and $h^{\mathrm{a}}$ are the valuation functions for state constants and event constants, respectively:

$$h^{\mathrm{f}}\colon S \rightarrow \mathrm{I}(\sigma^{\mathrm{f}}) \qquad \text{and} \qquad h^{\mathrm{a}}\colon A \rightarrow \mathrm{I}(\sigma^{\mathrm{a}})$$

$h^{\mathrm{f}}(s)$ is an interpretation of $\sigma^{\mathrm{f}}$, i.e., a function which assigns to every constant $f$ in $\sigma^{\mathrm{f}}$ a value $v$ in $dom(f)$, and $h^{\mathrm{a}}(\varepsilon)$ is an interpretation of $\sigma^{\mathrm{a}}$, i.e., a function which assigns to every constant $a$ in $\sigma^{\mathrm{a}}$ a value $v$ in $dom(a)$. Accordingly, for every state $s$ in $S$ and event/label $\varepsilon$ in $A$ we have:

$$\begin{aligned} \mathcal{M}, s \models f{=}v \quad &\text{iff} \quad h^{\mathrm{f}}(s)(f) = v \\ \mathcal{M}, \varepsilon \models a{=}v \quad &\text{iff} \quad h^{\mathrm{a}}(\varepsilon)(a) = v \end{aligned}$$

and for every transition $\tau$ in $R$:

$$\mathcal{M}, \tau \models a{=}v \quad \text{iff} \quad \mathcal{M}, label(\tau) \models a{=}v$$

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

$$\mathcal{M}, t \models a{=}v \quad \text{iff} \quad h^R(\tau)(a) = v$$

We will not bother with that extension here. Event constants in $\sigma^{\mathrm{a}}$ 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 $\sigma^{\mathrm{a}}$ whereas transition formulas are more general (as defined above). Transition formulas will be extended with some additional constructs in Sect. 6.

When $\varphi$ is a formula of $\sigma^{\mathrm{a}}$ and $\tau$ is a transition in $R$ we say that $\tau$ is a transition of type $\varphi$ when $\tau$ satisfies $\varphi$, i.e., when $\mathcal{M}, \tau \models \varphi$, and sometimes that $\varphi$ is true at, or true in, the transition $\tau$. 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 $\tau$ 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 $\varphi$.

*State formulas:*

$$\mathcal{M}, s \models [\varphi]F \quad \text{iff} \quad \mathcal{M}, \tau \models \varphi \text{ for every } \tau \in R \text{ such that } \mathit{prev}(\tau) = s.$$

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

*Transition formulas:*

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

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

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 $\mathcal{M}$, and $\varphi$ is *valid* in a model $\mathcal{M}$, written $\mathcal{M} \models \varphi$, when $\mathcal{M}, \tau \models \varphi$ for every transition $\tau$ in $\mathcal{M}$. A formula is *valid* if it is valid in every model $\mathcal{M}$ (written $\models F$ and $\models \varphi$, respectively).

$\mathcal{C}+$ [4] is a language for defining (a certain class of) transition systems of this type. The iCCalc implementation can be used to evaluate state, event, and transition formulas on transition systems defined by $\mathcal{C}+$ though it is not restricted to transition systems of that type.

Let us discuss the transition formulas first. 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 [10] 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 Sect. 6.4.

For example, let the state atom $\mathit{on\text{-}table}(\mathit{vase})$ represent that a certain vase is standing on a table. A transition of type $0{:}\,\mathit{on\text{-}table}(\mathit{vase}) \wedge 1{:}\neg\,\mathit{on\text{-}table}(\mathit{vase})$, equivalently, of type $0{:}\,\mathit{on\text{-}table}(\mathit{vase}) \wedge \neg 1{:}\,\mathit{on\text{-}table}(\mathit{vase})$ is one from a state in which the vase is on the table to one in which it is not on the table. Suppose that the event atom $\mathit{falls}(\mathit{vase})$ 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 domain will have the validity

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

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

The operators 0: and 1: are both normal[3]. 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 $\varphi$ is a transition formula, then $[\varphi]F$ is true at a state $s$ when every transition of type $\varphi$ from state $s$ results in a state where $F$ is true. $\langle\varphi\rangle F$ is true at a state $s$ when there exists at least one transition of type $\varphi$ from state $s$ whose resulting state satisfies $F$. $[\varphi]\bot$, equivalently $\neg\langle\varphi\rangle\top$, says that there is no transition of type $\varphi$ from the current state, and $\neg[\varphi]\bot$, equivalently $\langle\varphi\rangle\top$, that there is a transition of type $\varphi$ from the current state. When $\alpha$ is an event formula, that is, a propositional formula of $\sigma^{\mathrm{a}}$, then $\langle\alpha\rangle\top$, equivalently, $\neg[\alpha]\bot$ represents that an event of type $\alpha$ is executable in the current state.

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 $\varepsilon$ in an

[3] This is standard terminology. See e.g. [11, 12] or any introductory text on modal logic.

expression $[\varepsilon]F$ is a transition label/event $\varepsilon$ 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 $F \rightarrow [\varphi]G$ and the second to $F \wedge \langle\varphi\rangle G$.

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

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

as is easily confirmed, and hence

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

We also have validity of:

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

and of

$$[\bot]\bot$$

Sauro et al. [9] 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 $\sigma^{\mathrm{f}}$ and $\alpha$ is an event formula, that is, a propositional formula of $\sigma^{\mathrm{a}}$. We will not present 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

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

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 $\varphi$ 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 $\varphi$ from that state leads to a state which satisfies $G$. That is:

$$\mathcal{M} \models (0{:}F \wedge \varphi) \rightarrow 1{:}G$$

Properties of models can often be expressed equivalently as validities of state formulas or of transition formulas. This is because:

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

For example, suppose 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 $p$. Suppose that the (Boolean) event constant *toggle* represents that the light switch is toggled, and event constants $x$:*move* with domain $\{l, r, \mathsf{f}\}$ that agent $x$ moves in the direction $l$, $r$, or stays where it is. A model $\mathcal{M}$ modelling this domain would have the properties:

– state formulas

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

– transition formulas

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

We find transition formulas are generally more useful and clearer.

### 2.2 Norms and Coloured Transition Systems

A simple way of representing norms is to partition the states and transitions of a transition system into two categories. A *coloured transition system* [5, 6] is a structure of the form $\langle S, A, R, S_g, R_g\rangle$ where $\langle S, A, R\rangle$ is a labelled transition system of the kind discussed above, and where the two new components are

– $S_g \subseteq S$, 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. For example, Carmo and Jones [13] employ a structure which has both ideal/sub-ideal states and ideal/sub-ideal transitions (unlabelled). van der Meyden's 'Dynamic logic of permission' [14] employs a structure in which transitions, but not states, are classified as 'permitted/non-permitted'. van der Meyden's version was constructed as a response to problems of Meyer's 'Dynamic deontic logic' [15] which classifies transitions as 'permitted/non-permitted' by reference to the state resulting from a transition. 'Deontic interpreted systems' [16] classify states as 'green'/'red', where these states have further internal structure to model the local states of agents in a multi-agent context. Recently, Ågotnes et al. [17] have presented a language based on the temporal logic CTL. They partition transitions into those that comply with a set of norms and those that do not (that is, into 'green' and 'red' in our terminology). They then define a modified form of CTL for expressing temporal properties of paths/runs in which every transition

is 'green', or what we refer to as 'fully compliant behaviour' in Sect. 4.1 below. There are no constructs in the language for expressing properties of paths/runs in which some transition is not 'green'.

We require that a coloured transition system $\langle S, A, R, S_g, R_g \rangle$ must further satisfy the constraint that, for all states $s$ and $s'$ in $S$ and all transitions $\tau$ 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.

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 (i.e., that there is at least one transition from every state), or that there must be at least one green state, or that from every green state there must be at least one green transition, or that from every green state reachable from some specified initial state(s) there must be at least one green transition, and so on. These are examples of properties that might be of interest when analyzing a transition system. We can check for them but we do not assume they are always satisfied. We do assume that every coloured transition systems satisfies the *ggg* constraint.

Instead of introducing a special category of coloured transition systems, with extra components $S_g$ and $R_g$, we now prefer to speak of labelled transition systems generally and introduce colourings for states and transitions by means of suitably chosen constants in $\sigma^{\mathrm{f}}$ and $\sigma^{\mathrm{a}}$. This is more general and adds flexibility. In particular, we have a state constant *status* and an event constant *trans* both with domain $\{\mathit{green}, \mathit{red}\}$. The intended reading is that $\|\mathit{status}{=}\mathit{green}\|^{\mathcal{M}}$ denotes the 'green states' and $\|\mathit{status}{=}\mathit{red}\|^{\mathcal{M}} = S \setminus \|\mathit{status}{=}\mathit{green}\|^{\mathcal{M}}$ the '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 any 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

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

As further illustrations of the use of the language, here are the other properties mentioned earlier, expressed now as validities in a model $\mathcal{M}$.

- the transition relation must be serial

$$\mathcal{M} \models \langle\top\rangle\top$$

- there must be at least one green state

$$\mathcal{M} \not\models \mathit{status}{=}\mathit{red}, \quad \text{equivalently, } \mathcal{M} \not\models \neg(\mathit{status}{=}\mathit{green})$$

- from every green state there must be at least one green transition

$$\mathcal{M} \models \mathit{status}{=}\mathit{green} \rightarrow \langle \mathit{trans}{=}\mathit{green}\rangle\top$$

We cannot express, in this language, that from every green state reachable from some specified initial state(s) there must be at least one green transition since we have no way of expressing reachability (in the language). That could be fixed by extending the language but we will not do it here. Reachability properties in a model can be checked using the iCCalc system but are not expressible as formulas of the language.

$n\mathcal{C}+$ [5, 6] is a language for defining (a certain class of) transition systems of this type. The iCCalc implementation builds in the special treatment of 'red' and 'green' required to ensure that the *ggg* constraint is satisfied.

In [6] we presented a refinement where instead of the binary classification of states as red or green, states are ordered according to how well each complies with the state permission laws of an $n\mathcal{C}+$ action description. We also discussed possible generalisations of the *ggg* constraint for that case. In the current paper, we keep to the simple classification of states as green or red.

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.

## 3 Example (Rooms)

This example concerns the specification of norm-governed interactions between independently acting agents. It was discussed in a previous paper [1]. We now present it using the formalism introduced in previous sections.

In the example there are two categories of agents, male and female, who move around in a world of interconnecting rooms. The rooms are connected by doorways through which agents may pass. (The precise topography, and number of rooms, can vary.) Each doorway connects two rooms. Rooms can contain any number of male and female agents. The action atoms of $\sigma^{\mathrm{a}}$ will take the form $x{:}\mathit{move}{=}p$, where $x$ ranges over the agents in a particular example, and $p$ ranges

over a number of values representing directions in which agents can move, in addition to a value f: if a transition satisfies $x{:}move{=}\mathsf{f}$, that is to be taken to represent that agent $x$ does not move during that transition. Recall that by convention we write $\neg x{:}move$ as a shorthand for $x{:}move{=}\mathsf{f}$.

A normative element is introduced by insisting that a male agent and a female agent may never be alone together in a room; such configurations are physically possible, and the transition system will include states representing them, but all such states will be coloured red.

Although this example is relatively simple, it shares essential features with a number of real-world domains, in which there are large numbers of interacting agents or components which may be in different states, and where some of those combinations of states are prohibited. (These real-world examples are *not* restricted to domains where agents perform physical actions. Exactly the same points could be made for examples of institutions or virtual organisations, where the possible actions by agents are defined and constrained by institutional rules rather than physical constraints, and where actions by agents can be represented as transitions from one institutional state to another.)

For the purposes of illustration, we shall consider a concrete instance of the example in which there are just two rooms, on the left and right, with one connecting door, and three agents, two males $m_1$ and $m_2$, and a female $f_1$. We have deliberately made the example simple in order to concentrate on its essential features, and so that we can depict the transition system in its entirety. With more agents and more rooms the transition system is too big to be shown easily in diagrammatic form. We will also impose an additional constraint that only one agent can move through the doorway at once (the doorways are too narrow to let more than one agent pass through at the same time). This is a more significant restriction since it imposes constraints on possible interactions between the agents: if an agent moves from one room to another it thereby makes it impossible for other agents to pass through the same doorway.

The propositional language for this instance of the 'rooms' example contains state atoms $loc(x){=}l$ and $loc(x){=}r$, where $x$ ranges over $m_1, m_2, f_1$; $loc(m_1){=}l$ is true when the male agent $m_1$ is in the left-hand room, $loc(m_2){=}r$ is true when $m_2$ is in the right-hand room, and so on. The action atoms are, in line with previous remarks, $x{:}move{=}p$, where $x$ ranges over the agents and $p$ ranges over $l, r, \mathsf{f}$.

