Notions of Instrumentality in Agency Logic

Kees van Berkel(✉) and Matteo Pascucci(✉)

Institute of Logic and Computation, TU Wien, Vienna, Austria
kees@logic.at, matteo.pascucci@tuwien.ac.at

Abstract. We present a logic of agency called *LAE* whose language includes propositional constants for actions and expectations. The logic is based on Von Wright's theory of agency in general and his analysis of instrumentality in particular. An axiomatization of the logic, including an independence of agents axiom, is provided and soundness and completeness are shown with respect to its intended class of frames. The framework of *LAE* will allow us to formally define a manifold of concepts involved in agency theories, including Von Wright's four elementary forms of action, the notion of forbearance and notions of instrumentality that make reference to an agent's expectations.

Keywords: Action logic · Agency · Expectations · Instrumentality

1 Introduction

What do we mean when we ascribe *agency* to a human being? We most likely assert that this person has the ability to perform an action. This answer highlights two key aspects of agency: *ability* and *action*. A third key aspect of agency is that actions can be seen in most cases as means to an end; that is, as *instruments*. The present work provides a logical framework to reason about the interplay of these three aspects of agency. While the notions of ability and action have been formally addressed for the past few decades, the notion of instrumentality seems to have received minor attention in the literature thus far. Philosophical analyses of instrumentality as such are scarce, although the concept of 'means to an end' is paramount to any theory of agency. Despite these limitations, we believe that logical investigations around instrumentality should be established on firm philosophical grounds. The present work aims at providing a formal account of instrumentality within a framework of agency logic and will be largely based on ideas presented by Georg Henrik von Wright [13–15], who can be regarded as one of the founding fathers of the logic of action [2].

Two prominent formal frameworks have been developed for the last few decades with respect to the logical treatment of agency: *stit-logic* [4,10] and *propositional dynamic logic* (PDL) [7,8]. The main difference between the two approaches can be pinpointed as follows: in stit-logic the focus has been largely

T. Miller et al. (Eds.): PRIMA 2018, LNAI 11224, pp. 403–419, 2018.
https://doi.org/10.1007/978-3-030-03098-8_25

put on the formal treatment of (explicit) agents on the basis of available choices, whereas in PDL the focus has been put on the formal analysis of (explicit) actions, regarded as transitions between states. In this article we reconstruct both frameworks within a logic including propositional constants for actions and expectations called *LAE* (*logic of actions and expectations*); our contribution is related to previous proposals that aim either at extending one framework to include the other, such as [16], or at defining one framework within the other, such as [9]. Our main purpose is to use *LAE* in order to provide a formal definition of various notions of instrumentality that rely on Von Wright's ideas. Special attention will be paid to how these notions interact with an agent's expectations. The article is divided as follows: in Sect. 2 we present and elaborate on Von Wright's ideas; in Sect. 3 we introduce the system *LAE* and prove its soundness and completeness. Finally, in Sect. 4, we formally specify the main notions of the theory of agency and instrumentality at issue.

2 A Theory of Agency and Instrumentality

2.1 Acting

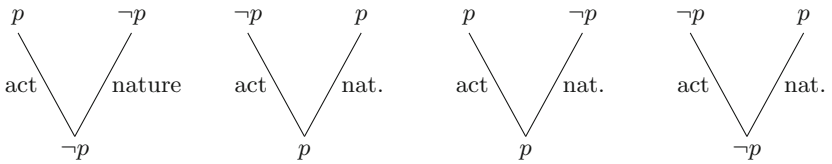
To 'bring about something' and to 'prevent something' are essential characteristics of what it is to act. What is brought about is a *state of affairs* and, for that reason, to 'see to it that p' means that one acts "in such a manner that the state of affairs that p is the result of one's action" [13, p. 37]. From this account it follows that acting is strongly related to the *emergence* of a particular result (perhaps not always the desired one). An account of action, hence, heavily depends on the notion of change.

A change is a *transition* from an initial state to an end-state. These transitions can be triggered by events in which agents play no role (e.g., a moon eclipse); however, in many cases they are triggered by an agent's behaviour. In particular, an agent may decide to act or not to act in a certain way in a given circumstance and this behaviour may produce several different results (at least, in a non-deterministic world). For this reason, we say that an action triggers a set of possible transitions from an initial state to a set of end-states. To act, then, is to provoke a specific form of change: it is a change brought about by the interference of an agent with the "course of nature" [15, p. 36]; one can *a posteriori* say that if the agent had not acted, the course of history *would* have been different. This is what Von Wright calls the *counterfactual element* of action [13, p. 43].

In order to understand how a result p is related to an action, one also has to take into account whether p holds or not in the initial state. Indeed, an agent may bring about p in two ways: either the initial state is $\neg p$ and the agent's behaviour changes it to the result p, or the initial state is p and the agent prevents it from changing to $\neg p$ [15, p. 42]. Summing up, the analysis in this section provides us with three main characteristics of action: (i) the initial state, (ii) the result of the agent's behaviour (i.e., the end-state) and (iii) the counterfactual 'course of nature'. Taking into account also the difference between p and $\neg p$ as atomic

results, Von Wright classifies *four elementary forms of agent behaviour*; the first two concern actions that 'bring about' something, the latter two are actions that 'prevent' something from happening [15, pp. 43–44]:

- *producing* p: constructively bringing p into 'existence' (Fig. 1a);[1]
- *destroying* p: without the agent's acting p would have 'prevailed' (Fig. 1b);
- *preserving* p: if the agent does not act, then p will 'perish' (Fig. 1c);
- *suppressing* p: if the agent does not act, then p will 'emerge' (Fig. 1d).



(a) producing (b) destroying (c) preserving (d) suppressing

Fig. 1. Von Wright's four elementary types of action.

2.2 Actions

Up until now we have been talking about 'acts' without specification. Commonly, a distinction is made between two sorts of actions: actions described in an impersonal, generic way (e.g. 'writing') and concrete, individual instances of these generic actions, as performed by a particular agent at a particular time (e.g. 'I am currently writing'). The former are frequently called '*action-types*', whereas the latter can be named '*action-tokens*'. Following Von Wright, generic actions (i.e., types) can be regarded as 'categories' to which individual 'cases' (i.e., tokens) belong [15, p. 36].

Here we will generalize this account of actions by considering also *negative actions* and *complex actions*. This will enable us to speak of, for instance, the action-type 'not opening the door' and the action-type 'not opening the door or closing the window'. Negative actions are usually not expressible in the language of propositional dynamic logic, but they are taken into account in other formal approaches to agency which make explicit reference to actions, such as [2] and [3]. We will regard both action-types and action-tokens as essential to our logic of agency: As was pointed out in the previous section, an agent's behaviour at a particular state triggers a set of possible transitions and, therefore, represents an action-token. Moreover, as we will clarify in the next section, a proper notion of instrumentality makes reference to action-types; that is, in order to determine whether an action is a good instrument for a given purpose, one has to consider the outcomes of previous transitions triggered by actions of that type.

[1] The term used by Von Wright for this behaviour is 'doing p'. We avoid this expression because we reserve 'doing' for actions, and use 'producing' for results.

2.3 Instrumentality

Actions can be regarded as instrument serving a particular purpose; they are 'means to an end'. For instance, 'pressing Y on the keyboard' and 'pulling the handbrake of a car' are respectively instruments to 'confirm a procedure on a computer terminal' and to 'perform an emergency stop'. In this section several distinct forms of instrumentality will be presented that will be formally addressed in subsequent sections. As a philosophical basis, we will borrow from and extend Von Wright's analysis of instrumental goodness, as presented in [14, pp. 19–40]. To avoid ambiguity, the term 'proper instrument' is here regarded as an appropriate synonym for 'good instrument' and they will be used interchangeably.

Let us call an *intended* state of affairs ϕ a purpose and an action Δ an instrument. Paraphrasing Von Wright, an action Δ will qualify as a ϕ-instrument if and only if Δ can serve the purpose ϕ [14, p. 21]. It is also important to distinguish between instruments that can serve the purpose ϕ *simpliciter* and those that can serve ϕ *well*. The former will be called ϕ-instruments and the latter *proper* ϕ-instruments.