We do not show the $n\mathcal{C}+$ formulation of the example here. (It can be found in [1].) The transition system, whether defined using $n\mathcal{C}+$ or by some other means, is depicted in Fig. 1. We have not drawn the transitions from states to themselves, where no agent moves, in order to keep the drawing clear; all such transitions are coloured green. Also, we have not shown labels for transitions. These can easily be deduced, for every arc in the diagram should have a label which makes precisely one $x{:}move{=}p$ atom true (for $p$ one of $l, r, \mathsf{f}$); for example, the (red) transition from the top-most state (`s1`) to the one immediately below and to the right (`s6`) has a label which makes $m_2{:}move{=}r$, $\neg m_1{:}move$, and $\neg f_1{:}move$ true.

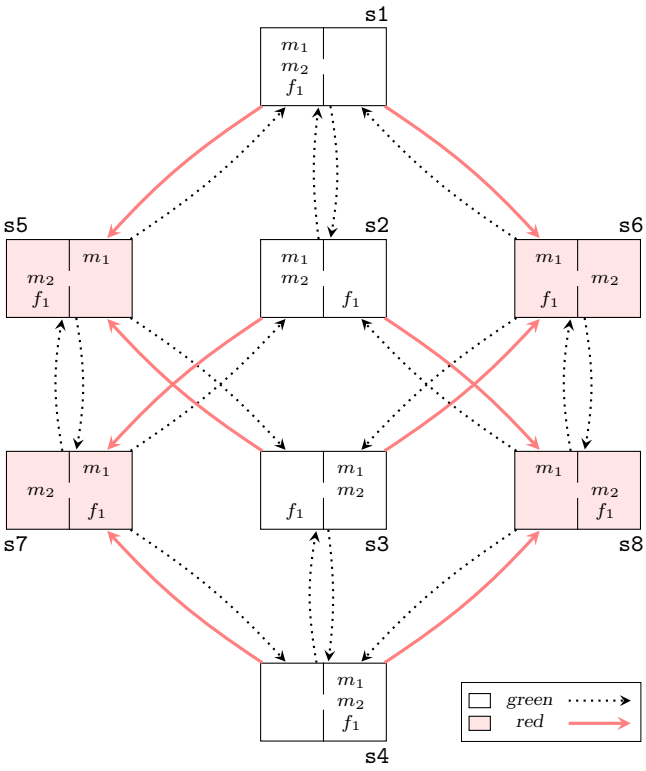


**Fig. 1.** A simple 'rooms' example.

One can see that the transition system satisfies the *ggg* constraint: since nothing further was said about the colouring of transitions, the red transitions are simply those where the system moves from a green state to a red state, i.e., to a state in which a male and a female are alone in a room together.

However, the example is also intended to demonstrate some important inadequacies. For consider, again, the transition from the top-most state (`s1`) where all agents are in the left-hand room, to the state below and to the right of it (`s6`) where $m_1$ and $f_1$ are left alone together after $m_2$ has exited to the right. In some sense, it is $m_2$ who has acted wrongly: $m_2$ has left the room, leaving $m_1$ and $f_1$ alone together, in a configuration which thus violates the norms governing the system. On the other hand, if we remove the restriction that at most one agent can pass through the doorway at one time, it is far from clear which of the three agents, if any, acted wrongly when $m_2$ exited: it might have been $m_2$ who acted badly, or it might have been $m_1$, who should have followed $m_2$ out, or it might have been $f_1$, who should have followed $m_2$ out. Or it might be that all of them, collectively, acted wrongly, or perhaps none of them. 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; indeed, there is no explicit concept of an individual agent in the semantics at all.

## 4 Agent-specific Norms

The language of the previous sections provides a means of representing when states and transitions satisfy, or fail to satisfy, a standard of legality, acceptability, desirability, and so on. Much can be said using the resources of this language. However, in representing systems in which there are multiple interacting agents (as with the simple 'rooms' example depicted in the previous section), it is often essential to be able to speak about an individual agent's behaviour: in particular, about whether individual agents' actions are in the right or wrong—whether they are conforming to norms which govern specifically *their* behaviour. In [1] we introduced a semantical structure which we called a *coloured agent-stranded transition system*. That had two components: a way of picking out an individual agent $x$'s actions from a transition—the agent $x$'s 'strand' in that transition, and a colouring of each such strand as $\mathit{green}(x)$ or $\mathit{red}(x)$ to represent the agent-specific norms for $x$. We will deal first with the agent-specific colourings $\mathit{green}(x)$ and $\mathit{red}(x)$ and defer discussion of the 'strand' component until Sect. 5.

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 *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 'rooms' example an agent-specific norm could not meaningfully prohibit an agent from acting in such a way that a male and female are alone in a room together. The agent cannot predict how other agents will act: just because a room is currently vacant, for example, does not mean that another agent will not enter it.

We now extend the 'rooms' example with some agent-specific norms. As a concrete example (one of many that could be devised) let us attempt to specify an (imperfect) protocol for recovery from red system states: whenever a male agent and a female agent are alone in a room, anywhere, every male agent is required to move to the room to its left (if there is one), and every female agent is required to move to the room to its right (if there is one).

Let $Ag$ be a finite set of agent names. In the present example $Ag = \{m_1, m_2, f_1\}$. For each agent $x \in Ag$, $\mathit{green}(x)$ is a subset of $R$ representing those transi-

tions where the actions of $x$ have been in accordance with norms specific for $x$. $red(x) = R \setminus green(x)$ are those transitions in which the actions of $x$ have failed to conform to $x$'s norms.

So in the example: suppose $s$ is a state of the system in which there is a male agent and a female agent alone in a room. For every male agent $x$ (anywhere), a transition from $s$ in which $x$ moves to the room on its left is coloured $green(x)$, a transition from $s$ in which $x$ stays where it is when there is no room to its left is $green(x)$, and any other transition from $s$ is $red(x)$. And similarly for female agents, but with 'left' replaced by 'right'. Further (let us suppose) in a state $s$ of the system where there is not a male agent and a female agent alone in a room, for any agent $x$, any transition from $s$ is $green(x)$. Thus, the agents are free to move around from room to room, but if ever the system enters a red global state, their individual norms require them to move to the left or right as the case may be; once the system re-enters a green global state they are free to move around again.

The precise mechanism by which agents detect that there is a male agent and a female agent alone in a room somewhere is not modelled at this level of detail. We will simply assume that there is some such mechanism—a klaxon sounds, or a suitable message is broadcast to all agents—the details do not matter for present purposes. Similarly, we are not modelling here how an agent determines which way to move. In a more detailed representation, we could model an agent's internal state, its perceptions of the environment in which it operates, how it determines where to move, and the mechanism by which it perceives that there is a male agent and a female agent alone in a room. We will not do so here: the simpler model is sufficient for present purposes.

In general, given a transition system modelling all 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 in a given system state will be designated as $green(x)$ and the others as $red(x)$. In the example as we have it, the agent-specific norms only constrain the agents' actions in a red system state. That is not essential. It is merely a feature of this particular example. A transition is designated as $green(x)$ when $x$'s actions in that transition comply with the agent-specific norms for $x$. We specify, separately, system norms which constrain various combinations of actions by individual agents, or other interactions of interest, by classifying transitions and states as globally red or green. So we have two separate layers of specification: (i) norms specific to agents governing their individual actions, and (ii) norms governing system behaviour as a whole. We are interested in examining the relationships, if any, between these two separate layers. We might be interested in verifying, for example, that all behaviour by agent $x$ compliant with the norms for $x$ guarantees that the system avoids globally red states, or produces only globally green runs, or always recovers from a global red state to a global green state, and so on. This is the setting we have in mind for discussion in this paper. We also want to identify several different categories of non-compliant behaviours, and generally, the conditions under which we can say that it is a particular agent

$x$'s actions that are responsible for, or the cause of, a transition being coloured (globally) red, or more generally, being of type $\varphi$.

As in the case of coloured transition systems discussed earlier, we prefer to speak of transition systems in general, and use suitably chosen event constants to represent the properties of interest. So, let $\sigma^{\mathrm{a}}$ contain (Boolean) event constants *green*$(x)$ for every agent $x \in Ag$, and let *red*$(x)$ be an abbreviation for $\neg$*green*$(x)$. A transition $\tau$ in $R$ 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, 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

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

The coherence constraint (3) is optional and not appropriate in all settings. We will adopt it in the examples discussed below. 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. We will give some examples presently.

There are other, more fundamental constraints that we must place on agent-specific colourings. We defer discussion of those until Sect. 5.

### 4.1 Fully Compliant Behaviour

As suggested above, we might now be interested in examining the relationship between system norms and individual agent-specific norms—in the present example, for instance, to determine whether the agent-specific norms expressed by the *green*$(x)$ specification do have the desired effect of guaranteeing recovery from a red system state to a green system state. Given a coloured transition system representing the system norms and the agent-specific norms, defined by an $n\mathcal{C}+$ action description or by some other means, we focus attention on those paths of the transition system that start at a red system state, and along which every agent always acts in accordance with its norms, i.e., those paths in which every transition is *green*$(x)$ for each of the agents $x$. A natural property to look

for is whether all such paths eventually pass through a green system state; if this property holds, it indicates that the agent-specific norms are doing a good job in ensuring that systems in violation of their global system norms eventually recover to a green state, assuming that all agents follow their individual norms correctly. (There is a further natural requirement: in the case where the system is initially in a red system state $s$, there should be at least one transition from that state. Otherwise, the requirement that all paths starting at $s$ eventually reach a green system state would be vacuously satisfied.)

In particular applications, it might not be a reasonable assumption to make that agents always act in accordance with their individual norms. This might be for several reasons. Sometimes physical constraints in the environment being modelled prevent joint actions in which all agents act well; in other circumstances, and noteworthy especially because we have in mind application areas in multi-agent systems, agents may not comply with the norms that govern them because it is more in their interests not to comply. In the latter case, penalties are often introduced to try and coerce agents into compliance. We leave that discussion to one side, however, as it is tangential to the current line of enquiry.

Consider now the 'rooms' example in particular, and what happens if we assume that all agents act in accordance with their individual norms. It is clear that the effectiveness of the protocol (if in a red state, males move to the left when possible, females move to the right when possible) in guaranteeing that the system will eventually reach a green state, depends on the topography of rooms and connecting doors. Let us assume that there is a finite number of rooms, each room has at least one connecting room to its left or one to its right, and that there are no cycles in the configuration, in the sense that if an agent continues moving in the same direction it will never pass first out of, then back into, the same room. Under these circumstances, and removing the restriction on how many agents can pass through a door at the same time, it is intuitive that there is always a recovery, in the sense defined, from every red system state. Since all agents act in accordance with their norms, every male will move to the left (if it can), and every female will move to the right (if it can). If the resulting system state is not green, they will move again. Eventually, in the worst case, the males and females will be segregated in separate rooms, which is a green system state.

Of course, we cannot guarantee that having reached a green system state, the agents will not re-enter a red state: in this example, the individual agent-specific norms only dictate how agents should behave when the system is globally red. Once the system has recovered, the agents may mingle again. It is easy to imagine how we might use a model-checker to verify this and similar properties on coloured transition systems; we will not discuss the details in this paper.

### 4.2 Non-compliant Behaviours

One must be careful not to assume that if an agent $x$ fails to comply with its individual norms—if some transition $\tau$ is $red(x)$—then it must be that agent $x$ acted wilfully, perhaps to seek some competitive advantage, or carelessly; or if it is a simple reactive device, that its constructors failed to implement it

correctly. This may be so, but an agent may also fail to comply with its norms because of factors beyond its control, because it is prevented from complying by the actions of other agents, or by extraneous factors in the environment. To illustrate: suppose we return to the version of the 'rooms' example in which it is impossible for more than one agent to pass through the same doorway at the same time. All other features, including the specification of system norms and agent-specific norms, remain as before. Clearly the situation can now arise where several agents are required by their individual norms to pass through the same doorway; at most one of them can comply, and if one does comply, the others must fail to comply.

Again, in order to keep diagrams of the transition system small enough to be shown in full, we will consider just the case of two interconnecting rooms, and three agents, $m_1$, $m_2$, and $f_1$, of whom the first two are male and the last is female. Figure 2 shows the coloured agent-stranded transition system for this version of the example. We have adopted here the local-global coherence constraint (3) which is why some transitions that were globally green in the version of Sect. 3 are now globally red. Nothing essential in what follows depends on this. Transition labels are omitted from the diagram: since at most one agent can move at a time, they are obvious from looking at the states. Annotations on the arcs indicate the three agent-specific colourings for each transition; where arcs have no such annotation the transition is $\mathit{green}(x)$ for each of the three agents $x$. Omitted from the diagram are reflexive arcs from the green system states to themselves, representing transitions in which no agent moves. These transitions are all globally green, and therefore also (given local-global coherence) $\mathit{green}(x)$ for each agent $x$. The significance of the asterisks in some of the annotations will be explained presently. For now they may be read as indicating that the transition is $\mathit{red}(x)$ for the agent $x$.

One can see from the diagram that the system exhibits the following kinds of behaviour, among others.