To qualify a particular instrument suitable for a particular purpose, we base our judgment on *past performance*; for example, with respect to questions of instrumentality we often make remarks such as 'it has worked before' and 'it has never disappointed me (thus far)'. In the first case, we recognize a weak criterion; that is, the instrument *has* served the purpose at least once and, for that reason, it *can* serve the purpose. In the latter case, we identify a stronger criterion for instrumentality; that is, there have been applications of the instrument and these applications *have always served* the purpose and, for that reason, the instrument serves the purpose *well*. Hence, notions of instrumentality are based on past experience. This experience, subsequently, can be either impersonal or personal (e.g., 'this machine has been tested' or 'I have used this tool before'). Thus far, we established two definitions of impersonal instrumentality:

(1) AGENT-INDEPENDENT BASIC INSTRUMENTALITY: action-type Δ is a basic ϕ-instrument if and only if Δ has served the purpose ϕ at least once in the past.
(2) AGENT-INDEPENDENT PROPER INSTRUMENTALITY: action-type Δ is a proper ϕ-instrument if and only if (i) Δ is a basic ϕ-instrument and (ii) Δ has always served the purpose ϕ in the past.

Hence, notions of instrumentality relate to both purpose and past performance. However, when we judge that 'these scissors are a proper instrument for me to cut this piece of paper', what do we mean? Von Wright briefly remarks that "judgments of instrumental goodness, usually, even if not necessarily, contain a conjectural element" [14, p. 27]. In other words, practical statements about instrumentality also contain reference to *expectations* about the instrument's future performance. Hence, agent-bound instrumentality is based on both (i) the past performance of particular action-tokens associated with a certain type and (ii) the expected continuation of this performance in the nearby future. In contrast to agent-independent statements of instrumentality, statements of this

form will vary over agents. What is more, the conjectural element of expected performance does not guarantee any future result: the agent might simply be wrong [14, p. 27]. The fact that the instrument has served the purpose well in the past, does not guarantee that it will not fail in the future. In our formal framework we will strongly emphasize these fundamental aspects of agent-bound instrumentality by investigating different notions of instrumentality that are restricted by the agent's expectations.

Lastly, we emphasize that expectations must be regarded as those future moments which the agent considers *more likely to happen*. An agent's expectations about the nearby future are therefore a subset of all possible next moments. We will accordingly introduce a formal restriction on expectations in Sect. 3.[2]

From the above we derive two agent-bound definitions of instrumentality:

(3) AGENT-BOUND BASIC INSTRUMENTALITY: An instrument Δ is a basic ϕ-instrument for agent α at moment m if and only if (i) Δ is a basic ϕ-instrument and (ii) α expects that Δ will serve ϕ at m.

(4) AGENT-BOUND PROPER INSTRUMENTALITY: An instrument Δ is a proper ϕ-instrument for agent α at moment m if and only if (i) Δ is a proper ϕ-instrument and (ii) α expects that Δ will serve ϕ at m.

The agent-independent and agent-dependent notions of instrumentality (1)–(4) will be formally addressed in Sect. 4.

In passing, *ability* can be regarded as an abstract form of agentive instrumentality; namely, saying that 'an agent is able to behave in a certain way which guarantees a result' is an abstraction of saying that 'there exists an instrument (action) which the agent can successfully employ to obtain that result'. Moreover, saying that an agent α is able to obtain ϕ through an action Δ, given that Δ has always led α to ϕ in the past, is essentially the same as saying that α *excels* at performing Δ to obtain ϕ. In this sense, Von Wright's concept of ability, 'being good *at* something', is strongly related to our concept of agent-bound proper instrumentality (cf. the analysis of 'technical goodness' as ability and skill in [14, pp. 32–39]).

3 The System *LAE*

We start our formal presentation with a boolean algebra of actions and subsequently introduce the language of the logic LAE, in which the performance of an action by an agent will be represented by a formula. Let $Action = \{\delta_1, ..., \delta_n\}$ be a finite set of atomic action-types. The set $Action^*$ of complex action-types is defined by the following BNF:

$$\Delta ::= \delta_i | \Delta \cup \Delta | \overline{\Delta}$$

[2] We want to stress that the term 'expectation' must not be regarded as an epistemic notion, such as knowledge. Although an agent can have expectations about the future, the agent might still have imperfect knowledge of these expected future states.

where $\delta_i \in Action$. The operations $\cup$ and — are respectively used to form *disjunctions of action-types* (e.g., 'turning-left or turning-right') and *negations of action-types* (e.g., 'not turning-right'). If $Agent = \{\alpha_1, ..., \alpha_m\}$ is a finite set of agent constants, an *agent-bound action-type* is an expression of kind Δ^{α_i}, where $\Delta \in Action^*$ and $\alpha_i \in Agent$. Let $Var = \{p_1, p_2, p_3, ...\}$ be a countable set of propositional variables; furthermore, for any $\alpha_i \in Agent$, let $Wit^{\alpha_i} = \{\mathfrak{d}_1^{\alpha_i}, ..., \mathfrak{d}_n^{\alpha_i}\}$ be a set of propositional constants respectively witnessing the performance of the atomic action-types δ_1, ..., δ_n by α_i and let $\mathfrak{e}^{\alpha_i}$ be a propositional constant witnessing the compatibility of a state with α_i's expectations.[3] Notice that $|Wit^{\alpha_i}| = |Action| = n$. The set $\bigcup_{\alpha_i \in Agent} Wit^{\alpha_i}$ can be simply denoted by Wit. The language $\mathcal{L}$ is defined by the following BNF:

$$\phi ::= p_i | \mathfrak{e}^{\alpha_j} | \mathfrak{d}_i^{\alpha_j} | \neg\phi | \phi \to \phi | \Box\phi | N\phi$$

for any $p_i \in Var$, $\alpha_j \in Agent$ and $\mathfrak{d}_i^{\alpha_j} \in Wit$. We can read $\Box\phi$ as 'in all successor states ϕ is the case' and $N\phi$ as 'in the actual successor state ϕ is the case'. We use standard definitions for additional boolean and modal operators. For instance, $\Diamond\phi$ abbreviates $\neg\Box\neg\phi$ and means 'in some successor state ϕ is the case'. Expressions like $\mathfrak{e}^{\alpha_j}$ and $\mathfrak{d}_i^{\alpha_j}$ mean respectively 'the most recent expectations of agent α_j are met' and 'agent α_j has just performed action δ_i'. The set of atomic propositional symbols in $\mathcal{L}$ is $Atom = Var \cup Wit \cup \{\mathfrak{e}^{\alpha_j} : \alpha_j \in Agent\}$.

Let t be a translation function mapping agent-bound action-types to formulas of $\mathcal{L}$ as below:

- for any $\delta_i \in Action$ and $\alpha_j \in Agent$, $t(\delta_i^{\alpha_j}) = \mathfrak{d}_i^{\alpha_j}$,
- for any $\Delta \in Action^*$ and $\alpha_i \in Agent$, $t(\overline{\Delta^{\alpha_i}}) = \neg t(\Delta^{\alpha_i})$;
- for any $\Delta, \Gamma \in Action^*$ and $\alpha_i, \alpha_j \in Agent$, $t(\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}) = t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j})$.

Let LAE be the system specified below:

A0 if ϕ is a propositional tautology, then $\vdash_{LAE} \phi$;
R0 $\phi, \phi \to \psi \vdash_{LAE} \psi$;
A1 $\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$;
R1 if $\vdash_{LAE} \phi$, then $\vdash_{LAE} \Box\phi$;
A2 $N(\phi \to \psi) \to (N\phi \to N\psi)$;
A3 $\neg N\phi \to N\neg\phi$;
A4 $\Box\phi \to N\phi$;
A5 for any list of (distinct) $\alpha_1, ..., \alpha_n \in Agent$ and list of (non-necessarily distinct) $\Delta_1, ..., \Delta_n \in Action^*$,
$(\Diamond t(\Delta_1^{\alpha_1}) \wedge ... \wedge \Diamond t(\Delta_n^{\alpha_n})) \to \Diamond(t(\Delta_1^{\alpha_1}) \wedge ... \wedge t(\Delta_n^{\alpha_n}))$;
A6 for any $\alpha_j \in Agent$, $\Diamond\mathfrak{e}^{\alpha_j} \to \Diamond\neg\mathfrak{e}^{\alpha_j}$.