(1) There are transitions coloured $\mathit{green}(x)$ for all three agents $x$ but which are nevertheless globally red. This is because, in this example, the agent-specific norms do not constrain agents' actions in green system states. Indeed, one can see from the diagram that in this example (though not in general) the globally red transitions which are $\mathit{green}(x)$ for all three agents $x$ are exactly those from a green system state to a red system state. The model $\mathcal{M}$ has the property:

$$\mathcal{M} \models \mathit{green}(m_1) \wedge \mathit{green}(m_2) \wedge \mathit{green}(f_1) \wedge \mathit{trans}{=}\mathit{red} \leftrightarrow 0{:}\mathit{status}{=}\mathit{green} \wedge 1{:}\mathit{status}{=}\mathit{red}$$

(2) There are globally green transitions from red system states to green system states (such as the one from state `s8` to state `s2` in which $m_2$ moves to the left and $m_1$ and $f_1$ stay where they are). These are transitions in which all three agents are able to comply with their individual norms. In this example, though not necessarily in other versions with more elaborate room configurations and

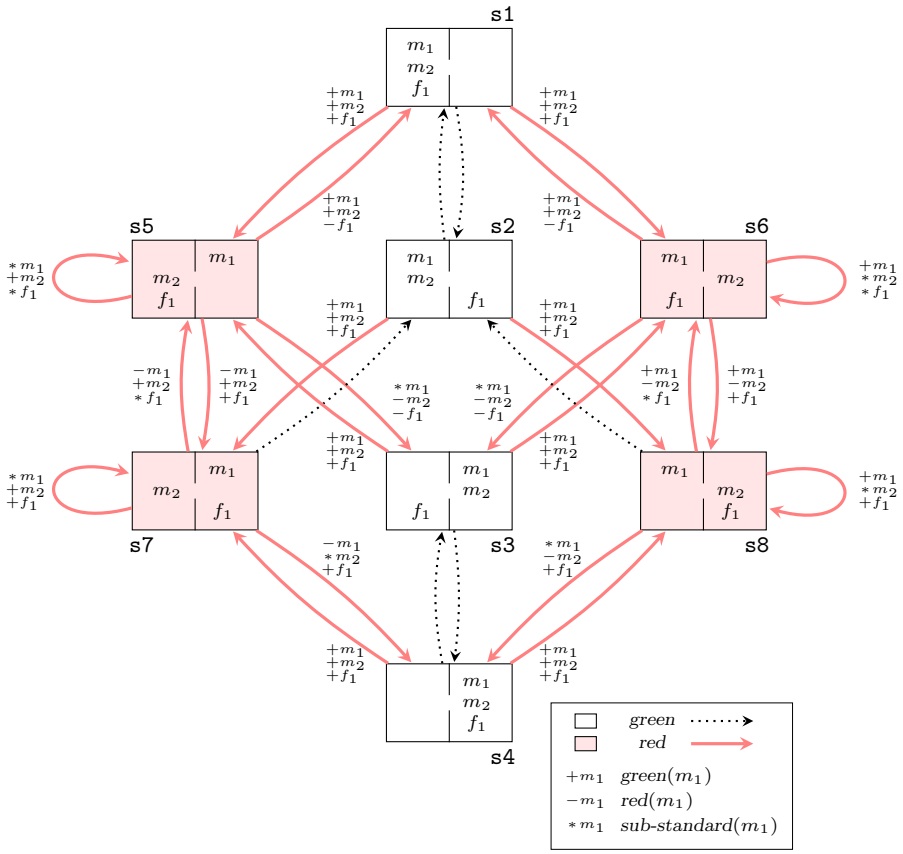


**Fig. 2.** Transitions without annotations are *green*($x$) for each of the three agents $x$. Reflexive arcs on green nodes, where no agent moves, are omitted from the diagram: they are all globally green, and *green*($x$) for each agent $x$. (The concept of a *sub-standard* strand is explained in Sect. 4.3.)

more agents, such transitions always recover from a red system state to a green system state. The system exhibits the property:

$$\mathcal{M} \models \ \mathit{green}(m_1) \wedge \mathit{green}(m_2) \wedge \mathit{green}(f_1) \wedge 0{:}\mathit{status}{=}\mathit{red} \rightarrow 1{:}\mathit{status}{=}\mathit{green}$$

(3) There are also globally red transitions in which at least one agent fails to comply with its individual norms but which lead from a red system state to a green system state (such as the one from state `s8` to state `s4` in which $m_1$ moves to the right and $m_2$ and $f_1$ stay where they are). These transitions recover from a red system state to a green system state but in violation of the individual agent-specific norms. These are transitions of type

$$\mathit{trans}{=}\mathit{red} \wedge (\mathit{red}(m_1) \vee \mathit{red}(m_2) \vee \mathit{red}(f_1)) \wedge 0{:}\mathit{status}{=}\mathit{red} \wedge 1{:}\mathit{status}{=}\mathit{green}$$

In fact, in the rooms example, though not in general, the system has the property:

$$\mathcal{M} \models \ \mathit{trans}{=}\mathit{red} \wedge 0{:}\mathit{status}{=}\mathit{red} \wedge 1{:}\mathit{status}{=}\mathit{green} \rightarrow (\mathit{red}(m_1) \vee \mathit{red}(m_2) \vee \mathit{red}(f_1))$$

(4) There are globally red transitions, such as the one from state `s6` to state `s3` in which $m_1$ moves to the right, and $f_1$ and $m_2$ stay where they are, in which no agent complies with its individual norms. These are transitions of type

$$\mathit{trans}{=}\mathit{red} \wedge \mathit{red}(m_1) \wedge \mathit{red}(m_2) \wedge \mathit{red}(f_1)$$

(5) And as the example is designed to demonstrate, there are globally red transitions where one agent complies with its individual norms but in doing so makes it impossible for one or both of the others to comply with theirs. For example, in the red system state `s6` where $m_1$ and $f_1$ are in the room on the left and $m_2$ is on the room on the right, there is no transition in which $m_2$ and $f_1$ can both comply with their individual norms: the following state formula is true at `s6`

$$\neg\langle \mathit{green}(m_2) \wedge \mathit{green}(f_1)\rangle\top \quad \text{equivalently} \quad [\mathit{green}(m_2) \wedge \mathit{green}(f_1)]\bot$$

$\mathit{green}(m_2) \wedge \mathit{green}(f_1)$ is not 'executable' in state `s6`.

In this version of the example, what are the possible system behaviours in the case where all agents do comply with their individual norms? Figure 3 shows the transition system that results if all $\mathit{red}(x)$ transitions are discarded, for all three agents $x$. The diagram confirms that when there is a constraint preventing more than one agent from moving through a doorway at a time, the system can enter a state from which there is no transition unless at least one agent fails to comply with its individual norms. In the diagram, these are the two red system states `s5` and `s6` where the female agent $f_1$ is in the left-hand room with a male. The ICCALC system provides facilities for undertaking this kind of analysis.

Now: one may think that there is a flaw in the way that the individual agent-specific norms in the example have been formulated, that their specification is wrong in that there are situations which make norm compliance impossible. A properly designed set of norms, it might be argued, must satisfy an 'ought implies can' principle; if it does not, it is flawed. That is not so. We are thinking here of a multi-agent system in which agents act independently, where there is no communication between agents, and where no agent can predict how other agents will act. If there were such communication it might be different, but suppose there is not. In these circumstances, it is quite impractical to try to anticipate every possible combination of behaviours by other agents, and in the environment, and to try to formulate agent-specific norms that make provision for each eventuality. It is quite impractical, even in examples as simple as this. It is realistic, however, to formulate agent-specific norms that will guide an individual agent's behaviour without reference to what other agents might do, and simply accept that there might be circumstances in which the agent-specific norms for $x$ conflict with those for $y$, and generally, that an agent may be prevented from complying with its individual agent-specific norms in some circumstances.

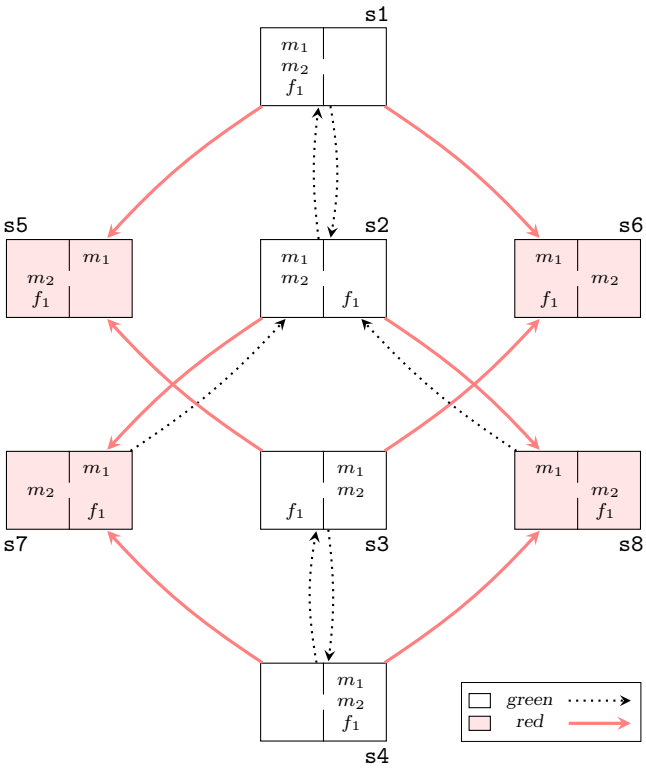


**Fig. 3.** System behaviour if all three agents comply with their individual norms. Reflexive arcs on green nodes are omitted from the diagram.

### 4.3 Sub-standard Behaviours

The example is designed to demonstrate several different categories of non-compliant agent behaviour. We pick out one for particular attention. Consider the state in which $m_1$ and $f_1$ are in the room on the left and $m_2$ is in the room on the right. (This is the red system state `s6` at the upper right of the diagram.) Because of the constraint on moving through the doorway, it is not possible for all three agents to comply with their individual norms. But suppose that each agent behaves in such a way that it will comply with its individual norms *in as much as it can*. A purely reactive agent, let us suppose, is programmed in such a way that it will attempt to act in accordance with its individual norms though it may not always succeed if something prevents it. A deliberative agent (human or computer) incorporates its individual norms in its decision-making procedures and takes them into account when planning its actions: it will always attempt to act in accordance with its individual norms though it may be unsuccessful. If all agents in the system behave in this way, then there are two

possible transitions from the red system state `s6`: either $f_1$ succeeds in moving to the right in accordance with its individual norms, or $m_2$ succeeds in moving to the left in accordance with its. The third possible transition from this system state, in which every agent stays where it is, can be ignored: it can only occur if no agent attempts to act in accordance with its individual norms, and this, we are supposing, is not how the agents behave. The exact mechanism which determines which of the two agents $m_2$ and $f_1$ is successful in getting through the doorway is not represented at the level of detail modelled here. At this level of detail, all we can say is that one or other of the agents $m_2$ and $f_1$ will pass through the doorway but we cannot say which.

Similarly, in the red system state `s8` at the lower right of the diagram, in which $m_1$ is on the left and $m_2$ and $f_1$ are on the right, we can ignore the transition in which $m_1$ moves to the right, if $m_1$'s behaviour is such that it always attempts to comply with its individual norms. The transition in which $f_1$ moves to the left can also be ignored, if $f_1$'s behaviour is to attempt to comply with its individual norms. And the transition in which $m_2$ stays where it is can be ignored, if $m_2$'s behaviour is to attempt to comply with its individual norms. This leaves just one possible transition, in which $m_2$ attempts to move to the left; this will succeed because the other two agents will not act in such a way as to prevent it. (We are tempted to refer to this kind of behaviour as behaviour in which every agent 'does the best that it can'. The term has too many unintended connotations, however, and so we avoid it.)

We are not suggesting, of course, that agents *always* behave in this way, only that there are circumstances where they do, or where it can be reasonably assumed that they do, or simply where we are interested in examining what system behaviours result if we suppose that they do.

We now make these ideas a little more precise. We will say that $x$'s behaviour in a particular transition $\tau$ from a state $s$ is *sub-standard*$(x)$ if the transition is *red*$(x)$ and, had $x$ acted differently in state $s$ while all other agents acted in the same way as they did in $\tau$, the transition from state $s$ could have been *green*$(x)$: $x$ could have acted differently in state $s$ and complied with its individual norms.

Alternatively, as another way of looking at it, we could say that a *red*$(x)$ transition $\tau$ from a state $s$ is *unavoidably-red*$(x)$ if every transition from state $s$ in which every agent other than $x$ acts in the same way as it does in $\tau$ is also *red*$(x)$: there is no *green*$(x)$ transition from state $s$ if every agent other than $x$ acts in the way it does in $\tau$. This is closer to the informal discussion above. One can see, informally for now, that every *red*$(x)$ transition is *sub-standard*$(x)$ if and only if it is not *unavoidably-red*$(x)$, and indeed, that every *red*$(x)$ transition is either *sub-standard*$(x)$ or *unavoidably-red*$(x)$, but not both.

Notice that these definitions allow for the possibility of actions in the environment. It is easy to imagine other versions of the example where an agent may be unable to act in accordance with its individual norms not because of the actions of other agents but because of extraneous factors in the environment. (Suppose, for instance, that an agent is unable to move to the room on the left while it is raining.) And here is a reason why we prefer not to treat 'the environment' as a

kind of agent: we do not want to be talking about *sub-standard* behaviours of the environment, or of agents preventing the environment from acting in accordance with its individual norms. In this respect at least, 'the environment' is a very different kind of agent from the others.