The most relevant axioms of the system S are A3, which guarantees that every state has a unique successor, A4, which says that the actual successor of a state is within the set of its successors, A5, which represents the stit-logic principle known as *independence of agents*, and A6, which ensures that agents never expect

[3] The use of propositional constants in modal logic can be traced back at least to [1].

all possible future state-of-affairs to happen (if at a given state there are successor states satisfying an agent's expectations, then there are also successor states not satisfying the expectations).[4] The semantics for LAE will clarify that none of these axioms implies that a state has successors. Thus, the system can be used to reason about scenarios in which there are final possible states. Furthermore, it is noteworthy that the principle of 'independence of agents' is compatible with a scenario in which an agent ends in a state that does not meet that agent's (most recent) expectations.

We define the following additional operators:

E1 for any $\Delta \in Action^*$ and $\alpha_i \in Agent$,
$[\Delta^{\alpha_i}]^{would}\phi =_{def} \Box(t(\Delta^{\alpha_i}) \rightarrow \phi)$;
E2 for any $\Delta \in Action^*$ and $\alpha_i \in Agent$,
$[\Delta^{\alpha_i}]^{could}\phi =_{def} \Box(t(\Delta^{\alpha_i}) \rightarrow \phi) \wedge \Diamond t(\Delta^{\alpha_i})$;
E3 for any $\Delta \in Action^*$ and $\alpha_i \in Agent$,
$[\Delta^{\alpha_i}]^{will}\phi =_{def} \Box(t(\Delta^{\alpha_i}) \rightarrow \phi) \wedge \neg N \neg t(\Delta^{\alpha_i})$.

We can read the formula $[\Delta^{\alpha_i}]^{would}\phi$ as 'at the current state, by behaving in accordance with Δ, α_i would bring about ϕ'. (Notice that this does not ensure that α_i is currently able to behave in accordance with Δ.) The formula $[\Delta^{\alpha_i}]^{could}\phi$ means 'at the current state, by behaving in accordance with Δ, α_i would bring about ϕ and α_i could (i.e., is able to) behave in accordance with Δ'. Finally, the formula $[\Delta^{\alpha_i}]^{will}\phi$ means 'at the current state, by behaving in accordance with Δ, α_i would bring about ϕ and α_i will actually behave in accordance with Δ'.

A relational frame to interpret the language $\mathcal{L}$ is an ordered tuple $\mathfrak{F} = \langle W, \{W_{\mathfrak{d}_i^{\alpha_j}} : \mathfrak{d}_i^{\alpha_j} \in \mathcal{L}\}, \{W_{\mathfrak{e}^{\alpha_j}} : \mathfrak{e}^{\alpha_j} \in \mathcal{L}\}, R, R_N\rangle$, where $W = \{w_1, w_2, w_3, ...\}$ is a set of states, each $W_{\mathfrak{d}_i^{\alpha_j}}$ and each $W_{\mathfrak{e}^{\alpha_j}}$ is a subset of W and R and R_N are binary relations over W. The relation R captures the idea of a transition from a state to one of its immediate successors. As we pointed out in Sect. 2, a transition can be triggered by any event and so it does not require, in general, an active interference of an agent. The relation R_N represents transitions in the *course of events that can be considered actual with respect to a given state*; namely, we have wR_Nu only if u is an immediate successor of w and belongs to the actual future of w. Thus, the notion of actual future is *state-dependent*. This allows one to reason about the actual future of counterfactual states as well.[5]

[4] The 'independence of agents' axiom is central to stit-logic; it ensures that when choices are made *simultaneously*, an agent cannot *a priori* limit the choices available to the others; see e.g. [4, pp. 217–218]. Axiom A6 allows for the possibility that an agent has contradictory expectations about the future which cannot be realized.

[5] For instance, suppose that at w it started raining and I decided to take a walk without bringing an umbrella with me. Thus, I am in a state w' such that in the future of w' I will very likely get wet; however, had I decided to bring an umbrella with me at w, I would have ended in a state w'' such that in the future of w'' I would not have got wet. Therefore, one can also say that in the *actual future of the counterfactual state* w'' I would not have got wet.

A relational model to interpret $\mathcal{L}$ is an ordered tuple $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ where $\mathfrak{F}$ is a relational frame and V is a valuation function which maps atomic propositional symbols to sets of states and satisfies the following conditions:

- $V(\mathfrak{d}_i^{\alpha_j}) = W_{\mathfrak{d}_i^{\alpha_j}}$, for any $\mathfrak{d}_i^{\alpha_j} \in \mathcal{L}$;
- $V(\mathfrak{e}^{\alpha_j}) = W_{\mathfrak{e}^{\alpha_j}}$, for any $\mathfrak{e}^{\alpha_j} \in \mathcal{L}$.

Thus, propositional constants have the same interpretation in all models over a frame. Formulas of $\mathcal{L}$ are evaluated at a state of a model in the customary way. Truth-conditions are defined as follows:

- $\mathfrak{M}, w \vDash \chi$ iff $w \in V(\chi)$, for any $\chi \in Atom$;
- $\mathfrak{M}, w \vDash \neg\phi$ iff $\mathfrak{M}, w \nvDash \phi$;
- $\mathfrak{M}, w \vDash \phi \to \psi$ iff $\mathfrak{M}, w \nvDash \phi$ or $\mathfrak{M}, w \vDash \psi$;
- $\mathfrak{M}, w \vDash \Box\phi$ iff for all $v \in W$ s.t. wRv we have $\mathfrak{M}, v \vDash \phi$;
- $\mathfrak{M}, w \vDash N\phi$ iff for all $v \in W$ s.t. wR_Nv, we have $\mathfrak{M}, v \vDash \phi$.

Let $\mathfrak{F}, w \vDash \phi$ mean that $\mathfrak{M}, w \vDash \phi$ for all models $\mathfrak{M}$ over the frame $\mathfrak{F}$. The notion of validity of a formula with respect to (w.r.t.) a model, a frame, a class of models and a class of frames is defined in the standard way. Finally, for a given formula $\phi \in \mathcal{L}$, let $||\phi||^{\mathfrak{M}} = \{w \in W : \mathfrak{M}, w \vDash \phi\}$ and $||\phi||^{\mathfrak{F}} = \{w \in W : \mathfrak{F}, w \vDash \phi\}$. Due to the fixed interpretation of propositional constants and the definition of the translation function t, we have that, given a frame $\mathfrak{F}$ and an arbitrary model $\mathfrak{M}$ over it:

- $||t(\Delta^{\alpha_i})||^{\mathfrak{F}} = ||t(\Delta^{\alpha_i})||^{\mathfrak{M}}$, for any $\Delta \in Action^*$ and any $\alpha_i \in Agent$.

Let C_f be the class of all frames satisfying the following properties:

p(A3) for all $w \in W$, if there is $u \in W$ s.t. wR_Nu, then for all $v \in W$ s.t. wR_Nv, we have $v = u$;
p(A4) for all $w, v \in W$, if wR_Nv, then wRv;
p(A5) for all $w \in W$ and for all lists of distinct agents $\alpha_1, ..., \alpha_n$, if there are (non-necessarily distinct) action-types $\Delta_1, ..., \Delta_n$ s.t. for $1 \leq i \leq n$ there is $u_i \in W$ s.t. wRu_i and $u_i \in ||t(\Delta_i^{\alpha_i})||^{\mathfrak{F}}$, then there is $v \in W$ s.t. wRv and $v \in ||t(\Delta_1^{\alpha_1})||^{\mathfrak{F}} \cap ... \cap ||t(\Delta_n^{\alpha_n})||^{\mathfrak{F}}$;
p(A6) for all $w \in W$ and $\alpha_j \in Agent$, if there is $v \in W$ s.t. wRv and $v \in ||\mathfrak{e}^{\alpha_j}||^{\mathfrak{F}}$, then there is also $u \in W$ s.t. wRu and $u \notin ||\mathfrak{e}^{\alpha_j}||^{\mathfrak{F}}$.