It still remains to formalise these definitions. For this we need to be able to refer to actions by individual agents in transitions, which is not part of the LTS structure as we have it. Indeed, there is no explicit concept of an individual agent in the semantics at all. We defer further discussion until the next section. For now, we rely on the informal account just given.

The diagram of the transition system for this example was shown earlier in Fig. 2. The figure shows the sub-standard transitions for each agent. They are those in which the transition annotations are marked with an asterisk. For example, in the red system state `s6` at the upper right of the diagram, where $m_1$ and $f_1$ are on the left and $m_2$ is on the right, the transition in which all three agents stay where they are is *sub-standard*$(m_2)$, because there is a *green*$(m_2)$ transition from this state in which $m_1$ and $f_1$ act in the same way and $m_2$ acts differently, namely the transition in which $m_1$ and $f_1$ stay where they are and $m_2$ moves to the left in accordance with its individual norms. Similarly, the transition from state `s6` in which $m_1$ moves to the right and $m_2$ and $f_1$ stay where they are is *sub-standard*$(m_1)$ because the transition where all three agents stay where they are is *green*$(m_1)$. And likewise for the other transitions marked as sub-standard in the diagram. The *red*$(x)$ transitions not marked as *sub-standard*$(x)$ are *unavoidably-red*$(x)$.

Suppose we wish to examine what system behaviours result if all three agents comply, in as much as they can, with their individual norms. Suppose, in other words, that we disregard those transitions which are sub-standard for any of the three agents $x$. The ICCALC implementation supports this kind of analysis. The result is shown in Fig. 4. There are still *red* transitions in the diagram. Some, such as the one from `s4` to `s8`, are *green*$(x)$ for every $x$ but are nevertheless globally red. Those, such as the one from `s6` to `s8`, which are *red*$(x)$ for some agent $x$ are *unavoidably-red*$(x)$.

Many other variations of the example could be examined in similar fashion. If female agents are more reliable than male agents, for instance, we might be interested in examining what system behaviours result when there is never sub-standard behaviour by females though possible sub-standard behaviour by males.

Interestingly, when analysing the example using ICCALC, it turned out that if we assume there is no sub-standard behaviour by either of the two male agents $m_1$ and $m_2$, that is, if we assume that $m_1$ and $m_2$ always comply with their individual norms if they can, then there is no sub-standard behaviour by the female agent $f_1$ either. This is really an artefact of the simplicity of the example where there are just two rooms and very strong constraints on how the three agents can move between them. Nevertheless, it does demonstrate the possibility, in principle at least, that agents can sometimes be coerced into compliance by the behaviours of others, without resort to sanctions and other enforcement mechanisms.

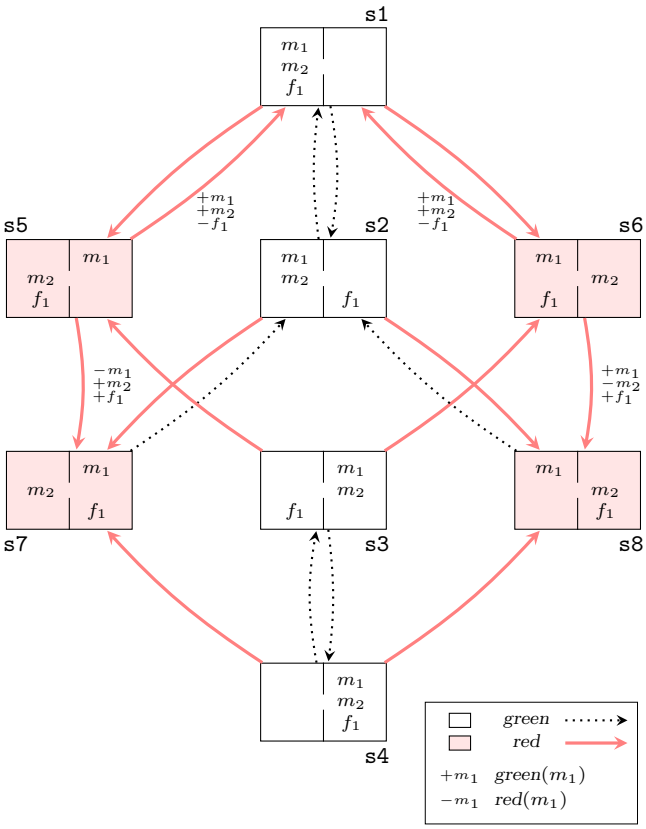


**Fig. 4.** System behaviour if all three agents comply with their individual norms, in as much as they can. Transitions without annotations are coloured *green*($x$) for each agent $x$. Reflexive arcs on green nodes are omitted from the diagram.

As a final remark, notice that what is *sub-standard*($x$) or *unavoidably-red*($x$) for an agent $x$ can depend on *normative* as well as *physical* constraints. Suppose (just for the sake of an example) that there is another individual norm for $m_1$ to the effect that it should never stay in a particular room (say, the room on the left) but should move out immediately if it enters it: a transition in which $m_1$ stays in the room on the left is *red*($m_1$), in every system state, red or green. With this additional constraint, some of the transitions that were globally green are now globally red because of the local-global coherence constraint (assuming we choose to adopt it). But further, the transition from the red system state `s6` at the upper right of the diagram in Fig. 2, in which $m_1$ moves to the right and $m_2$ and $f_1$ stay in the room on the right, was previously *sub-standard*($m_1$). It is no longer *sub-standard*($m_1$): there is now no *green*($m_1$) transition from this state when $m_2$ and $f_1$ stay where they are.

Clearly, in this example, if $m_1$ is in the room on the left in a red system state, it has conflicting individual norms: one requiring it to move to the right, and one requiring it to stay where it is. It cannot comply with both, so neither action is *sub-standard*$(m_1)$; both are *unavoidably-red*$(m_1)$.

How $m_1$ should resolve this conflict is an interesting question but not one that we intend to consider here. It is also a question that only has relevance when $m_1$ is a deliberative agent which must reason about what to do. If $m_1$ is a purely reactive device, then its behaviour in this case could perhaps be predicted by examining its program code. Both of these possibilities are beyond the level of detail of agent and system behaviours modelled in this paper. In the simplest case we could eliminate the conflict by simply specifying that one norm takes precedence over the other and adjusting the definition of *red*$(x)$ and *green*$(x)$ accordingly. Discussion of other possible mechanisms is beyond the scope of this paper.

Notice that, unlike the situation referred to earlier, where there was a conflict between agent-specific norms for two *different* agents, here we have a conflict between agent-specific norms for the *same* agent. It would be reasonable to say that there should be no conflicts of this type in any well-defined set of agent-specific norms.

There is thus a special category of *unavoidably-red*$(x)$ transitions in which *every* action performed by $x$ is *red*$(x)$.

- A *red*$(x)$ transition $\tau$ in $R$ is *degenerately-red*$(x)$ iff for every transition $\tau' \in R$ such that *prev*$(\tau) =$ *prev*$(\tau')$ we have $\mathcal{M}, \tau' \models$ *red*$(x)$.

Clearly

$$\textit{degenerately-red}(x) \subseteq \textit{unavoidably-red}(x)$$

When a transition $\tau$ is *degenerately-red*$(x)$ then its initial state $s =$ *prev*$(\tau)$ is such that there is no action that can be performed by $x$ in compliance with its individual norms. We call such a state a *red*$(x)$*-sink*:

- state $s$ is a *red*$(x)$*-sink* iff for every transition $\tau \in R$ such that *prev*$(\tau) = s$ we have $\mathcal{M}, \tau \models$ *red*$(x)$.

In formulas, a state $s$ in a model $\mathcal{M}$ is a *red*$(x)$-sink when:

$$\mathcal{M}, s \models \neg\langle \textit{green}(x)\rangle \top \quad \text{equivalently} \quad \mathcal{M}, s \models [\textit{green}(x)]\bot$$

*green*$(x)$ is not 'executable' in a *red*$(x)$-sink. Intuitively, $s$ is a *red*$(x)$-sink iff every transition $\tau$ from $s$ is *degenerately-red*$(x)$. A well-designed set of agent-specific norms should contain no *red*$(x)$-sinks. We can test for the presence of such states but we will not assume that they cannot occur. (Notice that a *red*$(x)$-sink is not necessarily a *red*$(y)$-sink for all other agents $y$.) There are no *red*$(x)$-sinks and no *degenerately-red*$(x)$ transitions in the 'rooms' example we have been discussing. (Though there are, as we have observed, states in which there are no transitions of type $\textit{green}(m_1) \wedge \textit{green}(m_2) \wedge \textit{green}(f_1)$.)

Similarly, we can say that a system state $s$ is a (global) *red-sink* if there is no transition from $s$ that is globally green. A state $s$ is thus a red-sink when

$$\mathcal{M}, s \models \neg\langle \mathit{trans}{=}\mathit{green}\rangle\top \quad \text{equivalently} \quad \mathcal{M}, s \models [\mathit{trans}{=}\mathit{green}]\bot$$

*trans*=*green* is not 'executable' in a red-sink state.

One might think that any well designed set of system norms will have no red-sinks. That is not so. The local-global coherence constraint (if it is adopted) means that every $\mathit{red}(x)$-sink is also a red-sink. But even if there are no $\mathit{red}(x)$-sinks there can still be global red-sinks—that is one of the points we are making with the rooms example. Red-sink states may be undesirable/unwanted but we do not want to insist that they cannot occur. They can occur even in a well-designed set of agent-specific and system norms.

As an aside, note that a red-sink state can be (globally) green: a green state from which all transitions are red (or from which there are no transitions at all) is a red-sink state. We have considered extending the 'green-green-green' constraint: we could say that any transition to a (global) red-sink is undesirable/unwanted and should therefore be (globally) red. That seems natural and straightforward but its implications remain for future investigation and are not built-in to the framework as we have it now.

## 5 Agent-stranded Transition Systems

Although the transition systems as they currently stand allow us to colour transitions $\mathit{green}(x)$ and $\mathit{red}(x)$, we are only able to give informal definitions of concepts such as $\mathit{sub\text{-}standard}(x)$ and $\mathit{unavoidably\text{-}red}(x)$. This is because there is no way of referring to an individual agent's actions in a transition. There is no explicit concept of an individual agent in the semantics at all. We would like to be able to extract from, or represent in, a transition system that it is specifically one agent's actions that are responsible for, or the cause of, a transition's having a certain property $\varphi$.

Let $Ag$ be a (finite) set of agent names. An *agent-stranded LTS* is a structure

$$\langle S, A, R, Ag, \mathit{strand}\rangle$$

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

The new component is *strand*, which is a function on $Ag \times A$. $\mathit{strand}(x, \varepsilon)$ picks out from a transition label/event $\varepsilon$ the component or 'strand' that corresponds to agent $x$'s contribution to the event $\varepsilon$. We will write $\varepsilon_x$ for $\mathit{strand}(x, \varepsilon)$. 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 $A_{\mathrm{env}}$ 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

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

However, it is not necessary to restrict attention to transition labels/events 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 $\varepsilon$ of $A$. As usual, $\varepsilon_x$ may represent several concurrent actions by $x$, or actions with non-deterministic effects (by which we mean that there could be transitions $\tau$ and $\tau'$ with $prev(\tau) = prev(\tau')$, $\varepsilon_x = \varepsilon'_x$ where $\varepsilon$ and $\varepsilon'$ are the labels of $\tau$ and $\tau'$ respectively, and $post(\tau) \neq post(\tau')$).

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

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

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

We do not, at this stage, introduce more granularity into the structure of states or consider norms which regulate the (local) state of an individual agent. These are possible developments for further work. Our interest here is to study the norm-governed *behaviour* of agents, and how this may be related to the norms pertaining to the system as a whole. To that end, we will concentrate on the transitions which are used to represent agents' actions.

We are now able to formalise the sub-standard and unavoidably-red categories of non-compliant behaviours, amongst other things. But first we turn to a fundamental feature of agent-specific norms that we were unable to discuss previously.

We assume as before that there is a constant *status* in $\sigma^{\text{f}}$ for colouring states (globally) *red* or *green*, an event constant *trans* in $\sigma^{\text{a}}$ for colouring transitions (globally) *red* or *green*, and (Boolean) event constants $green(x)$ and $red(x)$ in $\sigma^{\text{a}}$ for each agent $x$ in $Ag$, with $red(x)$ as an abbreviation for $\neg green(x)$.

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 $\mathcal{M} \models red(x) \rightarrow red$ is optional. However, we do impose the following constraint on agent-specific colourings: if $\tau$ is a $green(x)$ (resp., $red(x)$) transition from a state $s$ in model $\mathcal{M}$, then every transition $\tau'$ from state $s$ in which agent $x$ behaves in the same way as it does in $\tau$ must also be $green(x)$ (resp., $red(x)$). In other words, for all transitions $\tau$ and $\tau'$ in a model $\mathcal{M}$, and all agents $x \in Ag$:

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

(And hence also $\mathcal{M}, \tau \models red(x)$ iff $\mathcal{M}, \tau' \models red(x)$ whenever $prev(\tau) = prev(\tau')$ 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 $\tau$ and $\tau'$ 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 $x$.