The class C_f is non-empty. Indeed, the following is a very simple frame belonging to it: $\mathfrak{F} = \langle W, \{W_{\mathfrak{d}_i^{\alpha_j}} : \mathfrak{d}_i^{\alpha_j} \in \mathcal{L}\}, \{W_{\mathfrak{e}^{\alpha_j}} : \mathfrak{e}^{\alpha_j} \in \mathcal{L}\}, R, R_N \rangle$, where $W = \{w_1, w_2\}$, $W_{\mathfrak{d}_i^{\alpha_j}} = \{w_2\}$ for any $\mathfrak{d}_i^{\alpha_j} \in \mathcal{L}$, $W_{\mathfrak{e}^{\alpha_j}} = \emptyset$ for any $\mathfrak{e}^{\alpha_j} \in \mathcal{L}$ and $R = R_N = \{(w_1, w_2)\}$. It is straightforward to verify that p(A3)-p(A6) are satisfied by $\mathfrak{F}$.

Theorem 1. *The system LAE is sound w.r.t. the class C_f.*

Proof. Axioms A0, A1 and A2 are valid in all relational frames and rules R0 and R1 preserve validity in all relational frames. In the case of A3, take an arbitrary frame $\mathfrak{F} \in C_f$ and a model $\mathfrak{M}$ over it. Assume $\mathfrak{M}, w \vDash \neg N\phi$ for some $w \in W$;

from this we can infer that there is $v \in W$ s.t. $wR_N v$ and $\mathfrak{M}, v \vDash \neg\phi$; by p(A3), it follows that for all $u \in W$ s.t. $wR_N u$, $u = v$. Therefore, $\mathfrak{M}, w \vDash N\neg\phi$. In the case of A4, assume $\mathfrak{M}, w \vDash \Box\phi$; then, for all $v \in W$ s.t. wRv we have $\mathfrak{M}, v \vDash \phi$. By p(A4), we can infer that for all $u \in W$ s.t. $wR_N u$ we have $\mathfrak{M}, u \vDash \phi$. Hence, $\mathfrak{M}, w \vDash N\phi$. In the case of A5, let, for some distinct $\alpha_1, ..., \alpha_n \in Agent$ and some (non-necessarily distinct) $\Delta_1, ..., \Delta_n \in Action^*$, $\mathfrak{M}, w \vDash \Diamond t(\Delta_1^{\alpha_1}) \wedge ... \wedge \Diamond t(\Delta_n^{\alpha_n})$. From this we can infer that there are (non-necessarily distinct) $v_1, ..., v_n \in W$ s.t., for $1 \leq i \leq n$, wRv_i and $v_i \in ||t(\Delta_i^{\alpha_i})||^{\mathfrak{F}}$. By p(A5), we can infer that there is $u \in W$ s.t. wRu and $u \in ||t(\Delta_1^{\alpha_1})||^{\mathfrak{F}} \cap ... \cap ||t(\Delta_n^{\alpha_n})||^{\mathfrak{F}}$. Hence, $\mathfrak{M}, w \vDash \Diamond(t(\Delta_1^{\alpha_1}) \wedge ... \wedge t(\Delta_n^{\alpha_n}))$. In the case of A6, assume $\mathfrak{M}, w \vDash \Diamond\mathfrak{e}^{\alpha_j}$ for some $\mathfrak{e}^{\alpha_j} \in \mathcal{L}$. Then, there is $v \in W$ s.t. wRv and $\mathfrak{M}, v \vDash \mathfrak{e}^{\alpha_j}$. By p(A6), we can infer that there is $u \in W$ s.t. wRu and $\mathfrak{M}, u \vDash \neg\mathfrak{e}^{\alpha_j}$; hence, $\mathfrak{M}, w \vDash \Diamond\neg\mathfrak{e}^{\alpha_j}$.

Let $\mathfrak{F}^{LAE}$ be the canonical frame for LAE, defined as follows:

- W^{LAE} is the set of all maximally LAE-consistent sets of formulas;
- for any $w, v \in W^{LAE}$, $wR^{LAE}v$ iff $\{\phi : \Box\phi \in w\} \subseteq v$;
- for any $w, v \in W^{LAE}$, $wR_N^{LAE}v$ iff $\{\phi : N\phi \in w\} \subseteq v$;
- for any $\mathfrak{d}_i^{\alpha_j} \in \mathcal{L}$, $W^{LAE}_{\mathfrak{d}_i^{\alpha_j}} = \{w \in W^{LAE} : \mathfrak{d}_i^{\alpha_j} \in w\}$;
- for any $\mathfrak{e}^{\alpha_j} \in \mathcal{L}$, $W^{LAE}_{\mathfrak{e}^{\alpha_j}} = \{w \in W^{LAE} : \mathfrak{e}^{\alpha_j} \in w\}$.

The canonical model for LAE, denoted by $\mathfrak{M}^{LAE}$, is obtained by adding a valuation function V^{LAE} s.t.:

- for any $\chi \in Atom$, $V^{LAE}(\chi) = \{w \in W^{LAE} : \chi \in w\}$.

Any alternative valuation function V on the canonical frame must satisfy the aforementioned restrictions on propositional constants (namely, $V(\mathfrak{d}_i^{\alpha_j}) = W^{LAE}_{\mathfrak{d}_i^{\alpha_j}}$, etc.). By usual properties of canonical models, for any formula $\phi \in \mathcal{L}$ and any state $w \in W^{LAE}$, we have $\mathfrak{M}^{LAE}, w \vDash \phi$ iff $\phi \in w$.

The following theorem illustrates some properties of the frame $\mathfrak{F}^{LAE}$.

Theorem 2. *Let $R^{LAE}_{\Delta^{\alpha_i}}$ be a binary relation over W^{LAE} s.t., for any $w, v \in W^{LAE}$, $wR^{LAE}_{\Delta^{\alpha_i}}v$ iff $\{\phi : [\Delta^{\alpha_i}]^{would}\phi \in w\} \subseteq v$; we show some of the properties of this relation:*

(I) $R^{LAE}_{\Delta^{\alpha_i}} \subseteq R^{LAE}$;
(II) $R^{LAE}_{\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}} = R^{LAE}_{\Delta^{\alpha_i}} \cup R^{LAE}_{\Gamma^{\alpha_j}}$;
(III) $R^{LAE}_{\overline{\Delta^{\alpha_i}}} = R^{LAE} \cap \overline{R^{LAE}_{\Delta^{\alpha_i}}}$.

Proof. Let w be an arbitrary world in the canonical model of LAE.
(I) Assume $wR^{LAE}_{\Delta^{\alpha_i}}v$; then, $\{\phi : [\Delta^{\alpha_i}]^{would}\phi \in w\} \subseteq v$. Furthermore, let $\neg(wR^{LAE}v)$; then there is $\Box\psi \in w$ s.t. $\psi \notin v$. From this and ordinary modal reasoning it follows that $\Box(t(\Delta^{\alpha_i}) \rightarrow \psi) \in w$ and $[\Delta^{\alpha_i}]^{would}\psi \in w$, so $\psi \in v$, which represents a contradiction.
(II) Assume $wR^{LAE}_{\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}}v$. Then, $\{\phi : [\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}]^{would}\phi \in w\} \subseteq v$, which entails $\{\phi : \Box((t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j})) \rightarrow \phi) \in w\} \subseteq v$ and $\{\phi : \Box(t(\Delta^{\alpha_i}) \rightarrow \phi) \wedge \Box(t(\Gamma^{\alpha_j}) \rightarrow$

$\phi) \in w\} \subseteq v$, so $\{\phi : [\Delta^{\alpha_i}]^{would}\phi \wedge [\Gamma^{\alpha_j}]^{would}\phi) \in w\} \subseteq v$. Suppose $\neg(wR^{LAE}_{\Delta^{\alpha_i}} \cup R^{LAE}_{\Gamma^{\alpha_j}}v)$; then, there are $[\Delta^{\alpha_i}]^{would}\psi, [\Gamma^{\alpha_j}]^{would}\chi \in w$ s.t. $\psi, \chi \notin v$. From this it follows that $\Box(t(\Delta^{\alpha_i}) \rightarrow \psi), \Box(t(\Gamma^{\alpha_j}) \rightarrow \chi) \in w$. Since $\Box(t(\Delta^{\alpha_i}) \rightarrow (t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j}))) \wedge \Box(t(\Gamma^{\alpha_j}) \rightarrow (t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j}))) \in w$, then $t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j}) \in v$, which means that either $t(\Delta^{\alpha_i}) \in v$ or $t(\Gamma^{\alpha_j}) \in v$. Since we know that $wR^{LAE}_{\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}}v$ entails $wR^{LAE}v$, then $\{\phi : \Box\phi \in w\} \subseteq v$. This means that if $t(\Delta^{\alpha_i}) \in v$, then $\psi \in v$; if $t(\Gamma^{\alpha_j}) \in v$, then $\chi \in v$. A contradiction arises in both cases.