To take a simple example: suppose that when $x$ fires a loaded gun at $y$, the action may result in the killing of $y$, or the shot may miss, or the gun may misfire, and $y$ survives: the shooting action is non-deterministic. We may take the view, as system designers, that a shooting transition is *red* if it results in the killing of $y$, and *green* if it does not. However, since $x$'s action is the same whether the shooting is fatal or not, an agent-specific norm for $x$ must either make both transitions $\mathit{green}(x)$ or both $\mathit{red}(x)$.

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 explicitly. These features can be added but raise more questions than we have space for here. We leave that refinement for another occasion. For now, we insist on the 'absence of moral luck' constraint (4) as a minimal requirement for agent-specific norms.

### Sub-standard behaviours

We are now able to formalise the notion of sub-standard and unavoidably-red behaviours of agent $x$.

*Definition* Let $\mathcal{M} = \langle S, A, R, Ag, \mathit{strand}, h^{\mathrm{f}}, h^{\mathrm{a}} \rangle$ be an agent-stranded model, with event constants $\mathit{green}(x)$ and $\mathit{red}(x)$ representing the agent-specific norms for every $x$ in $Ag$.

Let *unavoidably-red* and *sub-standard* be functions from the set of agents $Ag$ to $\wp(R)$. For every agent $x \in Ag$ and every transition $\tau \in R$:

- $\tau \in \mathit{unavoidably\text{-}red}(x)$ iff $\mathcal{M}, \tau \models \mathit{red}(x)$ and, for every transition $\tau' \in R$ such that $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\tau_y = \tau'_y$ for every agent $y \in Ag \setminus \{x\}$, we have $\mathcal{M}, \tau' \models \mathit{red}(x)$.
- $\tau \in \mathit{sub\text{-}standard}(x)$ iff $\mathcal{M}, \tau \models \mathit{red}(x)$ and there exists $\tau' \in R$ such that $\mathit{prev}(\tau) = \mathit{prev}(\tau')$ and $\tau_x \neq \tau'_x$ and $\tau_y = \tau'_y$ for every agent $y \in Ag \setminus \{x\}$ and $\mathcal{M}, \tau' \models \mathit{green}(x)$.

Notice that the definition of *sub-standard*$(x)$ makes reference to agent $x$ acting differently in the transitions $\tau$ and $\tau'$. If we assume the 'absence of moral luck' property (4)—as we do—then the definition can be simplified. If $\mathcal{M}, \tau \models$ *red*$(x)$ and $\mathcal{M}, \tau' \models$ *green*$(x)$ for a transition $\tau'$ from the same initial state as $\tau$ (*prev*$(\tau)$ = *prev*$(\tau')$) then the condition $\tau_x \neq \tau'_x$ is implied: if $\tau_x = \tau'_x$ then the 'absence of moral luck' constraint would be violated. The following simpler definition is equivalent to the original:

- $\tau \in$ *sub-standard*$(x)$ iff $\mathcal{M}, \tau \models$ *red*$(x)$ and there exists $\tau' \in R$ such that *prev*$(\tau)$ = *prev*$(\tau')$ and $\tau_y = \tau'_y$ for every agent $y \in Ag \setminus \{x\}$ and $\mathcal{M}, \tau' \models$ *green*$(x)$.

We will use this simpler definition of *sub-standard*$(x)$ from now on. Agent-specific colourings must satisfy the 'absence of moral luck' property; without it the notion of *sub-standard*$(x)$ is not meaningful.

One can see from the definitions that, as indicated informally earlier, every *red*$(x)$ transition is *sub-standard*$(x)$ if and only if it is not *unavoidably-red*$(x)$, and that every *red*$(x)$ transition is either *sub-standard*$(x)$ or *unavoidably-red*$(x)$, but not both. In other words

$$\textit{sub-standard}(x) = \textit{red}(x) \setminus \textit{unavoidably-red}(x)$$

Recall that there is a special sub-category of *degenerately-red*$(x)$ transitions in which *every* action performed by $x$ is *red*$(x)$. Since *degenerately-red*$(x) \subseteq$ *unavoidably-red*$(x)$, the *red*$(x)$ transitions can be partitioned into three disjoint sub-types:

- *sub-standard*$(x)$
- *degenerately-red*$(x)$
- *unavoidably-red*$(x) \setminus$ *degenerately-red*$(x)$

We do not give a name to this third category: a well-formed set of agent-specific norms will have no *degenerately-red*$(x)$ transitions, and then it is only the distinction between *sub-standard*$(x)$ and *unavoidably-red*$(x)$ that matters.

There are a number of other questions that we might now consider. For instance:

- Are there any other categories of non-compliant behaviour that could usefully be identified?
- Is it meaningful to talk about *sub-standard*$(x)$ behaviour of an agent $y$ other than $x$? What could this mean?
- If a transition is (globally) red, can we determine which of the agents, if any, is responsible for that transition's being (globally) red? In the 'rooms' example, if agent $m_2$ exits a room and leaves $m_1$ and $f_1$ alone together, can we determine which of the agents, if any, violated the system norms?
- If a transition is *unavoidably-red*$(x)$ (but not *degenerately-red*$(x)$) is it possible to identify the subset of agents $Ag$ whose actions prevent $x$ from complying with its agent-specific norms?

The last question concerns forms of *collective* action/agency that will not be addressed in this paper. They are investigated elsewhere [3]. The first three questions are answered below. However the present notation is too cumbersome. We now extend the language so these and other properties can be expressed as *formulas*.

## 6 A Modal Language for Agency in Transitions

In this section we introduce a modal language for talking about the agent-specific components of transitions (their 'strands'). We extend the transition formulas of Sect. 2 with a (unary) operator $[_{\text{alt}}]$, and (unary) operators $[x]$ and $[\backslash x]$ for every agent $x \in Ag$. This will allow us to express concepts such as *sub-standard*$(x)$ and *unavoidably-red*$(x)$, and others, as formulas. In Sect. 6.2 we will introduce two 'brings it about' modalities.

### 6.1 A Logic of Agent Strands

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

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

$\langle_{\text{alt}}\rangle$ is the dual of $[_{\text{alt}}]$.

$[_{\text{alt}}]\varphi$ is satisfied by, or 'true at', a transition $\tau$ when all alternative transitions *from the same initial state* as $\tau$ satisfy $\varphi$. $\langle_{\text{alt}}\rangle\varphi$ is true at a transition $\tau$ if there exists an alternative transition from the same initial state as $\tau$ of type $\varphi$.

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

$$[_{\text{alt}}]\varphi \rightarrow \varphi$$
$$[_{\text{alt}}]\varphi \rightarrow [_{\text{alt}}][_{\text{alt}}]\varphi$$
$$\neg[_{\text{alt}}]\varphi \rightarrow [_{\text{alt}}]\neg[_{\text{alt}}]\varphi$$

Clearly the following is valid

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

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

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

$\langle x \rangle$ 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:

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

We will say, for short, that when $[x]\varphi$ is true at a transition $\tau$, $\varphi$ is necessary for how $x$ acts in $\tau$; and when $[\backslash x]\varphi$ is true at $\tau$, that $\varphi$ is necessary for how the agents $Ag \setminus \{x\}$ collectively act in $\tau$. (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.) $\langle x\rangle\varphi$ is true at a transition $\tau$ if there is a transition $\tau'$ of type $\varphi$ from the same initial state as $\tau$ in which $x$ acts in the same way as it does in $\tau$. Clearly $\varphi \rightarrow \langle x\rangle\varphi$ is valid. $\varphi \wedge \langle x\rangle\neg\varphi$ is true at a transition $\tau$ if $\tau$ is of type $\varphi$ but there is an alternative transition of type $\neg\varphi$ from the same initial state as $\tau$ in which $x$ acts in the same way as it does in $\tau$. $\varphi \wedge \langle x\rangle\neg\varphi$ is equivalent to $\varphi \wedge \neg[x]\varphi$. And similarly, $\varphi \wedge \langle\backslash x\rangle\neg\varphi$ is true at a transition $\tau$ if $\tau$ is of type $\varphi$ and there is an alternative transition of type $\neg\varphi$ from the same initial state as $\tau$ in which every other agent besides $x$ acts in the same way as it does in $\tau$.

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

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

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

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

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

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

The other direction is *not* valid:

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

This is important. Here is a simple example. Consider the (green) state in the 'rooms' example in which all three agents are on the left (this is the state `s1` in the diagrams), and the transition $\tau$ from that state in which the female $f_1$ moves to the right. The resulting state is also green, and so the transition $\tau$ is (globally) green too (*trans=green* is true at $\tau$). Clearly in all transitions from `s1` in which $f_1$ moves (there is only one), *trans=green* is true, and so $[f_1]$*trans=green* is true at $\tau$. $[\backslash f_1]$*trans=green* is also true at $\tau$. There are two transitions from `s1` in which $m_1$ and $m_2$ both act as they do in $\tau$: $\tau$ itself, and the transition in which

$m_1$ and $m_2$ stay where they are and so does $f_1$. Both of these transitions have *trans=green* true.

But consider $[m_1]$*trans=green*. There are three transitions from state `s1` in which $m_1$ acts as it does in $\tau$: $\tau$ itself, the transition in which $f_1$ moves to the right and $m_1$ and $m_2$ stay where they are, and the transition in which $m_2$ moves to the right and $m_1$ and $f_1$ stay where they are. The last of these is a transition from a green system state to a red system state and so is of type *trans=red*. So $[m_1]$*trans=green* is false at $\tau$. By exactly the same argument $[m_2]$*trans=green* is false at $\tau$ as well. So here we have an example where $[\backslash f_1]$*trans=green* is true but neither $[m_1]$*trans=green* nor $[m_2]$*trans=green* is true. In general $[\backslash x]\varphi$ is true at a transition $\tau$ because $\varphi$ is necessary for how the agents $Ag \setminus \{x\}$ *collectively* act in $\tau$, but that is not the same as saying that $[y]\varphi$ is true at $\tau$ for some individual agent $y \in Ag \setminus \{x\}$.

For the special case where there are exactly two agents in $Ag$, $Ag = \{x, y\}$, the following is valid

$$[\backslash x]\varphi \leftrightarrow [y]\varphi \qquad (Ag = \{x, y\})$$

But that is merely a special case. 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 $[\backslash x]\varphi \rightarrow [x]\varphi$.

The language can be generalised to allow expressions $[G]\varphi$ for any $G \subseteq Ag$. $[x]\varphi$ is then shorthand for $[\{x\}]\varphi$, $[\backslash x]\varphi$ is shorthand for $[Ag \backslash \{x\}]\varphi$, and $[\text{alt}]\varphi$ is shorthand for $[\emptyset]\varphi$. The generalisation actually simplifies the technical development but since we are not discussing technical details in this paper we will not use the generalised form $[G]\varphi$ in what follows. We will note only that the logic of these operators is very familiar: the logic of $[G]\varphi$ is exactly that of 'distributed knowledge' (of type $S5$) of a group of agents $G$. (See e.g. [18].) Soundness, completeness, and complexity results are immediately available. We leave further discussion of technical properties to one side. See [3] for details. Our aim in this paper is to illustrate the expressiveness and uses of the language.

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

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

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

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

It is *degenerately-red*$(x)$ when

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

and hence *unavoidably-red*$(x)$ but not *degenerately-red*$(x)$ when

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

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? Expressing the definition given earlier as a formula, transition $\tau$ in a model $\mathcal{M}$ is *sub-standard*$(x)$ when

$$\mathcal{M}, \tau \models red(x) \wedge \langle \backslash x \rangle\, green(x)$$

that is, equivalently, when:

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

Consider now the 'absence of moral luck' constraint in a model $\mathcal{M}$, that is, the validity $\mathcal{M} \models red(x) \rightarrow [x]\, red(x)$. Agent-specific colourings must have this property as the minimal requirement for agent-specific norms of the type we are discussing. With this constraint we have $\mathcal{M} \models red(x) \leftrightarrow [x]\, red(x)$, and this in turn means that a transition $\tau$ in a model $\mathcal{M}$ is *sub-standard*$(x)$ when

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

Implicit 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 introducing two new defined operators for expressing *agency* of an agent $x$ in bringing it about that a transition is of a particular type.

$\mathrm{E}_x$ and $\mathrm{E}^+_x$ are defined operators:

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

Both may be read as expressing a sense in which $x$ brings it about that (a transition is of type) $\varphi$. We will explain the difference between them below. Essentially, $\mathrm{E}^+_x$ takes into account possible actions by other agents whereas $\mathrm{E}_x$ does not but treats them merely as part of the environment in which $x$ acts.

With the 'absence of moral luck' constraint, a transition $\tau$ in a model $\mathcal{M}$ is *sub-standard*$(x)$ when

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

So, a transition is *sub-standard*$(x)$ when $x$ brings it about that, or is responsible for, the transition's being *red*$(x)$.

The notation $\mathrm{E}_x \varphi$ is chosen because its definition bears a very strong resemblance to Ingmar Pörn's [2] 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 $\varphi$. Pörn's logic does not have the analogue of $\mathrm{E}^+_x\varphi$. There are nevertheless some striking similarities, but also some very significant technical differences. See [3] for further discussion.

What about $\mathrm{E}_x\,red(x)$? What kind of non-compliant behaviour does that express? $\mathrm{E}_x\,red(x)$ is $[x]red(x) \wedge \neg[\text{alt}]red(x)$. Assuming the 'absence of moral luck' property for $red(x)$, which we do, this is equivalent to $red(x) \wedge \neg[\text{alt}]red(x)$, which is just $red(x)$ 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 $\mathrm{E}^+_x(trans{=}red)$ and $\mathrm{E}_x(trans{=}red)$ which express that an agent $x$ brings it about, or is responsible for, a transition's being (globally) red. These are not representations of agent-specific norms. Although $\mathrm{E}^+_x(trans{=}red)$ and $\mathrm{E}_x(trans{=}red)$ both satisfy the required 'absence of moral luck' property—both of the following are valid in any model $\mathcal{M}$:

$$\mathrm{E}^+_x(trans{=}red) \rightarrow [x]\mathrm{E}^+_x(trans{=}red)$$
$$\mathrm{E}_x(trans{=}red) \rightarrow [x]\mathrm{E}_x(trans{=}red)$$

we are regarding this property as the *minimal* requirement for agent-specific norms; the other requirements, concerned with what an agent can actually sense/perceive and what actions it can actually perform, are not modelled at the level of detail we have in the present framework. The point is that $\mathrm{E}^+_x(trans{=}red)$ and $\mathrm{E}_x(trans{=}red)$ are unlikely to satisfy these other requirements, since both are expressed in terms of a global transition property ($trans{=}red$) and this is not something that an individual agent is likely to be able to sense/perceive. On the other hand, $\mathrm{E}^+_x(trans{=}red)$ and $\mathrm{E}_x(trans{=}red)$ both express properties that might be of interest from the system designer's point of view. We will see other examples of similar properties when we look at some examples later.

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 $\mathrm{E}_y\mathrm{E}^+_x\,red(x)$ or $\mathrm{E}^+_y\mathrm{E}^+_x\,red(x)$ for agents $x \neq y$. Certainly the simpler expressions $\mathrm{E}^+_y\,red(x)$ and $\mathrm{E}_y\,red(x)$ 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 $\mathrm{E}_y\mathrm{E}^+_x\,red(x)$ and $\mathrm{E}^+_y\mathrm{E}^+_x\,red(x)$ are not satisfiable in any model $\mathcal{M}$; both of the following are valid

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

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

### 6.2 A Logic of 'Brings it about'

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

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

The study of logics of this type has a very long tradition. In computer science the best known examples are perhaps the 'stit' ('seeing to it that') family (see e.g. [19–21]). Segerberg [22] provides a summary of early work in this area, and Hilpinen [23] 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.

We will not present a full account of the logical properties of the agency operators $\mathrm{E}_x$ and $\mathrm{E}^+_x$ here. They 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.

We will simply remark that the definitions of $\mathrm{E}_x$ and $\mathrm{E}^+_x$ have two ingredients typical of logics of agency. The first conjunct is a 'necessity condition': $\mathcal{M}, \tau \models [x]\varphi$ says that all transitions from $\mathit{prev}(\tau)$ in which $x$ acts in the same way as it does in $\tau$ are of type $\varphi$, or as we also say, that $\varphi$ is necessary for how $x$ acts in $\tau$. The other component is used to capture the concept of *agency* itself—the fundamental idea that $\varphi$ is, in some sense, caused by or is the result of actions by $x$. Most accounts of agency introduce a negative 'counteraction' or counterfactual 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 second conjunct in the definition of $\mathrm{E}_x$ adds the 'counteraction' requirement: had $x$ acted differently, then the transition might have been different. 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 $\tau$ only if $x$ acts differently than in $\tau$. Thus, $\mathrm{E}_x\varphi$ is true at a transition $\tau$ if and only if $\varphi$ is necessary for how $x$ acts in $\tau$, and had $x$ acted differently than in $\tau$ then the transition from $\mathit{prev}(\tau)$ might have been different (i.e., of type $\neg\varphi$). For $\mathrm{E}^+_x$, the counteraction condition is stronger: had $x$ acted differently than in $\tau$ then the transition from $\mathit{prev}(\tau)$ might have been different, *even if all other agents*, besides $x$, had acted in the same way as they did in $\tau$.

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) $\varphi$. Clearly the following is valid:

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

What is the difference? It is easy to check that, because $[y]\varphi \to [\backslash x]\varphi$ is valid for any $x \neq y$, the following is valid

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

and hence also:

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

So $\mathrm{E}^+_x\varphi$ expresses that it is $x$, and $x$ *alone*, who brings it about that $\varphi$. In contrast, $\mathrm{E}_x\varphi$ leaves open the possibility that some other agent $y \neq x$ also brings it about that $\varphi$: 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. The 'rooms' example has several instances, as will be demonstrated in Sect. 7. Notice that the conjunction $\mathrm{E}_x\varphi \wedge \mathrm{E}_y\varphi$ is equivalent to

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

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. The logic of (unwitting) collective action/agency is investigated in [3]. We do not have space to summarise that here.

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. In the present framework, $\mathrm{E}_x\mathrm{E}_y\varphi$ 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:

$$\mathrm{E}_x\mathrm{E}_y\varphi \to \mathrm{E}_x\varphi \tag{5}$$

is valid for $\mathrm{E}_x$. Informally, it says that if $x$ acts in such a way that it unwittingly brings it about that $y$ unwittingly brings it about that $\varphi$, then $x$ also unwittingly brings it about that $\varphi$. What of $\mathrm{E}^+_x$ and $\mathrm{E}^+_y$ 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 $\mathrm{E}^+_x \mathrm{E}^+_y \varphi \rightarrow \mathrm{E}^+_x \varphi$.

Clearly $\mathrm{E}_x$ and $\mathrm{E}^+_x$ express a notion of *successful* action: if agent $x$ brings it about that (a transition is of type) $\varphi$ then it is indeed the case that $\varphi$. Or to put it another way (paraphrasing Hilpinen [23] quoting Chellas [24]): $x$ can be held responsible for its being the case that $\varphi$ only if it is the case that $\varphi$. $\mathrm{E}_x$ and $\mathrm{E}^+_x$ are both 'success' operators: both of the following schemes are valid:

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

Sergot [3] examines other properties of these 'brings it about' operators and provides a sound and complete axiomatisation of the logic. Further details can be found there. They are not essential for the purposes of this paper.

### 6.3 Example: 'The others made me do it'

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

One possibility:

$$[x]\varphi \wedge [\backslash x]\varphi \wedge \neg [_{\mathrm{alt}}]\varphi \tag{6}$$

This might be read as '$x$ did $\varphi$, but the others $Ag \setminus \{x\}$ between them acted in such a way as to make $\varphi$ unavoidable'. It can be checked that (6) is equivalent to

$$\mathrm{E}_x \varphi \wedge \neg \mathrm{E}^+_x \varphi \tag{7}$$

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

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

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

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 is $\varphi$. 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' $\varphi$ 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 (8) is false. (8) can be expressed equivalently as

$$[\backslash x]\mathrm{E}_x\varphi \tag{9}$$

Moreover, the following is valid:

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

In other words, 'the others made me do it' (8)–(9) implies 'the others made me do it' (6)–(7), 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:

$$\bigvee_{y \in Ag\setminus\{x\}} \mathrm{E}_y\mathrm{E}_x\varphi \tag{10}$$

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 (10) implies, but is not implied by, (9).

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 (10) implies the second (8)–(9) which implies the first (6)–(7).

### 6.4 Bringing about and Sustaining

$\mathrm{E}_x\varphi$ and $\mathrm{E}^+_x\varphi$ represent that $x$ brings it about that a transition is of type $\varphi$. 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 $F$. 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. This sense of

'brings it about' can be expressed as a *state formula*. We omit the details. It is transitions that are of primary interest in this paper.

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, $[\mathrm{alt}]\varphi \wedge \mathrm{E}_x \varphi' \to \mathrm{E}_x(\varphi \wedge \varphi')$ is also valid (and similarly for $\mathrm{E}^+_x$), 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, some suggestions by Segerberg [22] and Hilpinen [23] following an idea of von Wright [25, 26]. We will follow the terminology of Hilpinen's version; the others are essentially the same. Hilpinen sketches an account with two components: first, that actions are associated with transitions between states; and second, to provide the counterfactual 'counteraction' condition required to capture the notion of agency, a distinction 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

$$\begin{aligned} \mathrm{E}_x(0{:}\neg F \wedge 1{:}F) \quad &\text{or} \quad \mathrm{E}^+_x(0{:}\neg F \wedge 1{:}F) \\ \mathrm{E}_x(0{:}F \wedge 1{:}F) \quad &\text{or} \quad \mathrm{E}^+_x(0{:}F \wedge 1{:}F) \end{aligned}$$

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:

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

(Or as above, but with $\mathrm{E}^+_x$ in place of $\mathrm{E}_x$.)

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:

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

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 $\mathrm{E}^+_x$).

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 $\mathrm{E}^+_x$ in place of $\mathrm{E}_x$ if we wish to express that $x$ is the sole agent responsible in each case.

## 7 Example (Rooms, contd)

This section illustrates how the formal language presented in the paper may be applied to the analysis of the 'rooms' example. It presents a transcript of the outputs from the ICCALC system. These transcripts are produced by specifying a list of formulas expressing properties of interest. ICCALC evaluates these formulas on all transitions in the example. It is also possible to specify formulas to be evaluated on states. We show only a small extract of state annotations here to keep the transcripts manageable.

We have also modified the example slightly. In the version discussed here, the agent-specific norms apply only to those male agents and female agents who are in a room alone together, and not, as in the previous version, to male agents and female agents in other rooms as well. So concretely: in this version, whenever a male agent $x$ and a female agent $y$ are alone in a room together, a transition from that state is *green*($x$) if the male agent $x$ moves to the left, if there is a room to the left, *green*($x$) if it does not move when there is no room to the left, and *red*($x$) otherwise; it is *green*($y$) for the female agent $y$ with 'left' replaced by 'right'. There could be several such rooms in a system state (though not in the simple example where there are just two rooms and three agents); the agent-specific norms apply to all such male-female pairs. All other transitions are *green*($x$) for all agents $x$. All other features of the example are exactly as before. The system norms colour any state where there is a male agent and female agent alone in a room (globally) red (*status=red*); all other system states are (globally) green (*status=green*). Transitions are coloured (globally) red (*trans=red*) by the *ggg* constraint and by the local-global coherence constraint that every *red*($x$) transition is also globally red; all other transitions are globally green (*trans=green*). We also have the physical constraint that no more than one agent can pass through the same doorway in any one transition. If there are many interconnecting rooms, agents could pass through different doorways in the same transition, but no more than one through any single doorway at the same time. In the simple example to be considered here, where there are just two rooms as before, this cannot happen.

There is nothing particularly significant about the change in the example. The version discussed here is arguably more realistic, since it requires only that agents are able to detect when they are alone in a room with a member of the opposite sex; there is no need to assume that klaxons or other devices exist to inform agents that the situation has arisen in other rooms. The main reason for choosing the modified version, however, is simply that features of the original example, including in particular what is *sub-standard*($x$) and *unavoidably-red*($x$) there, have already been discussed. The modified version provides a slightly different example.

The states and transitions for the modified version are exactly the same as those for the original. The global colouring of states is the same; the global colouring of transitions is slightly different because that is partly determined by the local-global coherence constraint and in this version of the example the

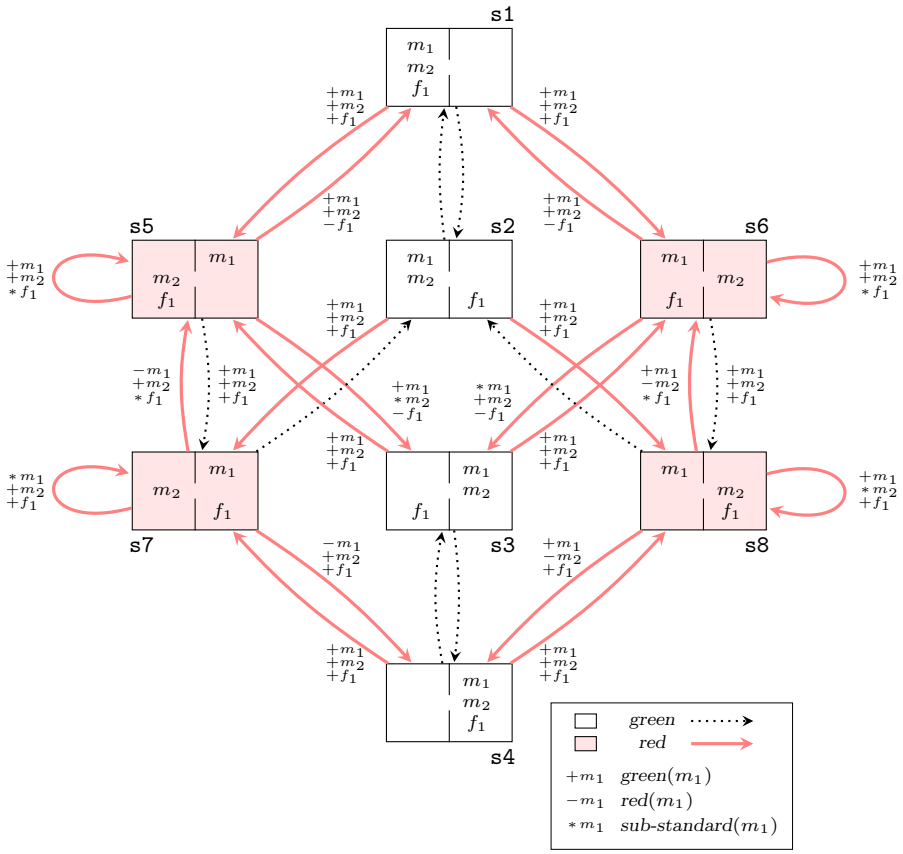


**Fig. 5.** The modified 'rooms' example. Transitions without annotations are coloured *green*($x$) for each agent $x$. Reflexive arcs on green nodes are omitted from the diagram.

agent-specific norms are different. We include a diagram of the transition system in Fig. 5 for ease of reference.