Assume $\neg(wR^{LAE}_{\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}}v)$; then, there is $[\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}]^{would}\psi \in w$ s.t. $\psi \notin v$. Therefore, $\Box((t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j})) \rightarrow \psi) \in w$. Suppose $wR^{LAE}v$ (otherwise the intended result trivially follows); then, since $\Box(\neg\psi \rightarrow \neg(t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j}))) \in w$, then $\neg(t(\Delta^{\alpha_i}) \vee t(\Gamma^{\alpha_j})) \in v$, whence $\neg t(\Delta^{\alpha_i}), \neg t(\Gamma^{\alpha_j}) \in v$, so $\{\phi : [\Delta^{\alpha_i}]^{would}\phi \in w\} \nsubseteq v$ and $\{\phi : [\Gamma^{\alpha_j}]^{would}\phi \in w\} \nsubseteq v$, hence $\neg(wR^{LAE}_{\Delta^{\alpha_i}} \cup R^{LAE}_{\Gamma^{\alpha_j}}v)$.

(III) Let $wR^{LAE}_{\overline{\Delta^{\alpha_i}}}v$; then, $\{\phi : [\overline{\Delta^{\alpha_i}}]^{would}\phi \in w\} \subseteq v$; we know that from this it is possible to infer $wR^{LAE}v$. Since $[\Delta^{\alpha_i}]^{would}t(\Delta^{\alpha_i}), [\overline{\Delta^{\alpha_i}}]^{would}\neg t(\Delta^{\alpha_i}) \in w$, then $\neg t(\Delta^{\alpha_i}) \in v$ and $t(\Delta^{\alpha_i}) \notin v$, so $\neg(wR^{LAE}_{\Delta^{\alpha_i}}v)$, which is $w\overline{R^{LAE}_{\Delta^{\alpha_i}}}v$, and $wR^{LAE} \cap \overline{R^{LAE}_{\Delta^{\alpha_i}}}v$.

Let $\neg(w\overline{R^{LAE}_{\Delta^{\alpha_i}}}v)$; then, there is $[\overline{\Delta^{\alpha_i}}]^{would}\psi \in w$ s.t. $\psi \notin v$. Assume $wR^{LAE}v$; since $\Box(\neg t(\overline{\Delta^{\alpha_i}}) \rightarrow \psi) \in w$, then $\neg t(\Delta^{\alpha_i}) \rightarrow \psi \in v$, so $t(\Delta^{\alpha_i}) \in v$. Let $[\Delta^{\alpha_i}]^{would}\chi \in w$; then, $\Box(t(\Delta^{\alpha_i}) \rightarrow \chi) \in w$ and $\chi \in v$; thus, $\{\phi : [\Delta^{\alpha_i}]^{would}\phi \in w\} \subseteq v$, which means $wR^{LAE}_{\Delta^{\alpha_i}}v$ and $\neg(wR^{LAE} \cap \overline{R^{LAE}_{\Delta^{\alpha_i}}}v)$.

Theorem 3. *The frame $\mathfrak{F}^{LAE}$ belongs to the class C_f.*

Proof. We need to show that $\mathfrak{F}^{LAE}$ satisfies the properties p(A3)–p(A6). In the case of p(A3), suppose $wR^{LAE}_N v$, $wR^{LAE}_N u$ and $v \neq u$. Then, there is ϕ s.t. $\phi \in v$ and $\phi \notin u$. In the canonical model $\mathfrak{M}^{LAE}$ we have $\mathfrak{M}^{LAE}, v \vDash \phi$ and $\mathfrak{M}^{LAE}, u \vDash \neg\phi$, so $\mathfrak{M}^{LAE}, w \vDash \neg N\phi$ and, by A3, $\mathfrak{M}^{LAE}, w \vDash N\neg\phi$, which entails $\mathfrak{M}^{LAE}, v \vDash \neg\phi$, whence $\phi, \neg\phi \in v$: contradiction. In the case of p(A4), suppose that $wR^{LAE}_N v$ and $\neg wR^{LAE}v$. Then there is $\Box\phi \in w$ s.t. $\phi \notin v$; however, by A4, $N\phi \in w$ and this entails $\neg wR^{LAE}_N v$: contradiction. In the case of p(A5), suppose that for a list of distinct agents $\alpha_1, ..., \alpha_n$ and for a list of (non-necessarily distinct) action-types $\Delta_1, ..., \Delta_n$, we have that there are (non-necessarily distinct) worlds $u_1, ..., u_n$ s.t., for $1 \leq i \leq n$, $wR^{LAE}u_i$ and $u_i \in ||t(\Delta^{\alpha_i}_i)||^{\mathfrak{F}}$. Then, $w \in ||\Diamond t(\Delta^{\alpha_1}_1)||^{\mathfrak{F}} \cap ... \cap ||\Diamond t(\Delta^{\alpha_n}_n)||^{\mathfrak{F}}$, which entails $\Diamond t(\Delta^{\alpha_1}_1) \wedge ... \wedge \Diamond t(\Delta^{\alpha_n}_n) \in w$ and, by A4, we get $\Diamond(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n)) \in w$. Assume that there is no maximally LAE-consistent set v s.t. $\{\phi : \Box\phi \in w\} \cup \{(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n)) \subseteq v$; then, $\vdash_{LAE} (\phi_1 \wedge ... \wedge \phi_m) \rightarrow \neg(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n))$ for some $\phi_1, ..., \phi_m \in \{\phi : \Box\phi \in w\}$. From this one can infer $\vdash_{LAE} \Box(\phi_1 \wedge ... \wedge \phi_m) \rightarrow \Box\neg(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n))$, so $\vdash_{LAE} (\Diamond t(\Delta^{\alpha_1}_1) \wedge ... \wedge \Diamond t(\Delta^{\alpha_n}_n) \wedge \Box(\phi_1 \wedge ... \wedge \phi_m)) \rightarrow \neg\Diamond(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n))$; however, this is impossible since we know that $\vdash_{LAE} (\Diamond t(\Delta^{\alpha_1}_1) \wedge ... \wedge \Diamond t(\Delta^{\alpha_n}_n)) \rightarrow \Diamond(t(\Delta^{\alpha_1}_1) \wedge ... \wedge t(\Delta^{\alpha_n}_n))$. Hence, we can conclude that there is a maximally LAE-consistent set v s.t. $wR^{LAE}v$ and $v \in ||(t(\Delta^{\alpha_1}_1)||^{\mathfrak{F}} \cap ... \cap ||t(\Delta^{\alpha_n}_n))||^{\mathfrak{F}}$. In the case of p(A6), assume that $wR^{LAE}v$ and $v \in ||\mathfrak{e}^{\alpha_i}||^{\mathfrak{F}}$ for some $\mathfrak{e}^{\alpha_i} \in \mathcal{L}$; then, suppose that the set $\{\phi : \Box\phi \in w\} \cup \{\neg\mathfrak{e}^{\alpha_i}\}$ is not LAE-consistent. From this one can infer that $\vdash_{LAE} \Box(\phi_1 \wedge ... \wedge \phi_n) \rightarrow \neg\Diamond\neg\mathfrak{e}^{\alpha_i}$ for some $\phi_1, ..., \phi_n \in \{\phi : \Box\phi \in w\}$; hence, $\vdash_{LAE} (\Box(\phi_1 \wedge ... \wedge \phi_n) \wedge \Diamond\mathfrak{e}^{\alpha_i}) \rightarrow \neg\Diamond\neg\mathfrak{e}^{\alpha_i}$, which contradicts A6. Then,

there is a maximally LAE-consistent set u s.t. $\{\phi : \Box\phi \in w\} \cup \{\neg\mathfrak{e}^{\alpha_i}\} \subseteq u$, which means $u \in ||\neg\mathfrak{e}^{\alpha_i}||^{\mathfrak{F}}$ and $wR^{LAE}u$.