Notice that there are symmetries in the transition system because of symmetry in the example, between the two male agents $m_1$ and $m_2$, and between left and right. For that reason it is sufficient to look at transitions from just four states of the system and not all of them. We will show the transcripts for the transitions from the states in the top right quadrant of diagrams, that is, the two green states labelled `s1` and `s2` in the diagram, and the two red states `s6` and `s8`. Properties of the other states and transitions in the system are easily reconstructed by interchanging $m_1$ and $m_2$, or left and right, as the case may be.

We might begin by checking whether there are *degenerately-red*($x$) transitions in the system, or (globally) red sink states. Here are the relevant queries and the output produced by iCCalc for the example:

```
?-  satisfiable [-]:red(X) where agent(X).
** not satisfiable

?-  satisfiable -executable(trans=green).
** not satisfiable
```

We trust that the iCCalc syntax is sufficiently close to the syntax of formulas used in the paper that it requires no explanation. (`[-]` is the syntax for $[_{\text{alt}}]$.) `executable`($\varphi$) is shorthand for $\langle\varphi\rangle\top$. The first of the queries above is a transition formula asking whether there are any *degenerately-red*($x$) transitions. The second is a state formula asking whether there are any (globally) red sink states. (The query for *red*($x$)-sinks would be `satisfiable -executable(green(X))`. There are no *red*($x$)-sinks in the example.)

Here we see a difference between this version of the example and the original. As discussed earlier, the original version does have (globally) red sinks. There are two of them: one where $m_1$ and $f_1$ are on the left and $m_2$ is on the right (state `s6`), and another (by symmetry) where $m_2$ and $f_1$ are on the left and $m_1$ is on the right (state `s5`). These are not global red sinks in the modified example because, unlike in the original, when $m_1$ and $f_1$ are on the left and $m_2$ is on the right the agent-specific norms for $m_2$ do not require it to move left. In the original version of the example there are no globally green (*trans=green*) transitions from these states because of the local-global coherence constraint.

Further: in the original there are states from which there is no transition unless at least one agent fails to comply with its agent-specific norms. The following iCCalc query on the original version of the example

```
?-  satisfiable -executable(green(m1) & green(m2) & green(f1)).
```

finds two states where the formula is satisfied: they are also the two global red sinks. One can check the equivalence as follows:

```
?-  valid -executable(green(m1) & green(m2) & green(f1))
                                    <-> -executable(trans=green).
** valid
```

Note that this is not the same as:

```
?-  valid (green(m1) & green(m2) & green(f1)) <-> (trans=green).
** not valid
```

In the modified version of the example, in contrast, we get

```
?-  satisfiable -executable(green(m1) & green(m2) & green(f1)).
** not satisfiable
```

```
**transition t17:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
 :[m1:move=r,green(m1),green(m2),green(f1),trans=red]
1:[loc(m1)=r,loc(m2)=l,loc(f1)=l,alone(m2,f1),status=red]
    E+(m1):(trans=red)
    E+(m1):(0:(status=green) & 1:(status=red))

**transition t18:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
 :[m2:move=r,green(m1),green(m2),green(f1),trans=red]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
    E+(m2):(trans=red)
    E+(m2):(0:(status=green) & 1:(status=red))

**transition t19:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
 :[f1:move=r,green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
    E(f1):(0:(status=green) & 1:(status=green))

**transition t20:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
 :[green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
```

**Fig. 6.** Transitions from the green state `s1` (all three agents on the left)

Now let us look at the transcripts from iCCalc when we ask for annotations of the transitions (for the modified version of the example). We will consider first transitions from the two green states `s1` and `s2`.

Figure 6 shows the transitions from the state `s1`, where all three agents are on the left. The numbering of the transitions in the transcript is not significant. These identifiers are generated by iCCalc when the transition system is calculated from the $nC+$ formulation of the example. They are included merely for ease of reference.

There are no *unavoidably-red*($x$) or *sub-standard*($x$) transitions from this state. The state is green and so the agent-specific norms in the example do not impose any constraints on how the agents may move. However, one can see that in transitions `t17` and `t18`, where one of the male agents moves to the right and leaves the other alone with the female, the one who moves is (solely) responsible for bringing it about that the transition is globally red (*trans=red*). In both cases the male who moves is also (solely) responsible for bringing it about that the system state becomes red (*status=red*) in Hilpinen's sense.

```
**transition t13:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
 :[m1:move=r,green(m1),green(m2),green(f1),trans=red]
1:[loc(m1)=r,loc(m2)=l,loc(f1)=r,alone(m1,f1),status=red]
    E+(m1):(trans=red)
    E+(m1):(0:(status=green) & 1:(status=red))

**transition t14:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
 :[m2:move=r,green(m1),green(m2),green(f1),trans=red]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
    E+(m2):(trans=red)
    E+(m2):(0:(status=green) & 1:(status=red))

**transition t16:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
 :[f1:move=l,green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
    E(f1):(0:(status=green) & 1:(status=green))

**transition t15:
0:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
 :[green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
```

**Fig. 7.** Transitions from the green state `s2` ($m_1$ and $m_2$ on the left, $f_1$ on the right)

In contrast, when the female agent $f_1$ moves to the right (transition `t19`) she is responsible for sustaining the case that the system state is green (*status=green*). She is not solely responsible for sustaining it, however, since it also depends on how the male agents act: if the male agents both act as they do in `t19` (neither moves) then the system state remains green whether the female agent $f_1$ moves or not. The transition `t20` is the one where no agent moves in this state. There is nothing that we particularly want to say about it.

Now let us look at the other green state, `s2`. Figure 7 shows the transitions from this state. Although this state is not symmetrical to `s1` (the male agents and the female agent are in separate rooms here) the annotation of the transitions turns out to be the same as for `s1`. (There could of course be a difference if we specified a more extensive set of formulas to appear in annotations of transitions.)

Now let us look at the state `s6` where $m_1$ and $f_1$ are on the left and $m_2$ is on the right. Here the system is in a red state and the agent-specific norms impose some constraints on the behaviours of $m_1$ and $f_1$. Unlike the original version of the example, there are no agent-specific norms constraining $m_2$'s behaviours in this state since $m_2$ is in a different room from the other two.

The annotation produced by ICCALC for this state is as follows:

```
**state s6: [loc(m1)=l, loc(m2)=r, loc(f1)=l, alone(m1,f1), status=red]
    oblig(m1,-m1:move) = executable(-m1:move) & -permitted(m1,-(-m1:move))
    prohib(m1,m1:move=r) = executable(m1:move=r) & -permitted(m1,m1:move=r)
    oblig(f1,f1:move=r) = executable(f1:move=r) & -permitted(f1,-f1:move=r)
    prohib(f1,-f1:move) = executable(-f1:move) & -permitted(f1,-f1:move)
```

The (Boolean) state constant $alone(m_1, f_1)$ has the obvious interpretation. It is convenient to include $alone(x, y)$ constants in $nC+$ formulations of larger versions of the example, where there are many rooms and more agents.

The state annotation also shows some further notational abbreviations that we find convenient. Let $\alpha$ be a formula of $\sigma^{\mathrm{a}}$, that is, a propositional formula of event atoms. It is natural to say that $\alpha$ is permitted for $x$ in a state $s$ according to the agent-specific norms for $x$ when there is a transition of type $\alpha$ from $s$ which is $green(x)$. Accordingly, we define:

$$permitted(x, \alpha) =_{\mathrm{def}} executable(\alpha \wedge green(x))$$

Here, as usual, $\varphi$ is 'executable' means only that there exists a transition of type $\varphi$ from the current state: $executable(\varphi)$ is shorthand for the state formula $\langle\varphi\rangle\top$. In practice, $\alpha$ in an expression $permitted(x, \alpha)$ will always be a propositional formula of atoms of the form $x{:}a{=}v$.

We can define a sense of 'obligatory' and 'prohibited' action in similar fashion. As a first stab, an event of type $\alpha$ is *prohibited* for $x$ in a state $s$ according to the agent-specific norms for $x$ if every transition of type $\alpha$ from state $s$ is $red(x)$. However, that would mean that if there is no transition of type $\alpha$ in state $s$ at all then $\alpha$ is prohibited for $x$. It is more informative if we add that there must be at least one transition of type $\alpha$ from $s$:

$$prohib(x, \alpha) =_{\mathrm{def}} executable(\alpha) \wedge \neg executable(\alpha \wedge green(x))$$

(where 'executable' has its usual meaning). The above is equivalently expressed as

$$prohib(x, \alpha) =_{\mathrm{def}} executable(\alpha) \wedge \neg permitted(x, \alpha)$$

which is the form that appears in the state annotation shown.

Similarly, it is natural to say that $\alpha$ is *obligatory* for $x$ in a state $s$ according to the agent-specific norms for $x$ when there is at least one transition of type $\alpha$ from state $s$, and every $green(x)$ transition from $s$ is of type $\alpha$ (equivalently, there are no $green(x)$ transitions from state $s$ of type $\neg\alpha$). This can be expressed as:

$$oblig(x, \alpha) =_{\mathrm{def}} executable(\alpha) \wedge \neg executable(\neg\alpha \wedge green(x))$$

which is also equivalent to:

$$oblig(x, \alpha) =_{\mathrm{def}} executable(\alpha) \wedge \neg permitted(x, \neg\alpha)$$

This is the definition that is shown in the state annotation above.

The state annotation shown may give the impression that it is not necessary to have both *oblig* and *prohib*: one seems to repeat what the other says. But that is just a feature of the simplicity of this particular example. In this particular example, an agent in the left hand room can only move to the right or stay where it is: it must do one or the other. In more complicated examples, it may have many other options, and then the difference between *oblig* and *prohib* becomes marked.

It should be noted that these defined forms express only *one* sense in which $\alpha$ could be said to be permitted/obligatory/prohibited for $x$ according to the agent-specific norms for $x$. We do not have space to discuss any other possibilities in this paper.

The transitions from state `s6` are shown in Fig. 8. In transition `t21` the male agent $m_1$ moves right in contravention of the agent-specific norms that require it to stay where it is in this state. The transition is *sub-standard*($m_1$) because $m_1$ could have complied with its agent-specific norms but does not in this transition. (It is also the case that $\mathrm{E}^+_x\mathrm{E}^+_x$*red*($m1$) is true at `t21`; the ICCALC annotation would show if it were false. Compare transition `t24` below.) Transition `t21` is also *unavoidably-red*($f_1$): $f_1$ is prevented from complying with her agent-specific norms by the actions of others in this transition. In this particular case, the transcript shows that it is $m_1$'s actions that prevent $f_1$ from moving right as required by her agent-specific norms. Transition `t21` also provides an example where two different agents ($m_1$ and $f_1$ here) both bring about, are both responsible for, the transition's being globally red (*trans=red*) and thus in contravention of the system norms. We also see that $m_1$ is solely responsible in this transition for bringing it about that the system state becomes green, that is, moves from a red state to a green state. So here we have an example where the system recovers from a red system state to a green system state, but where the transition itself is (globally) red, and therefore in contravention of the system norms, and where the agent $m_1$ who is solely responsible for the recovery from a red system state to a green system state does so by acting in contravention of its own agent-specific norms.

Transition `t22`, in which $m_2$ moves left, is similar but not symmetric with `t21`. Here, the agent-specific norms for $m_2$ do not require it to stay where it is because $m_2$ is not in the same room as $m_1$ and $f_1$. $m_2$ is thus free to move according to its agent-specific norms, but if it does move, then it makes it impossible for the female agent $f_1$ to comply with hers: the transition is *unavoidably-red*($f_1$), and as the transcript shows, it is $m_2$ who is responsible (though not solely responsible) for making it so. Transition `t22` is also another example of two different agents ($m_2$ and $f_1$) both bringing it about that a transition is of a particular type (globally red). $m_2$ is also solely responsible for bringing it about that the system recovers (becomes green), though unlike in `t21`, not in contravention of its own agent-specific norms.

Transition `t23` is straightforward. Here all three agents comply with their agent-specific norms. $f_1$ however, although acting in compliance with her agent-specific norms by moving to the right, nevertheless is thereby responsible (though

```
**transition t21:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
 :[m1:move=r,red(m1),green(m2),red(f1),trans=red]
1:[loc(m1)=r,loc(m2)=r,loc(f1)=l,status=green]
    substandard(m1) = E+(m1):red(m1)
    unavoidably_red(f1) = [-f1]:red(f1)
        E(m1):red(f1)
    E(m1):(trans=red)
    E(f1):(trans=red)
    E+(m1):(0:(status=red) & 1:(status=green))

**transition t22:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
 :[m2:move=l,green(m1),green(m2),red(f1),trans=red]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=l,status=green]
    unavoidably_red(f1) = [-f1]:red(f1)
        E(m2):red(f1)
    E(m2):(trans=red)
    E(f1):(trans=red)
    E+(m2):(0:(status=red) & 1:(status=green))

**transition t23:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
 :[f1:move=r,green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
    E(f1):(0:(status=red) & 1:(status=red))

**transition t24:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
 :[green(m1),green(m2),red(f1),trans=red]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
    substandard(f1) = E+(f1):red(f1)
        -E+(f1):E+(f1):red(f1)
    E+(f1):(trans=red)
        -E+(f1):E+(f1):(trans=red)
```

**Fig. 8.** Transitions from the red state `s6` ($m_1$ and $f_1$ on the left, $m_2$ on the right)

not solely responsible) for sustaining the case that the system remains in a red system state. As with other similar examples, one should be very careful not say that an agent behaves badly if it is responsible for sustaining, or bringing about, that a system state remains, or becomes, a red system state. It may also act well in the same transition, in the sense that it complies with its agent-specific norms. System norms and agent-specific norms are related, for instance by local-global coherence, but they express different standards of legality, acceptability,

desirability, and therefore different standards of what it means to say that an agent acts well or acts badly.

Finally, transition `t24`, in which no agent moves, is *sub-standard*($f_1$) because here $f_1$ could have complied with her agent-specific norms but did not. She is also solely responsible for bringing about that the transition is globally red. Note though, that although $\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is true at `t24` (this is what *sub-standard*($f_1$) means), $\mathrm{E}^+_{f_1}\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is not true. In general $\mathrm{E}^+_x\varphi \rightarrow \mathrm{E}^+_x\mathrm{E}^+_x\varphi$ is not valid. Here we have an example. We can see that $[f_1]\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is not true at `t24`. If it were, that would mean $\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is true at every transition from state `s6` in which $f_1$ acts as she does in `t24`, i.e., does not move. Transitions `t21` and `t22` are both like this, but $\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is not true at either of them: neither of them is *sub-standard*($f_1$). And if $[f_1]\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$ is not true at `t24` then neither is $\mathrm{E}^+_{f_1}\mathrm{E}^+_{f_1}\, \mathit{red}(f_1)$. Similarly for $\mathrm{E}^+_{f_1}\, (\mathit{trans}{=}\mathit{red})$; $[f_1]\mathrm{E}^+_{f_1}\, (\mathit{trans}{=}\mathit{red})$ is not true at `t24`, as is easily confirmed.

To complete the picture, here is the ICCALC output for the other red state, `s8`.

```
**state s8: [loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
    oblig(m2,m2:move=l) = executable(m2:move=l) & -permitted(m2,-m2:move=l)
    prohib(m2,-m2:move) = executable(-m2:move) & -permitted(m2,-m2:move)
    oblig(f1,-f1:move) = executable(-f1:move) & -permitted(f1,-(-f1:move))
    prohib(f1,f1:move=l) = executable(f1:move=l) & -permitted(f1,f1:move=l)
```

The transitions from this state are shown in Fig. 9. We do not provide a commentary. Although the details are different, the general points we wish to make have already been discussed. (When a formula $\mathrm{E}^+_x\varphi$ is true, $\mathrm{E}^+_x\mathrm{E}^+_x\varphi$ is also true unless shown otherwise.)

## 8 Conclusion

We have presented a modal-logical language for talking about properties of states and transitions of a labelled transition system and, by introducing agent 'strands' as a component of transitions, for talking about what transition properties are necessary for how a particular agent, or group of agents, acts in a particular transition. This allows us in turn to introduce two defined 'brings it about' modalities. The novel feature is that 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. We are thereby able to make explicit the notions of agency that underpin various forms of norm compliant or non-compliant behaviour, and to be able to discuss relationships between system norms and agent-specific norms using the formal language. 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. We are also able to express when an agent, or group of agents, is responsible, solely

```
**transition t9:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
 :[m1:move=r,green(m1),red(m2),green(f1),trans=red]
1:[loc(m1)=r,loc(m2)=r,loc(f1)=r,status=green]
    unavoidably_red(m2) = [-m2]:red(m2)
        E(m1):red(m2)
    E(m1):(trans=red)
    E(m2):(trans=red)
    E+(m1):(0:(status=red) & 1:(status=green))

**transition t10:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
 :[m2:move=l,green(m1),green(m2),green(f1),trans=green]
1:[loc(m1)=l,loc(m2)=l,loc(f1)=r,status=green]
    E+(m2):(0:(status=red) & 1:(status=green))

**transition t12:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
 :[f1:move=l,green(m1),red(m2),red(f1),trans=red]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=l,alone(m1,f1),status=red]
    unavoidably_red(m2) = [-m2]:red(m2)
        E(f1):red(m2)
    substandard(f1) = E+(f1):red(f1)
    E(m2):(trans=red)
    E(f1):(trans=red)
    E(f1):(0:(status=red) & 1:(status=red))

**transition t11:
0:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
 :[green(m1),red(m2),green(f1),trans=red]
1:[loc(m1)=l,loc(m2)=r,loc(f1)=r,alone(m2,f1),status=red]
    substandard(m2) = E+(m2):red(m2)
        -E+(m2):E+(m2):red(m2)
    E+(m2):(trans=red)
        -E+(m2):E+(m2):(trans=red)
```

**Fig. 9.** Transitions from the red state `s8` ($m_1$ on the left, $m_2$ and $f_1$ on the right)

or otherwise, for bringing about that a transition complies with system norms, for bringing it about that the system recovers from a red system state to a green system state, for sustaining the case that the system remains in a green system state, and so on.

Besides the generalisation to (unwitting) collective agency [3] there are three main directions of current work.

(1) *Scaleability* It might be felt that the 'rooms' example used in this paper is too simple to be taken seriously as representative of real-world domains. We deliberately chose the simplest configuration of rooms and agents that allowed us to make the points we wanted to make while still being able to be depicted in their entirety. The example works just as well with more rooms, more than two categories of agents, and a wider repertoire of actions that the agents are able to perform. Generally, the issues we have addressed arise whenever we put together a complex system of interacting agents, acting independently, whose behaviours are subject to their own agent-specific norms, and where we wish to impose further system norms to regulate possible interactions.

Nevertheless, it is clear that serious issues of scaleability remain, and that in particular we confront the same state explosion problems that arise in all modelling approaches of this kind. These are problems, however, that are the subject of extensive current research. There is nothing that prevents us from applying emerging techniques and solutions to agent-stranded transition systems too.

One promising direction that we are exploring is the use of agent-centric projections. Roughly, given a model $\mathcal{M}$ describing system behaviour, it is possible to define a projection $\mathcal{M}_x$ in which all states and transitions indistinguishable for an agent $x$ are collapsed into one, and all other states and transitions are discarded. $\mathcal{M}_x$ thus models system behaviour from an individual agent $x$'s perspective. Some information is lost, but (depending of course on what is indistinguishable for an individual $x$), the projection $\mathcal{M}_x$ is much smaller and more manageable than the full model $\mathcal{M}$.

(2) *Agent-specific norms* One fundamental feature of agent-specific norms, as we see it, is that to be effective or even meaningful in guiding the actions of an individual agent $x$ they must be formulated in terms of what the agent $x$ can actually sense/perceive of its environment, and the actions that an agent $x$ can actually perform. We referred to the 'absence of moral luck' constraint as the minimal requirement we must impose on agent-specific norms. To do a proper job it is necessary to refine and extend the semantical structures in order to model these features explicitly. This part is not so difficult. We will present the details in another paper. There is also the further question of how agent-specific norms once formulated can be incorporated into an agent's implementation—in the case of a 'lightweight' reactive agent, how to modify its program code to take agent-specific norms into account, and in the case of a deliberative agent, how to represent the agent-specific norms in a form that the agent can use in its reasoning processes. We have very little to say about that yet.

(3) *The representation of norms* We gave in the paper one simple formulation of what it can mean to say that an action is obligatory or permitted for $x$ according to the agent-specific norms for $x$. There are many other variations and distinctions that can be expressed using the resources of the language. Generally, the logic of norms and the logic of action/agency have often been studied together, and it remains to explore how the full resources of the language can be used

to articulate distinctions and issues that have previously been discussed in the literature. Further, it is well known in the field of deontic logic that a simple binary classification of states and/or transitions into green/red (ideal/sub-ideal, permitted/not permitted) is too simple to deal adequately with many kinds of norms. In [6], for instance, we presented a refinement of the current approach in which the states of a transition systems were ordered depending on how well each complied with a set of explicitly stated norms. Much more remains to be done along these lines.

## Acknowledgements

The characterisation of non-compliant behaviours, agent-specific norms, the 'rooms' example, and the iCCalc implementation is joint work with Robert Craven. I am grateful to Alex Artikis for helpful comments on an earlier draft.

## References

1. Craven, R., Sergot, M.: Agent strands in the action language nC+. Journal of Applied Logic **6**(2) (June 2008) 172–191
2. Pörn, I.: Action Theory and Social Science: Some Formal Models. Number 120 in Synthese Library. D. Reidel, Dordrecht (1977)
3. Sergot, M.: The logic of unwitting collective agency. Technical Report 2008/6, Department of Computing, Imperial College London (2008)
4. Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N., Turner, H.: Nonmonotonic causal theories. Artificial Intelligence **153**(1–2) (2004) 49–104
5. Sergot, M.: $(\mathcal{C}+)^{++}$: An action language for modelling norms and institutions. Technical Report 2004/8, Department of Computing, Imperial College London (2004)
6. Sergot, M., Craven, R.: The deontic component of action language nC+. In Goble, L., Meyer, J.J.C., eds.: Deontic Logic and Artificial Normative Systems. Proc. 8th International Workshop on Deontic Logic in Computer Science (DEON'06), Utrecht, July 2006. LNAI 4048, Springer Verlag (2006) 222–237
7. Große, G., Khalil, H.: State Event Logic. Journal of the IGPL **4**(1) (1996) 47–74
8. Venema, Y.: Points, lines and diamonds: a two-sorted modal logic for projective planes. Journal of Logic and Computation **9**(5) (1999) 601–621
9. Sauro, L., Gerbrandy, J., van der Hoek, W., Wooldridge, M.: Reasoning about action and cooperation. In: Proceedings of the Fifth International Joint Conference on Autonomous agents and Multiagent Systems: AAMAS'06, New York, NY, USA, ACM (2006) 185–192
10. von Wright, G.H.: Norm and Action—A Logical Enquiry. Routledge and Kegan Paul, London (1963)
11. Chellas, B.F.: Modal Logic—An Introduction. Cambridge University Press (1980)
12. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press (2001)
13. Carmo, J., Jones, A.J.I.: Deontic database constraints, violation and recovery. Studia Logica **57**(1) (1996) 139–165
14. Meyden, R.: The dynamic logic of permission. Journal of Logic and Computation **6**(3) (1996) 465–479

15. Meyer, J.J.C.: A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. Notre Dame Journal of Formal Logic **29**(1) (1988) 109–136
16. Lomuscio, A., Sergot, M.J.: Deontic interpreted systems. Studia Logica **75**(1) (October 2003) 63–92
17. Ågotnes, T., van der Hoek, W., Rodriguez-Aguilar, J.A., Sierra, C., Wooldridge, M.: On the logic of normative systems. In Veloso, M.M., ed.: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), California, AAAI Press (2007) 1175–1180
18. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning about Knowledge. MIT Press, Cambridge (1995)
19. Belnap, N., Perloff, M.: Seeing to it that: a canonical form for agentives. Theoria **54** (1988) 175–199
20. Horty, J.F., Belnap, N.: The deliberative stit: a study of action, omission, ability, and obligation. Journal of Philosophical Logic **24**(6) (1995) 583–644
21. Horty, J.F.: Agency and Deontic Logic. Oxford University Press (2001)
22. Segerberg, K.: Getting started: Beginnings in the logic of action. Studia Logica **51**(3–4) (1992) 347–378
23. Hilpinen, R.: On action and agency. In Ejerhed, E., Lindström, S., eds.: Logic, Action and Cognition—Essays in Philosophical Logic. Volume 2 of Trends in Logic, Studia Logica Library. Kluwer Academic Publishers, Dordrecht (1997) 3–27
24. Chellas, B.F.: The Logical Form of Imperatives. Dissertation, Stanford University (1969)
25. von Wright, G.H.: An essay in deontic logic and the general theory of action. Number 21 in Acta Philosophica Fennica. (1968)
26. von Wright, G.H.: Practical Reason. Blackwell, Oxford (1983)