An immediate consequence of Theorem 3 is that LAE is complete w.r.t. the class C_f; hence, together with Theorem 1, this entails that LAE is characterized by the class C_f. Furthermore, as a consequence of Theorem 2 and Theorem 3, the following schemata, which capture the properties of a boolean algebra of action-types, are provable in LAE:[6]

T1 $[\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}]^{would}\phi \equiv [\Gamma^{\alpha_j} \cup \Delta^{\alpha_i}]^{would}\phi$;
T2 $[\Delta^{\alpha_i} \cup (\Gamma^{\alpha_j} \cup \Sigma^{\alpha_k})]^{would}\phi \equiv [(\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}) \cup \Sigma^{\alpha_k}]^{would}\phi$;
T3 $[\overline{\Delta^{\alpha_i}} \cup \Gamma^{\alpha_j} \cup \overline{\Delta^{\alpha_i} \cup \Gamma^{\alpha_j}}]^{would}\phi \equiv [\Delta^{\alpha_i}]^{would}\phi$.

We will now show that the system LAE is also characterized by a subclass of C_f that includes only tree-like frames which resemble more familiar structures used in the literature for logics of agency, in particular, diverse stit-logics (e.g. [4,6,10,16]). A *branching-time frame with immediate successors* is an ordered tuple $\mathfrak{F} = \langle T, \{T_{\mathfrak{d}_i^{\alpha_j}} : \mathfrak{d}_i^{\alpha_j} \in \mathcal{L}\}, \{T_{\mathfrak{e}^{\alpha_j}} : \mathfrak{e}^{\alpha_j} \in \mathcal{L}\}, < \rangle$ where $T = \{m_1, m_2, m_3, ...\}$ is a set of moments, each $T_{\mathfrak{d}_i^{\alpha_j}}$ and each $T_{\mathfrak{e}^{\alpha_j}}$ is a subset of T and $<$ is a binary asymmetric, intransitive and backward linear relation over T, namely:

- $\forall m, m' \in T : (m < m' \rightarrow \neg(m' < m))$;
- $\forall m, m', m'' \in T : ((m < m' \wedge m' < m'') \rightarrow \neg(m < m''))$;
- $\forall m, m', m'' \in T : (m' < m \wedge m'' < m) \rightarrow m' = m''$.

We define the usual machinery related to branching-time frames. Let $\ll$ be the transitive closure of $<$; then, T is partially ordered by $\ll$ and any $\ll$-maximal chain of moments can be called a history. Let H be the set of histories in a given branching-time frame $\mathfrak{F}$ and $H_m = \{h \in H : m \in h\}$ the set of all histories in $\mathfrak{F}$ 'passing through' a moment m. A model over a branching-time frame with immediate successors is an ordered tuple $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$, where $\mathfrak{F}$ is the underlying frame and V a valuation function mapping atomic propositional symbols to moments and satisfying the usual restrictions on propositional constants.[7] Formulas of $\mathcal{L}$ are in this case evaluated with reference to a moment/history pair in a model.[8] Let *actual* be a function which associates to a moment m the only successor of m (*if any*) which belongs to the actual future of m, then:[9]

[6] Future work can be devoted to extensions of the language of LAE including operators for concatenations and iterations of action-types, in the spirit of [7,8].

[7] In the context of 'next moment' agency logic there is no need to assign atomic symbols to moment/history pairs, as observed in [6].

[8] Reference to histories provides a general framework suitable to express more complex notions related to indeterminism; for instance, one could add to the language of LAE an operator saying that something will always hold in one history passing through a given moment. Such an operator is not definable in terms of $\Box$ in infinite trees.

[9] Notice that, by definition, *actual* can be a *partial* function (a moment may have no actual successor even if an agent expects it to have some) and has some remarkable difference with the 'thin red line' function of the stit-logic literature [4]; indeed,

- $\mathfrak{M}, (m/h) \vDash \chi$ iff $m \in V(\chi)$, for any $\chi \in Atom$;
- $\mathfrak{M}, (m/h) \vDash \neg\phi$ iff $\mathfrak{M}, (m/h) \nvDash \phi$;
- $\mathfrak{M}, (m/h) \vDash \phi \rightarrow \psi$ iff $\mathfrak{M}, (m/h) \nvDash \phi$ or $\mathfrak{M}, (m/h) \vDash \psi$;
- $\mathfrak{M}, (m/h) \vDash \Box\phi$ iff for all $m' \in T$ s.t. $m < m'$ and all $h' \in H_{m'}$ we have $\mathfrak{M}, (m'/h') \vDash \phi$;
- $\mathfrak{M}, (m/h) \vDash N\phi$ iff $\mathfrak{M}, (actual(m)/h') \vDash \phi$ for all $h' \in H_{actual(m)}$.

Notice that according to the definition of $actual(m)$, if m has no actual successor, then $\mathfrak{M}, (m/h) \vDash N\phi$ for every $\phi \in \mathcal{L}$. In order to formally specify a class of branching time frames with immediate successors contained in C_f, we define the relations R and R_N in terms of moment/history pairs and the relation $<$, as follows:

- $(m/h)R(m'/h')$ iff $m < m'$, $h \in H_m$ and $h' \in H_{m'}$;
- $(m/h)R_N(m'/h')$ iff $m' = actual(m)$, $h \in H_m$ and $h' \in H_{m'}$.

The last two semantic clauses are then respectively equivalent to:

- $\mathfrak{M}, (m/h) \vDash \Box\phi$ iff for all (m'/h') s.t. $(m/h)R(m'/h')$, $\mathfrak{M}, (m'/h') \vDash \phi$;
- $\mathfrak{M}, (m/h) \vDash N\phi$ iff for all (m'/h') s.t. $(m/h)R_N(m'/h')$, $\mathfrak{M}, (m'/h') \vDash \phi$.

Let us say that a branching-time frame with immediate successors is an *lae-frame* iff it satisfies the properties p(A3)-p(A6). The class of all lae-frames can be denoted by C_f^{lae}; clearly, $C_f^{lae} \subset C_f$. In order to claim that LAE is also characterized by C_f^{lae}, one needs to show that the additional properties of lae-frames cannot be forced by any formula of the language $\mathcal{L}$. But this follows from well-known results concerning the correspondence theory of propositional modal languages. We sketch the proof below, relying on notions illustrated in [5].

Theorem 4. *For any $\phi \in \mathcal{L}$, if $C_f^{lae} \vDash \phi$, then $C_f \vDash \phi$.*

Proof. By contraposition, assume that ϕ is not valid in some model $\mathfrak{M}$ over a frame $\mathfrak{F}$ in C_f. This means that for some world w^* in the domain of $\mathfrak{M}$, we have $\mathfrak{M}, w^* \vDash \neg\phi$. Let $\mathfrak{M}'$ be the submodel of $\mathfrak{M}$ generated by w^*; then $\mathfrak{M}', w^* \vDash \neg\phi$. $\mathfrak{M}'$ can be transformed into a model $\mathfrak{M}^t$ over an asymmetric, intransitive tree $\mathfrak{F}^t$ rooted in w^*, whose set of states W^t consists of the sequences $\langle w_1, ..., w_n\rangle$ s.t. $w_1, ..., w_n \in W'$, $w_1 = w^*$ and $w_1 R' w_2, ..., w_{n-1} R' w_n$ (W' and R' being respectively the domain and the accessibility relation associated with $\Box$ in $\mathfrak{M}'$) and whose relations R^t and R_N^t are defined as follows:

- for any $u, v \in W^t$, uR^tv iff $u = \langle w_1, ..., w_n\rangle$, $v = \langle w_1, ..., w_n, w_{n+1}\rangle$ and $w_n R' w_{n+1}$;

---

the thin red line function assigns to each moment m a unique history to which m belongs (the actual history w.r.t. m), whereas *actual* assigns to m only its actual successor, if the latter exists. This solves some objections raised in [4] against the use of functions to represent actuality in branching-time; for instance, while there are problems of 'thin red line inheritance' among states related by $<$, there is no problem of 'actual successor inheritance', since any two states related by $<$ have different actual successors (if any).

– for any $u, v \in W^t$, $uR^t_N v$ iff $u = \langle w_1, ..., w_n \rangle$, $v = \langle w_1, ..., w_n, w_{n+1} \rangle$ and $w_n R'_N w_{n+1}$.

Let Π be a function from W' to $\wp(W^t)$ s.t. $\Pi(u) = \{\langle w_1, ..., w_n \rangle : w_n = u\}$; then, for all $u \in W'$ and all $\psi \in \mathcal{L}$ we have $\mathfrak{M}', u \vDash \psi$ iff $\mathfrak{M}^t, x \vDash \psi$ for every $x \in \Pi(u)$. Therefore, since $\Pi(w^*) \supseteq \{w^*\}$, we get $\mathfrak{M}^t, w^* \vDash \neg\phi$. Finally, let H^t be the set of histories in $\mathfrak{M}^t$; transform $\mathfrak{M}^t$ into a model $\mathfrak{M}^{fin}$ obtained by replacing every state $u \in W^t$ with a state $u_\sim = \{(w/h) : u \in \Pi(w)$ and $h \in H^t_u\}$. Define a binary relation $<^{fin}$ over W^{fin} s.t. $u_\sim <^{fin} v_\sim$ iff $uR^t v$; it follows that, for all $u_\sim \in W^{fin}$, $H^{fin}_{u_\sim} = H^t_u$. Let R^{fin} and R^{fin}_N be defined in terms of $<^{fin}$ as in branching-time frames with immediate successors, where $actual(w_\sim) = w'_\sim$ iff $wR^t_N w'$. $\mathfrak{M}^{fin}$ is a model over an lae-frame by construction. It can be easily proved that for all $u \in W^t$ and all $\psi \in \mathcal{L}$, we have $\mathfrak{M}^t, u \vDash \psi$ iff $\mathfrak{M}^{fin}, (w/h) \vDash \psi$ for every $(w/h) \in u_\sim$ iff $\mathfrak{M}^{fin}, u_\sim \vDash \psi$, hence $\mathfrak{M}^{fin}, w^*_\sim \vDash \neg\phi$.

We conclude with some theorems of LAE involving the operators in $E1-E3$:

T4 $([\Delta^{\alpha_1}_1]^{could}\phi_1 \wedge ... \wedge [\Delta^{\alpha_n}_n]^{could}\phi_n) \rightarrow [\Delta^{\alpha_1}_1 \cap ... \cap \Delta^{\alpha_n}_n]^{could}(\phi_1 \wedge ... \wedge \phi_n)$, where α_1,...,α_n are distinct;
T5 $[\Delta^{\alpha_i}]^{could}\phi \rightarrow \neg[\Delta^{\alpha_i}]^{could}\neg\phi$;
T6 $[\Delta^{\alpha_i}]^{will}\phi \rightarrow \neg[\Delta^{\alpha_i}]^{will}\neg\phi$;
T7 $[\Delta^{\alpha_i}]^{will}\phi \rightarrow [\Delta^{\alpha_i}]^{could}\phi$;
T8 $[\Delta^{\alpha_i}]^{could}\phi \rightarrow [\Delta^{\alpha_i}]^{would}\phi$.

T4 expresses the familiar 'independence of agents' principle in its agency appearance; T4 equivalents for 'will' and 'would' are also provable in LAE. T5 and T6 express that the defined operators for 'could' and 'will' behave in accordance with seriality. Clearly, we do not have a T5 equivalent for 'would'. T7 and T8 are bridge-theorems that express the relations between 'will', 'could' and 'would'. Finally, notice that the operators in E1-E3 can be modified by taking into account also agents' expectations, as illustrated below:

– $\mathfrak{M}, (m/h) \vDash [\Delta^{\alpha_i}]^{would}_{ex}\phi$ iff $\mathfrak{M}, (m/h) \vDash \Box((t(\Delta^{\alpha_i}) \wedge \mathfrak{e}^{\alpha_i}) \rightarrow \phi)$;
– $\mathfrak{M}, (m/h) \vDash [\Delta^{\alpha_i}]^{could}_{ex}\phi$ iff $\mathfrak{M}, (m/h) \vDash \Box((t(\Delta^{\alpha_i}) \wedge \mathfrak{e}^{\alpha_i}) \rightarrow \phi)$ and $\mathfrak{M}, (m/h) \vDash \Diamond(t(\Delta^{\alpha_i}) \wedge \mathfrak{e}^{\alpha_i})$;
– $\mathfrak{M}, (m/h) \vDash [\Delta^{\alpha_i}]^{will}_{ex}\phi$ iff $\mathfrak{M}, (m/h) \vDash \Box((t(\Delta^{\alpha_i}) \wedge \mathfrak{e}^{\alpha_i}) \rightarrow \phi)$ and $\mathfrak{M}, (actual(m)/h') \vDash t(\Delta^{\alpha_i}) \wedge \mathfrak{e}^{\alpha_i}$ for all $h' \in H_{actual(m)}$.

4 Discussion and Final Remarks

Performing Actions. Several concepts pertaining to the theory of agency introduced in this paper can be formally specified within the syntactical and semantic framework of the logic LAE. Recall (Sect. 2) that, by making reference to initial states, end-states, and counterfactual states, Von Wright derives four elementary forms of action: producing, destroying, preserving and suppressing. Although a formal approach to these terms is not

new (cf. [2,11]), the logic *LAE* allows us to expand them to more complex notions interacting with actions, expectations, instrumentality and ability:

(a) $m/h \models [\Delta^{\alpha_i}]^{prod} p$ iff $m/h \models \neg p$ and $m/h \models [\Delta^{\alpha_i}]^{will} p$ and $\exists m', \exists h' \in H_{m'}$ s.t. $m < m'$ and $m'/h' \models \neg p$

(b) $m/h \models [\Delta^{\alpha_i}]^{destr} p$ iff $m/h \models p$ and $m/h \models [\Delta^{\alpha_i}]^{will} \neg p$ and $\exists m', \exists h' \in H_{m'}$ s.t. $m < m'$ and $m'/h' \models p$

(c) $m/h \models [\Delta^{\alpha_i}]^{pres} p$ iff $m/h \models p$ and $m/h \models [\Delta^{\alpha_i}]^{will} p$ and $\exists m', \exists h' \in H_{m'}$ s.t. $m < m'$ and $m'/h' \models \neg p$

(d) $m/h \models [\Delta^{\alpha_i}]^{supp} p$ iff $m/h \models \neg p$ and $m/h \models [\Delta^{\alpha_i}]^{will} \neg p$ and $\exists m', \exists h' \in H_{m'}$ s.t. $m < m'$ and $m'/h' \models p$

The above formulae allow us to make explicit reference to the instruments that lead to producing, destroying, preserving and suppressing p, respectively. (Notice that (a)–(d) refer to atomic results.) We provide the intuitive reading of (a), the others will be similar: 'at the current state, by behaving in accordance with Δ, α_i *produces* p' means that '(i) $\neg p$ is currently the case; (ii) α actually behaves in accordance with Δ; (iii) p will actually be the case immediately after and (iv) $\neg p$ could otherwise be the case immediately after'.

Von Wright's reading of the four actions is stronger than ours, since he represents them in a *binary setting*: through agent α's conduct p will be the case, whereas through α's not-acting $\neg p$ would be the case. We believe that this account is too strong: it gives the agent α complete power over the faith of p. Definitions (a)–(d), instead, exemplify that α has the capability of determining the faith of p with some behaviour Δ, but cannot determine the faith of p through not acting.

Furthermore, observe that in our framework we can also redefine these four elementary actions in terms of *could* and *would*, as well as with reference to an agent's expectations. For the sake of discussion, we only provide the definition of 'agent α could destroy p by behaving in accordance with Δ':

(-) $m/h \models [\Delta^{\alpha_i}]^{could}_{destr} p$ iff $m/h \models p$ and $m/h \models [\Delta^{\alpha_i}]^{could} \neg p$ and $\exists m', \exists h' \in H_{m'}$ s.t. $m < m'$ and $m'/h' \models p$

Definitions (a)–(d) entail that propositions true at every next state, can neither be brought about nor prevented by any agent. Such definitions can therefore be seen as strong notions of *deliberative* action (cf. 'dstit' in [10]). This result brings us to the concept of *forbearance* (omission). Following Von Wright [15, p. 45], to forbear is stronger than to merely not act. In fact, it presupposes the ability to perform what is forborne. We introduce the following definition:

(e) $m/h \models [\Delta^{\alpha_i}]^{forb} \top$ iff $m/h \models [\Delta^{\alpha_i}]^{could} \top$ and $m/h \models [\overline{\Delta^{\alpha_i}}]^{will} \top$

Forbearance explicitly refers to actions: the usage of $\top$ (i.e., 'tautology') in (e) refers to the possibility to behave in accordance with action Δ and is interpreted as 'agent α forbears to behave in accordance with Δ' if and only if 'α could behave in accordance with Δ, but will behave in accordance with $\overline{\Delta}$ instead'.

Definitions (a)–(e) can be easily extended to formal notions of forbearance relating to results. We only illustrate the notion of 'forbearing to destroy p':

(-) $m/h \models [\Delta^{\alpha_i}]^{forb}_{destr} p$ iff $m/h \models p$ and $m/h \models [\Delta^{\alpha_i}]^{could}\neg p$ and $m/h \models [\overline{\Delta^{\alpha_i}}]^{will}\top$ and $\exists m'$, $\exists h' \in H_{m'}$ s.t. $m<m'$ and $m'/h' \models p$

Instrumentality. In Sect. 2 we made a distinction between weak and strong concepts of instrumentality, as well as agent-independent and agent-bound concepts. We will now provide their formalizations in the framework of LAE:

basic instrumentality

(f) $m/h \models [\Delta]^{b-instr}\phi$ iff $\exists m'$ s.t. $m'<m$ and for some $\alpha_i \in Agent$ we have $m'/h \models [\Delta^{\alpha_i}]^{will}\phi$

proper instrumentality

(g) $m/h \models [\Delta]^{p-instr}\phi$ iff (i) $m/h \models [\Delta]^{b-instr}\phi$ and (ii) $\forall m', \forall h'$ s.t. $m'<m$ and $h' \in H_{m'}$ and for all $\alpha_i \in Agent$ we have $m'/h' \models [\Delta^{\alpha_i}]^{would}\phi$

basic α-instrumentality

(h) $m/h \models [\Delta^{\alpha_i}]^{b-instr}_{ex}\phi$ iff (i) $m/h \models [\Delta^{\alpha_i}]^{could}_{ex}\phi$ and (ii) $\exists m'$ s.t. $m'<m$ and $m'/h \models [\Delta^{\alpha_i}]^{will}\phi$

proper α-instrumentality

(i) $m/h \models [\Delta^{\alpha_i}]^{p-instr}_{ex}\phi$ iff (i) $m/h \models [\Delta^{\alpha_i}]^{b-instr}_{ex}\phi$ and (ii) $\forall m', \forall h'$ s.t. $m'<m$ and $h' \in H_{m'}$ we have $m'/h' \models [\Delta^{\alpha_i}]^{would}\phi$

Definitions (f) and (g) employ the *will*-operator to ensure that, in the past, ϕ has been the actual result of behaviour in accordance with Δ and not just the result of lucky coincidence. Furthermore, (f) and (g) express instrumentality independent of past expectations. Moreover, (g) requires that, everywhere in the past, behaviour in accordance with Δ would have led to ϕ.

Definitions (h) and (i), instead, introduce respectively weak and strong agent-bound notions of instrumentality, the difference with the former two is that (h) and (i) consist of both future expectations and past experience: the agent expects the *continuation* of the instrument's past performance. We don't limit past experience to past expectations since an agent might discover concrete rules of instrumentality through the experience of unexpected results and actions. Observe that agent-bound instrumentality is defined through *all* three terms 'could', 'will' and 'would', relating respectively to 'the present state', 'a past state' and 'all past states'. Lastly, we emphasize that all formal definitions (f)–(i) allow for the agent to be disenchanted; that is, even proper-instruments might presently fail to lead to the intended result and agents might end in a state in which their expectations are not met.

In conclusion, taking both agent-dependent expectations and actions as the basis of our logic of agency we were able to construct three different notions of agency: *would*, *could* and *will*, each with its corresponding expectation-variant.

Together, these concepts were sufficient to address several extensions of Von Wright's elementary actions, including forbearance, as well as several formal definitions of instrumentality. As a final remark, we mention that both the process of generalizing actions and deriving notions of instrumentality are associated with induction and, for that reason, with the problems that come with it. Here, we only accentuate that the above formalization is in line with Von Wright's division of the problem of induction into two distinct problems [12]. First, there is the problem of justifying whether generalized statements are true for all observed cases (i.e., with respect to the past). This part is formally represented by definition (g). Secondly, there is the problem of using these generalized statements for future predictions. Von Wright remarks that here we seem to be satisfied with something less stringent: "Scarcely anybody would pretend that predictions, even when based upon the safest inductions, might not fail sometimes" [12, p. 51]. The latter is captured through the formal behaviour of expectations in LAE and the first clauses of definitions (h) and (i).

Acknowledgements. This work was funded by the WWTF project MA16-28.

References

1. Anderson, A.R.: A reduction of deontic logic to alethic modal logic. Mind **67**(265), 100–103 (1958)
2. Åqvist, L.: Old foundations for the logic of agency and action. Studia Logica **72**(3), 313–338 (2002)
3. Bentzen, M.M.: Action type deontic logic. J. Log. Lang. Inf. **23**(4), 397–414 (2014)
4. Belnap, N., Perloff, M., Xu, M.: Facing the Future. Agents and Choices in our Indeterminist World. Oxford University Press, Oxford (2001)
5. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001)
6. Broersen, J.: A logical analysis of the interaction between 'Obligation-to-do' and 'Knowingly Doing'. In: van der Meyden, R., van der Torre, L. (eds.) DEON 2008. LNCS (LNAI), vol. 5076, pp. 140–154. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-70525-3_12
7. Fischer, M., Ladner, R.: Propositional dynamic logic of regular programs. J. Comput. Syst. Sci. **18**(2), 194–211 (1979)
8. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. Cambridge University Press, Cambridge (2000)
9. Herzig, A., Lorini, E.: A dynamic logic of agency I: STIT, capabilities and powers. J. Log. Lang. Inf. **19**(1), 89–121 (2010)
10. Horty, J.: Agency and Deontic Logic. Oxford University Press, Oxford (2001)
11. Segerberg, K.: Getting started: beginnings in the logic of action. Studia Logica **51**(3), 347–378 (1992)
12. von Wright, G.H.: The Logical Problem of Induction. Barnes & Noble, New York (1957)
13. von Wright, G.H.: An Essay in Deontic Logic and the General Theory of Action. North-Holland Publishing Company, Amsterdam (1968)
14. von Wright, G.H.: The Varieties of Goodness. Routledge & Kegan Paul, London and Henley (1972). Fourth impression

15. von Wright, G.H.: Norm and Action: A Logical Enquiry. Routledge & Kegan Paul, London and Henley (1977). Fourth impression
16. Xu, M.: Combinations of STIT and actions. J. Log. Lang. Inf. **19**(4), 485–503 (2010)